Department of Computer Science

Methods and applications of mobile audio augmented reality

Robert Albrecht

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 122/2016

Methods and applications of mobile audio augmented reality

Robert Albrecht

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS1 of the school on 29 July 2016 at 12.

Aalto University
School of Science
Department of Computer Science

Supervising professor
Assoc. Prof. Tapio Lokki

Thesis advisor
Assoc. Prof. Tapio Lokki

Preliminary examiners
Assist. Prof. Cumhur Erkut
Aalborg University
Denmark

Assoc. Prof. Ian Oakley
Ulsan National Institute of Science and Technology
Republic of Korea

Opponents
Assoc. Prof. Federico Avanzini
University of Padova
Italy

Aalto University publication series
DOCTORAL DISSERTATIONS 122/2016

ISBN 978-952-60-6875-6 (printed)
ISBN 978-952-60-6876-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6876-3

Unigrafia Oy
Helsinki 2016

Finland

441 697
Printed matter

Author
Robert Albrecht

Name of the doctoral dissertation
Methods and applications of mobile audio augmented reality

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 122/2016

Field of research Media Technology

Manuscript submitted 15 March 2016 **Date of the defence** 29 July 2016

Permission to publish granted (date) 30 May 2016 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

In augmented reality, virtual objects are presented as if they were a part of the real world. In mobile audio augmented reality, sounds presented with headphones are perceived as if they originated from the surrounding environment. This thesis investigates potential applications of mobile audio augmented reality and different methods that are needed in these applications. The two main topics studied are distance presentation and spatial audio guidance.

Reverberation is known to be an important factor affecting the perceived distance of sound sources. Here, a practical method for modifying the perceived distance of virtual sound sources is investigated, where the temporal envelopes of binaural room impulse responses (BRIRs) are modified. In a listening test, speech sources were presented using these modified BRIRs. The results show that the perceived distance is controlled most effectively by modifying an early-to-late energy ratio with the first 50–100 ms of the BRIR included in the early energy.

Presenting large distances in an audio augmented reality environment is difficult, since people underestimate the distances of distant sound sources and very distant sound sources cannot even be heard. In a user study, the presentation of points of interest (POIs) outdoors using auditory distance cues was compared with a voice saying the distance in meters. The results suggest that distances should be given in meters if fairly accurate distance estimates are needed without prior training. With training, however, the user study participants were able to estimate the distances of the POIs fairly accurately based on the provided auditory distance cues, performing the task faster than when the distances were presented in meters.

In addition to the presentation of POIs, another type of spatial audio guidance is investigated: using spatialized music to guide pedestrians and cyclists to their destination. Two forms of guidance, route and beacon guidance, were tested in different environments. The user studies showed that music guidance is a pleasant and effective aid for navigation. Both route and beacon guidance were effective methods, but suitable for different environments and circumstances.

This thesis also investigates a mobile teleconferencing scenario, where participants can move freely from one location to another. With hear-through headphones, co-located participants can hear each other naturally. To avoid transmitting the speech of the participants to other participants in the same room – as this would be perceived as an echo – acoustic co-location detection is applied. In a user study, utilization of acoustic co-location detection was shown to improve the clarity of communication. Together, the studies presented in this thesis provide methods and guidelines for the development of mobile audio augmented reality applications.

Keywords Audio augmented reality, spatial audio, auditory distance perception, guidance, navigation, communication

ISBN (printed) 978-952-60-6875-6 **ISBN (pdf)** 978-952-60-6876-3

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2016

Pages 190 **urn** http://urn.fi/URN:ISBN:978-952-60-6876-3

Aalto-universitetet

Sammandrag

Aalto-universitetet, PB 11000, 00076 Aalto www.aalto.fi

Författare
Robert Albrecht

Doktorsavhandlingens titel
Med ljud förstärkt verklighet – mobila applikationer och metoder

Utgivare Högskolan för teknikvetenskaper

Enhet Institutionen för datateknik

Seriens namn Aalto University publication series DOCTORAL DISSERTATIONS 122/2016

Forskningsområde Mediateknik

Inlämningsdatum för manuskript 15.03.2016 **Datum för disputation** 29.07.2016

Beviljande av publiceringstillstånd (datum) 30.05.2016 **Språk** Engelska

☐ **Monografi** ☒ **Sammanläggningsavhandling**

Sammandrag

I förstärkt verklighet presenteras virtuella objekt som om de var en del av verkligheten. I med ljud förstärkt verklighet uppfattas ljud presenterade med t.ex. hörlurar som om de anlände från olika ställen i omgivningen. Denna avhandling undersöker potentiella mobila applikationer av med ljud förstärkt verklighet och olika metoder som behövs i dessa applikationer. Huvudtemana för avhandlingen är presentation av avstånd och vägledning med hjälp av spatialt ljud.

Efterklang är en faktor som inverkar på det uppfattade avståndet till en ljudkälla. Här undersöks en praktisk metod för att inverka på det uppfattade avståndet till virtuella ljudkällor genom att modifiera enveloppen på binaurala rumsimpulssvar. I ett experiment presenterades människoröster med hjälp av dessa modifierade impulssvar. Resultaten visar att det uppfattade avståndet kan kontrolleras effektivast genom att modifiera förhållandet mellan tidig och sen energi, där de första 50–100 ms av impulssvaret räknas som tidig energi.

Att presentera långa avstånd med ljud i förstärkt verklighet är svårt, eftersom människor underskattar avstånden till avlägsna ljudkällor och mycket avlägsna ljudkällor inte ens kan höras. I en användarstudie där intressepunkter presenterades utomhus jämfördes utnyttjandet av olika akustiska faktorer som inverkar på det uppfattade avståndet till en ljudkälla med att en röst sade avståndet i meter. Resultaten antyder att avstånd borde ges i meter för att de skall kunna uppskattas relativt noggrannt utan träning. Med hjälp av träning kunde dock deltagarna i användarstudien uppskatta avstånden till intressepunkterna tämligen noggrannt på basen av akustiska faktorer. Detta gjorde de snabbare än när avstånden gavs i meter.

Förutom presentationen av intressepunkter undersöks här också en annan typ av vägledning med hjälp av spatialt ljud: musik hörd från en viss riktning använd för att leda fotgängare och cyklister till sin destination. Två former av vägledning, fyr- och ruttvägledning, testades i olika omgivningar. Användarstudierna visade att musikvägledning är ett angenämt och effektivt hjälpmedel för navigation. Både fyr- och ruttvägledning var effektiva metoder, men de passade olika bra i olika omgivningar.

I denna avhandling studeras också ett mobilt telekonferensscenario, där deltagarna kan röra sig fritt från en plats till en annan. Med hjälp av akustiskt transparenta hörlurar kan deltagare i samma rum höra varandra normalt. På basen av mikrofonsignaler bestäms det vilka deltagare som befinner sig i samma rum, och deras tal skickas inte till andra i samma rum, eftersom det skulle uppfattas som ett eko. I en användarstudie visades detta förbättra kommunikationens klarhet. Tillsammans presenterar studierna i denna avhandling metoder och riktlinjer för utvecklandet av framtida mobila applikationer av med ljud förstärkt verklighet.

Nyckelord Med ljud förstärkt verklighet, spatialt ljud, uppfattande av avstånd, vägledning, navigation, kommunikation

ISBN (tryckt) 978-952-60-6875-6 **ISBN (pdf)** 978-952-60-6876-3

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

Utgivningsort Helsingfors **Tryckort** Helsingfors **År** 2016

Sidantal 190 **urn** http://urn.fi/URN:ISBN:978-952-60-6876-3

Preface

As I write these final pages of my thesis, I am filled with many different feelings. I feel glad and satisfied that I was able to put this thesis together. I feel a bit sad that my work here at the Department of Computer Science (and the Department of Media Technology before that) has almost come to its end. At the same time, I feel excited about the new opportunities that await.

First of all, I want to thank Prof. Tapio Lokki for giving me the opportunity to work under his supervision, doing research that finally ended up in the thesis you are now reading. I am grateful for the encouragement, guidance, feedback, and help he provided me with, when needed. I also thank Prof. Lauri Savioja for support during the beginning of my research.

The work done for this thesis was funded and supported by Nokia Technologies (formerly Nokia Research Center). I thank Nokia for the opportunity to work on this interesting topic, and especially all the people from Nokia that I have worked together with: Riitta Väänänen, Sampo Vesa, Jussi Virolainen, Jussi Mutanen, and Matti S. Hämäläinen. I also want to thank the other primary funder of my research work, Helsinki Doctoral Programme in Computer Science (Hecse).

For giving me a great place to work, I want to thank my colleagues in the virtual acoustics team over the years: Aki, Alex, Antti, Hannes, Henna, Jonathan, Jukka P., Jukka S., Mikael, Philip, Pierre, Raine, Sakari, Samuel, Sebastian, and everyone else I forgot to mention. Special thanks go to Kai Saksela for proofreading this thesis. I also want to thank the support staff and all the other people working at the department.

My warm thanks go to all the people who volunteered for the user studies and listening tests that I conducted, especially those of you who sat outside in close-to-freezing temperatures and patiently listened to noise

bursts and shouting voices all around you. Without you, this would not have been possible.

I thank the preliminary examiners, Prof. Cumhur Erkut and Prof. Ian Oakley, for providing me with encouraging comments and constructive criticism that helped me improve the thesis.

Finally, I want to thank my friends and family, and especially Marjukka, for giving me things to think about other than research.

Espoo, June 7, 2016,

Robert Albrecht

Contents

List of Publications

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

I Robert Albrecht and Tapio Lokki. Adjusting the perceived distance of virtual speech sources by modifying binaural room impulse responses. In *Proceedings of the 19th International Conference on Auditory Display*, Łódź, Poland, pages 233–241, July 2013.

II Robert Albrecht and Tapio Lokki. Auditory distance presentation in an urban augmented reality environment. *ACM Transactions on Applied Perception*, volume 12, issue 2, article no. 5, 19 pages, March 2015.

III Robert Albrecht, Riitta Väänänen, and Tapio Lokki. Guided by music: pedestrian and cyclist navigation with route and beacon guidance. *Personal and Ubiquitous Computing*, volume 20, issue 1, pages 121–145, February 2016.

IV Robert Albrecht, Sampo Vesa, Jussi Virolainen, Jussi Mutanen, and Tapio Lokki. Continuous mobile communication with acoustic co-location detection. In *134th Convention of the Audio Engineering Society*, Rome, Italy, paper no. 8895, 10 pages, May 2013.

Author's Contribution

Publication I: "Adjusting the perceived distance of virtual speech sources by modifying binaural room impulse responses"

The article investigates modification of the temporal envelope of binaural room impulse responses as an efficient and practical method for adjusting the perceived distance of virtual sound sources. In an experiment, participants were asked to judge the relative distances of virtual speech sources presented over headphones, with different parts of the utilized binaural room impulse responses either amplified or attenuated. The results indicate that modification of an early-to-late energy ratio, where approximately 50–100 ms of the impulse response is included in the early energy, alters the perception of distance more effectively than modification of the direct-to-reverberant energy ratio.

The present author conceived the idea, designed and conducted the experiment, and analyzed the results. The present author wrote the article with feedback provided by Prof. Tapio Lokki.

Publication II: "Auditory distance presentation in an urban augmented reality environment"

The article investigates the presentation of distances of points of interest in an outdoor augmented reality environment using audio. In a user study, presenting distances with a combination of several auditory distance cues was compared with a voice saying the distance in meters. Giving the distance in meters resulted in fairly accurate and precise distance estimates without prior training, but the participants could after a short period of training also make fairly good distance estimates utilizing only

auditory distance cues. Using auditory distance cues for distance presentation can thus be recommended in applications where it is beneficial to limit the length of the presented messages and where high precision is unnecessary.

The present author designed and conducted the experiment, as well as analyzed the results. The present author wrote the article with feedback provided by Prof. Tapio Lokki.

Publication III: "Guided by music: pedestrian and cyclist navigation with route and beacon guidance"

The article investigates pedestrian and cyclist navigation where the user follows the direction of spatialized music to reach the destination. Two different guidance types, route and beacon guidance, were evaluated in three different user studies. Most participants in the user studies were able to easily navigate using the auditory cues and enjoyed the experience. Route and beacon guidance were found to be appropriate for different situations depending on the preferences of the user.

The navigation software and hardware used in the user studies were provided by Nokia Technologies. The present author had the main responsibility for planning and conducting the user studies, with Dr. Riitta Väänänen participating in much of the process. The present author analyzed the results and wrote approximately 90% of the article.

Publication IV: "Continuous mobile communication with acoustic co-location detection"

The article proposes an acoustic co-location detection algorithm for mobile communication applications, in particular. The algorithm infers the co-location of two persons based on the signals from microphones that they carry. The algorithm was first evaluated with recordings of different mobile communication scenarios, where co-location was correctly identified most of the time. The algorithm was then integrated into a voice-over-IP conferencing system and evaluated in a user study, where it was shown to improve the clarity of communication.

The acoustic co-location detection algorithm was invented and implemented by Dr. Sampo Vesa, while Jussi Virolainen and Jussi Mutanen

were responsible for the conferencing system. The present author was responsible for making the recordings for evaluating the co-location detection algorithm, while the evaluation was performed by Dr. Sampo Vesa. The present author planned and conducted the user study, as well as analyzed the results. The present author wrote approximately 80% of the article.

List of Acronyms

AAR	Audio Augmented Reality
ACLD	Acoustic Co-Location Detection
AR	Augmented Reality
BRIR	Binaural Room Impulse Response
GPS	Global Positioning System
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
ILD	Interaural Level Difference
IQR	Interquartile Range
ITD	Interaural Time Difference
MFCC	Mel-Frequency Cepstral Coefficient
POI	Point Of Interest
RIR	Room Impulse Response
RMS	Root Mean Square
RMSE	Root Mean Square Error
VAD	Virtual Auditory Display
VoIP	Voice over IP

List of Symbols

θ	elevation angle
φ	azimuth angle
C_T	clarity, measured as an early-to-late energy ratio
d	distance
D/R	direct-to-reverberant energy ratio
$\bar{E}$	time-averaged energy
E/L	early-to-late energy ratio
$h(t)$	impulse response
n	number of observations or pairs
p	p-value
r_{ij}	correlation between i and j
S	sign test statistic
t	time
t_d	point in time immediately after the direct sound
V	circular variance
W	Wilcoxon rank sum test statistic
Z	Z-statistic

1. Introduction

Augmented reality is the extension of the real world with virtual objects (Azuma, 1997). This thesis discusses methods and applications of extending our acoustic environment with virtual sound sources.

1.1 Motivation

Most research on augmented reality has focused on augmenting our visual perception of the world around us, while comparatively very little research has been done on audio augmented reality. However, audio augmented reality has huge potential, both in combination with visually augmented reality and on its own.

One of the greatest advantages of audio is that sounds can be heard from any direction, whereas our field of view is limited. Current smartphone-based visual augmented reality applications are even further limited by the camera and the display of the device. Compared with such systems, auditory displays leave the hands and visual attention free for other tasks.

The aim of this thesis is to provide methods and knowledge for the development of future audio augmented reality applications. The usefulness of the developed methods is evaluated through real-time applications and user studies. These studies also provide insight into some of the perceptual aspects of mobile audio augmented reality applications. By examining both methods and applications thereof, the thesis provides a broad investigation of the subject that helps to pave the way for the widespread use of mobile audio augmented reality technology.

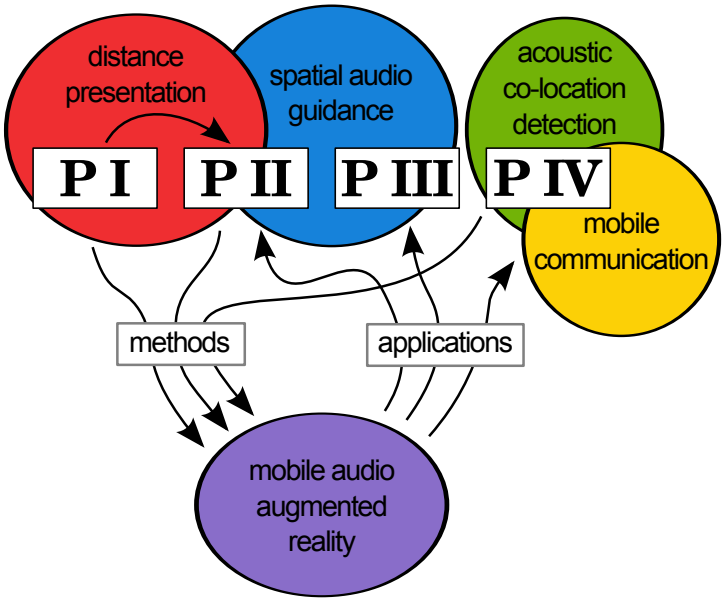


Figure 1.1. The main concepts covered by the publications included in this thesis and the relationship between them. The arrows show which publications provide methods for mobile audio augmented reality and which publications apply different augmented reality methods and technologies.

1.2 Scope of the thesis

Figure 1.1 illustrates the main concepts covered by this thesis and how these are interconnected. The methods presented in the thesis are the distance presentation methods investigated in Publications I and II, and the acoustic co-location detection method of Publication IV. The applications that are studied are two different types of spatial audio guidance in Publications II and III, and mobile communication in Publication IV.

Distance presentation methods and spatial audio guidance applications
To produce an auditory event somewhere in the space surrounding a listener, the percept of both direction and distance must be produced. While auditory localization in azimuth and elevation has been extensively studied, distance perception has received comparatively little attention (Zahorik et al., 2005). Even less research has focused on developing methods for presenting virtual sound sources at different distances. Publication I presents a practical method for distance presentation in mobile audio augmented reality applications, by modifying the early-to-late energy ratio of binaural room impulse responses. In doing so, the study also sheds some new light on the perception of distance.

Publication II combines this method with other auditory distance cues and investigates how these together can be utilized in an outdoor augmented reality environment to present the location of different points of interest. The main research question here is how we can take auditory distance cues that naturally produce auditory events at relatively short distances and use them to present points of interest at much larger distances. This study thus investigates both distance presentation methods and spatial audio guidance applications of augmented reality. The utilization of auditory distance cues is also compared with the simple alternative of having a voice announcing the distance to the points of interest.

Whereas the user study in Publication II focuses on the localization of points of interest in a static setting, Publication III utilizes mobile audio augmented reality technology to provide spatial audio guidance for pedestrians and cyclists. The study investigates if and how music can be used as a pleasant sound to guide users to their destination: the music is heard from the direction where the user should be heading.

Acoustic co-location detection methods and mobile communication applications

Publication IV presents another application of audio augmented reality technology: communication. Augmented reality methods lend themselves well to mobile teleconferencing, where some participants are co-located and some are at different locations. In this type of application, participants should be able to hear other participants at the same location unhindered, while the voices of participants at other locations could be presented with spatial audio, as though they were among the participants at the same location. Since co-located participants are able to hear each other naturally, co-location should be detected and audio should not be transmitted between these participants. In this study, the application of acoustic co-location detection was investigated in a mobile teleconferencing scenario.

The distance presentation method of Publication I is useful in many different augmented reality and spatial audio applications, such as the presentation of points of interest investigated in Publication II. While the mobile communication application used to evaluate the acoustic co-location detection method in Publication IV currently did not utilize spatial audio, the inclusion of spatial audio would both make the conversation more

natural and improve speech intelligibility. Distance presentation methods could be added, e.g., to indicate the proximity of the other participants, such as co-workers at a construction site. The navigation using music guidance in Publication III, on the other hand, is an example of an application where this type of distance presentation method might be disadvantageous: large modifications of the reverberation might reduce the enjoyability of the music, while only providing a vague indication of the remaining distance to the destination.

The applications studied in this thesis all benefit from one of the basic principles of mobile audio augmented reality: sounds from the environment should be heard together with the sounds presented over headphones. The reasons may, however, differ slightly depending on the application. In some applications, such as the presentation of points of interest (Publication II), virtual sound sources are added to the surrounding environment. In navigation applications (Publication III), sounds from the surroundings should be heard for safety reasons. In teleconferencing applications (Publication IV), co-located participants should be heard naturally at the same time as remote participants are heard over headphones.

The publications included in this thesis provide methods for mobile audio augmented reality applications as well as examples of applications that utilize these and other audio augmented reality techniques. Together, they will hopefully serve as an incentive for an increasing number of applications utilizing such methods and techniques.

1.3 Structure of the thesis

Chapter 2 discusses augmented reality and mobile audio augmented reality, in particular. General methods, technology, and applications of audio augmented reality are presented. Chapter 3 deals with distance presentation in audio augmented reality, presenting the results of Publications I and II. Chapter 4 presents the use of audio augmented reality techniques for spatial audio guidance and, in particular, the music guidance application investigated in Publication III. In Chapter 5, mobile communication applications of augmented reality are discussed and the results of Publication IV presented. Finally, Chapter 6 summarizes the research findings and presents possible directions for further research.

1.4 Ethics statement

The participants in the listening tests and user studies performed in Publications I, II, III, and IV were informed about the procedures of the studies and volunteered to participate. The gathered data were anonymized so that they cannot be linked back to the individual participants. The studies did not expose the participants to exceptionally strong stimuli or put the participants at a risk of harm beyond the risks encountered in normal life. According to the instructions of the Aalto University Research Ethics Committee, an ethical review was therefore not required.

2. Augmented Reality

Augmented reality (AR) is defined by Azuma (1997) as having three distinct features:

1. It combines real and virtual.
2. It is interactive in real time.
3. It is registered in three dimensions.

Another way to say this might be: in augmented reality, virtual objects are superimposed on our perception of the real world, having a location in the real world that is retained regardless of how we move or rotate. A stricter interpretation might also imply that the virtual objects should augment the real world in a meaningful way.

Most augmented reality research has dealt with visual augmented reality. Visual augmented reality can be achieved with the aid of, e.g., see-through displays, allowing the user to see both the real world and virtual objects superimposed on it. Mobile visual augmented reality can be achieved using head-mounted displays or simply by using a smartphone as a camera-based "see-through" display.

2.1 Audio augmented reality

Audio can be used to aid the visual modality in multimodal AR applications, but audio may also be used alone in unimodal audio augmented reality (AAR) applications. Such applications leave the visual sense completely free to observe the surrounding world. Utilizing the auditory modality also exhibits some notable differences compared with the visual modality.

Sarter (2006) explains the differences between auditory and visual channels of information presentation. First, our auditory perception has an omnidirectional character: we can perceive sounds from any direction any time. Secondly, sounds typically have a transient nature: a written sign can be read many times as long as it is visible, but a spoken sentence can only be heard once, unless it is repeated. Lastly, while we can shut out visual messages by closing our eyes or looking the other way, more drastic measures are required to shut out auditory messages. Kolarik et al. (2015) mentions a further difference: sound can normally travel around obstructing objects.

In augmented reality, the omnidirectional character of auditory perception is often an advantage compared with visual approaches. Applications can draw the user's attention and spatially present information in any direction any time. This advantage becomes even greater if we consider smartphone-based applications, where visual output is limited not only to the field of view but additionally to a small display. Virtual sound sources can also be perceived even if they are located behind other virtual or real objects. The transient nature of many sounds, on the other hand, may or may not be an advantage, depending on the application. Similarly, the fact that the auditory sense is our primary warning sense may be used to our advantage, but also misused (Edworthy and Hellier, 2005).

2.2 Mobile audio augmented reality

The architecture of a mobile audio augmented reality system is illustrated in Fig. 2.1. The hardware comprises acoustically transparent (hear-through) headphones, sensors for extracting orientation, location, and possibly context information, and a device, such as a smartphone, running the AAR application and providing a user interface for interacting with the application. The acoustically transparent headphones allow the user to hear the surrounding acoustic environment as well as virtual sound sources that augment it. These virtual sound sources are produced by applying spatial audio rendering techniques to the audio that is presented by the application. Utilizing information about the surrounding acoustic environment as well as the location and the orientation of the user, the sounds are presented so that they are perceived to originate from specific locations in the real world surrounding the user.

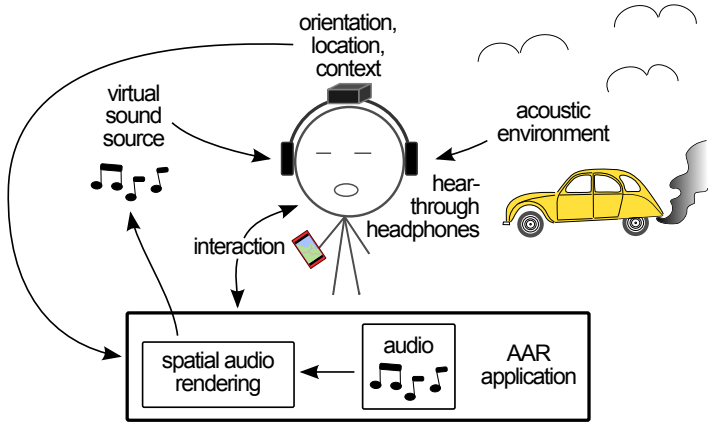


Figure 2.1. The architecture of a mobile audio augmented reality system.

2.2.1 Virtual auditory display

A sound that is heard by a human listener is affected by the path from the sound source to the eardrums. The head as well as the pinnae affects the sound reaching the eardrums in different ways. In addition, sound reaches the listener both through a direct path and through reflections via different surfaces. All these modifications to the sound emitted from the source not only help the listener to tell from where the sound originated, but also provide information about the space surrounding the source and the listener.

By means of spatial audio rendering techniques, the same types of modifications can be applied to sounds presented over headphones, to give the listener the perception that the sound originates not from the headphones, but from somewhere in the space surrounding the listener. This type of system is called a virtual auditory display (VAD) (Shinn-Cunningham, 1998).[1] AAR applications thus utilize VAD techniques to present virtual sound sources that are superimposed on the surrounding acoustic environment.

Presentation of direction

The two main cues used for localization in the horizontal plane are the interaural level difference (ILD) and the interaural time difference (ITD). The interaural level difference is due to the head shadowing the ear further away from the sound source. The interaural time difference is due to

[1]The terms virtual acoustic display (Wenzel, 1992) and virtual audio display (Brungart, 2002) have also been used.

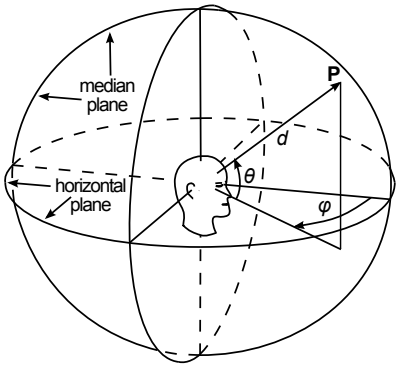


Figure 2.2. A head-centric coordinate system, showing the azimuth φ, elevation θ, and distance d of a point **P** in space, relative to the centre of the head. The median plane and the horizontal plane are also shown.

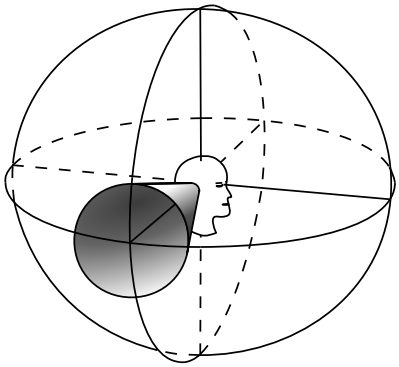


Figure 2.3. The cone of confusion. Interaural level and time differences provide ambiguous localization cues. For a single ILD or ITD, the sound source may lie somewhere on the surface of a cone.

the longer travel time to the ear further away. The interaural level and time differences are thus zero if the sound source is located in the median plane (see Fig. 2.2).

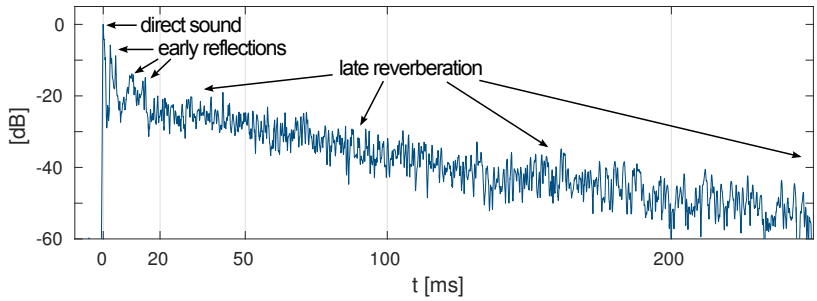
Adding ILD or ITD cues to sounds presented over headphones results in what is called lateralization (Plenge, 1974; Begault and Wenzel, 1992; Durlach et al., 1992). Depending on the cues, the sound is perceived as coming from the left or the right earpiece of the headphones, or somewhere between them. In other words, the sound is not perceived as coming from a direction outside the head, but rather from a source inside the head.

Instead of using only ILD and ITD cues, it is in many cases beneficial to present users with all the sound localization cues that are naturally available. Apart from the two mentioned cues, humans utilize other cues that aid in both horizontal and vertical localization. Many of these cues are due to the pinna affecting the temporal and spectral characteristics of sound entering the ear canal through reflections, shadowing, dispersion, diffraction, interference, and resonance (Blauert, 1997, p. 63). By utilizing these cues in addition to the ILD and ITD cues to present sounds, the sounds can be perceived as originating from outside the head. However, successful externalization of auditory events does not depend solely on these pinna cues. Utilizing pinna cues does not guarantee that the sound source is perceived outside the head, and neither are pinna cues always necessary to achieve externalized sound sources (Durlach et al., 1992).

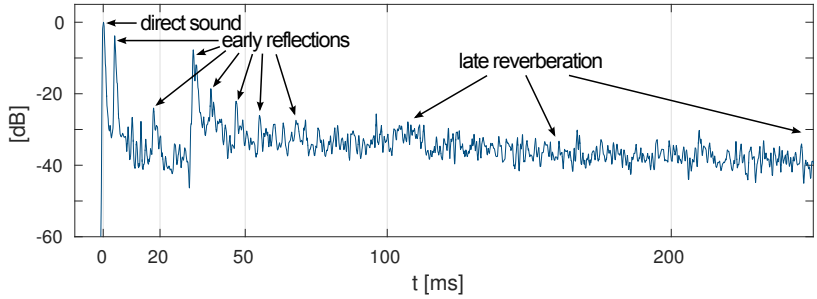
Since low-frequency sound diffracts effectively around the head, interaural level differences serve as a localization cue mainly at higher frequencies. Interaural time differences, on the other hand, serve as a localization cue mainly at low frequencies, since the phase differences these produce are ambiguous at higher frequencies. At frequencies in between, from 1000 to 4000 Hz, localization accuracy decreases (Rossing et al., 2002, p. 90). Pinna cues, which are available only at high frequencies, modify the spectrum of sound. Thus, prior knowledge about the spectrum is required to utilize these cues, unless additional cues are provided by head movements (Durlach et al., 1992).

The head-related transfer function (HRTF) and its time-domain counterpart, the head-related impulse response (HRIR), contain all these cues for a sound source in a specific direction. HRIRs are either simulations or, more often, measurements of the impulse response from a sound source to the ear canals of either a human subject or a dummy head. By convolving a sound with HRIRs measured in a certain direction, the sound can be presented with headphones so that the listener perceives it as arriving from that direction. While HRIRs are measured under anechoic conditions, binaural room impulse responses (BRIRs) also include sound that reaches the listener through reflection from different surfaces in a reverberant space. Even though the reflected sound, which arrives after the direct sound due to the longer path it travels before reaching the listener, contains localization cues of its own, the auditory system is able, within certain limits, to fuse the reflected sound to the perception of the direct sound. Consequently, the localization cues of the direct sound dominate the perception – a phenomenom called the precedence effect (Litovsky et al., 1999). However, if a reflection is strong enough and sufficiently separated in time from the direct sound, it is perceived as a separate echo.

Figure 2.4 shows two examples of BRIRs: one from a mildly reverberant office room and the other from a highly reverberant hall. BRIRs, as well as room impulse responses (RIRs) in general, can be divided into three parts: the direct sound, early reflections, and late reverberation. The boundary between early reflections and late reverberation is not clearly defined. Instead, there is a gradual transition from the discrete early reflections to the approximately diffuse sound field created by the late reverberation. This transition time, often called mixing time, can be estimated based on either objective measures or perceptual characteristics of the impulse response (Lindau et al., 2010). The mixing time generally increases as the



(a) A mildly reverberant office room.



(b) A highly reverberant hall.

Figure 2.4. Two examples of binaural room impulse responses, taken from the Aachen Impulse Response Database (Jeub et al., 2009, 2010). The first 250 ms of the responses of the right ear are shown.

room volume increases, as can be observed by comparing the two spaces in Fig. 2.4. The larger and more reverberant a space is, the slower the decay rate of the late reverberation also is. The properties of early reflections and late reverberation are discussed in more detail in Chapter 3.

The benefit of using HRTFs instead of the much more simple ILD and ITD cues is the ability to present sound sources that are perceived outside the head and at both different azimuths and different elevations (Begault and Wenzel, 1992) (see the coordinate system in Fig. 2.2). Additionally, sound sources in front of and behind a listener (e.g., at 30° and 150° azimuth) produce identical ILD and ITD cues, so the listener is not able to unambiguously localize the sound source without additional cues. In fact, the ILD and ITD cues imply that the sound source is somewhere on the surface of a cone, the so called cone of confusion (Fig. 2.3). HRTFs contain cues that help the listener to establish whether the sound source is in front of or behind him or her, or somewhere else on the cone of confusion. The main disadvantage of using HRTFs lies in the increase in complexity and required processing power.

Different databases of measured HRTFs are freely available. Two popular alternatives are the CIPIC (Algazi et al., 2001) and Listen (Warusfel, 2003) databases. Other options include databases by Gardner and Martin (1994), Blauert et al. (1998), Qu et al. (2009), Wierstorf et al. (2011), and Gómez Bolaños and Pulkki (2012). Since HRTF measurements are done in discrete directions, an interpolation scheme is also needed for producing HRTFs representing directions between these (Gamper, 2013). This is especially important for sound sources that move with respect to the listener.

Externalization

While the use of HRTFs instead of simple ILD and ITD cues should produce virtual sound sources that are perceived to be outside the head, this is not always the case. On the contrary, it is common that virtual sound sources presented straight in front of or behind the listener are instead perceived to be inside the head (Begault and Wenzel, 1992, 1993). This is referred to as in-the-head localization, or the lack of externalization.

Reverberation has been shown to improve externalization, and can be applied, e.g., using BRIRs. Both Begault et al. (2001) and Catic et al. (2015) found that early reflections (approximately the first 80 ms of the BRIRs) are sufficient for achieving externalization; late reverberation is not necessary. Catic et al. also found monaural reverberation cues to be sufficient for externalization of lateral ($\varphi = 30°$) sound sources, while binaural reverberation cues were needed to provide externalization for a frontal ($\varphi = 0°$) source. Head orientation tracking (discussed under Section 2.2.2) has also been shown to improve externalization in some cases (Brimijoin et al., 2013), but not in others (Begault et al., 2001). The use of individualized HRTFs, i.e., HRTFs measured for the individual in question, had no effect on externalization in the experiments of Begault et al. (2001), but other studies have found that the degree of externalization varies between different HRTFs (Seeber and Fastl, 2003; McMullen et al., 2012).

Front-back confusion

Since the primary localization cues, the ILD and the ITD, provide ambiguous information about the location of a sound source, confusion often arises in virtual auditory displays, where other localization cues, including visual cues, might be imperfect or completely missing. In particular,

it is common for a virtual sound source in front of the listener to be perceived in the corresponding position behind the listener (and, less commonly, vice versa), since the two positions have the same ILD and ITD. Similarly, up-down reversals might also take place due to this ambiguity. This type of reversals seem to be less common than front-back reversals (Wenzel et al., 1993; Bronkhorst, 1995; Carlile et al., 1997), although a large degree of up-down reversals has been observed under some conditions (Wenzel, 2001).

If the listener and the sound source stay still, the ILD and the ITD do not change either. Thus, they indicate that the sound source is somewhere on a cone of confusion (Fig. 2.3). However, if the listener rotates his or her head, the ILD and the ITD change accordingly, giving rise to a new cone of confusion. The change of the ILD and the ITD associated with the head rotation serves as a cue for the listener to infer the correct position of the sound source, which is located where the new and the old cones of confusion meet.

Wightman and Kistler (1999) and Begault et al. (2001) have shown that head rotation indeed reduces front-back reversals. Wightman and Kistler showed that front-back confusion also diminishes if the listener is able to move the sound source, since this equally results in a predictable change in the ILD and the ITD. On the other hand, if the source moves without the listener interacting with it, the front-back confusion remains.

Mariette (2007) investigated the mitigation of front-back confusion by body motion, in the absence of head rotation cues. The participants in the experiment walked short distances, during which the azimuth and the distance to virtual sound sources presented over headphones changed. The results displayed significantly improved front-back localization when the azimuth changed at least 12° or the distance change was more than 21% of the initial distance. Mariette notes that these changes correspond with the resolution of the spatial audio rendering and speculates that increasing the resolution would allow smaller azimuth and distance changes to mitigate front-back confusion. In the experiment, the participants were told to make their best guess concerning the location of the sound source, instead of indicating where they perceived it to be. Thus, it is unclear if body motion actually resolved perceptual front-back reversals, or if it only helped participants to make better guesses.

Since front-back confusion arises from inadequate localization cues, one might assume that using as naturalistic localization cues as possible in

virtual auditory displays would help to mitigate this problem. Since people's heads and ears have different shapes and sizes, the corresponding localization cues in the HRTFs also differ from person to person. The best localization cues should thus be provided by using the HRTFs of the person in question, since he or she is used to the localization cues that these provide.

The results of Wenzel et al. (1993) suggest that using nonindividualized instead of individualized HRTFs increases the rate of front-back reversals, as expected. However, Bronkhorst (1995) and Begault et al. (2001) did not find any effect of using individualized HRTFs on the amount of front-back reversals. It might thus be that other factors in the implementation of the virtual auditory display play a more important role in determining the amount of front-back confusion than the choice between individualized and nonindividualized HRTFs.

In any case, measuring individualized HRTFs is currently not practical for consumer applications. Rather, suitable HRTFs could be selected from a database using some appropriate method (Seeber and Fastl, 2003; Härmä et al., 2012; McMullen et al., 2012; Schönstein and Katz, 2012), based on criteria such as externalization, front-back discrimination, and up-down discrimination. Alternatively, parameterized HRTFs could be customized for each user based on, e.g., photographs of the user's pinnae (Spagnol et al., 2014).

Presentation of distance

To present virtual sound sources in three dimensions, not only do we need to control the azimuth and elevation, but also the distance of the sources. Several different factors affecting the perceived distance of sound sources have previously been identified, including intensity, reverberation, and binaural cues (Shinn-Cunningham, 2000; Zahorik et al., 2005; Kolarik et al., 2015).

Intensity is the most simple of cues: the closer a sound source is, the louder it sounds. On the other hand, it requires familiarity with the sound source, otherwise it can only serve as a relative cue (Shinn-Cunningham, 2000). Reverberation, on the other hand, can be used as an absolute cue. While the level of the direct sound from a sound source inside an enclosed space varies with distance, the level of the diffuse field caused by the late reverberation does not vary with distance. The human auditory system can take advantage of this fact and use the relationship between these

levels as a distance cue. Publication I investigates how such reverberation cues can be utilized in a virtual auditory display.

Binaural cues arise from the fact that the interaural level and time differences behave differently when the sound source is close to listener (at a distance less than approximately 1 m). While the ILD increases strongly as the sound source moves closer to the head, the ITD increases only slightly. Instead of the cone of confusion (Fig. 2.3) present at larger distances, ILD and ITD cues give rise to a torus of confusion at short distances (Shinn-Cunningham et al., 2000). This torus thus determines not only the possible directions of the sound source, but also the distance. These binaural cues can be used to create near-field virtual auditory displays (Brungart, 2002; Spagnol et al., 2012), but should be used in combination with other cues, since interaural differences are nonexistent in the median plane.

Other distance cues include spectral (Larsen et al., 2008), visual (Calcagno et al., 2012; Hládek et al., 2013), and speech cues (Brungart and Scott, 2001). Chapter 3 and Publications I and II discuss distance presentation in more detail.

Integration

Audio augmented reality combines the real acoustic environment with a virtual acoustic environment. In some applications, it might be beneficial that the virtual environment is clearly distinguishable from the real environment, but in other applications, such as games, the goal is to integrate the virtual sound sources into the surrounding environment. In other words, the virtual sounds should sound like they came not only from directions around the user, but they should have a suitable reverberation so that they sound like they actually originated from the acoustic environment around the user.

As of yet, little research has dealt with the issue of integration in audio augmented reality. Lindau and Weinzierl (2012) investigated the plausibility of virtual acoustic environments compared with real environments. In this study, both virtual and real acoustic events were presented and the participants were asked whether they thought that the events were real or not.

While this type of methodology could be applied to assess the virtual auditory displays in audio augmented reality, it might be a bit strict for such applications. First of all, it is not in practice possible to achieve a per-

fect match between the virtual acoustics and the real acoustics in a mobile AAR application, since the user may freely move around from one acoustic environment to another, completely different, environment. The best we can do is to extract information about the current environment from, e.g., microphone signals (Vesa and Härmä, 2005; Gamper and Lokki, 2009) or location information, and choose the most appropriate reverberation from a database or through real-time synthesis.

However, a perfect match between the real and the virtual acoustics is not necessary, either. Since visual stimuli corresponding to the auditory stimuli are missing, the user can only temporarily be given the impression that the virtual sound source is real. Thus, in mobile AAR applications, it is sufficient to give the user the "feeling" that the virtual sound source originates from the surrounding acoustic environment, rather than trying to convince him or her that it actually does.

When aiming to achieve natural timbre in AAR applications, one must take into account the non-ideal magnitude response of the headphones. Typical headphones are not designed to have a flat magnitude response, but the response is instead often designed to match that of stereo loudspeaker listening, either in a free or a diffuse field (Møller et al., 1995). In order to present sounds with a natural timbre, the headphone response must be compensated for (Schärer and Lindau, 2009; Lindau and Brinkmann, 2012).

2.2.2 Hardware

As shown in Fig. 2.1, the hardware needed for a mobile AAR system consists of acoustically transparent headphones, sensors for tracking head orientation and location, as well as a device that runs the AAR application and provides the user with an interface for interacting with the application.

Hear-through headphones

In augmented reality, virtual objects are superimposed on our perception of the surrounding world. In mobile audio augmented reality, this means, among other things, that we should hear the surrounding acoustic environment in addition to sounds that presented with headphones. Mobile AAR thus requires acoustically transparent headphones, which attenuate or modify sounds from the surroundings as little as possible, retaining not only their audibility, but also their localization cues.

Lindeman et al. (2007) divide techniques for achieving acoustic transparency into two categories: microphone hear-through and acoustic hear-through. Acoustic hear-through is achieved by using headphones that attenuate environmental sounds and distort their localization cues as little as possible. Regular open headphones can be used (Publication II), but improved perception of the environment may be achieved with special headphones (Martin et al., 2009). One alternative is to use bone-conduction headphones, which leave the ears completely free of any obstructions.

Microphone hear-through utilizes headphones that attenuate sounds from the environment to some extent. To achieve acoustic transparency, microphones are attached to the outside of the headphones. The environment can be perceived unattenuated by appropriately amplifying the microphone signals and passing them to the headphones. While binaural localization cues (ILDs and ITDs) are quite robust with respect to microphone placement, monaural cues are best preserved if the microphones are placed as close as possible to the ear canal entrances. This can be achieved by the use of insert headphones.

Since blocking the ear canal changes its resonances, equalization of the microphone signals is needed in order to achieve natural timbre (Tikander et al., 2008). Equalization is also needed to compensate for the headphone response and the leakage of low-frequency sound into the ear canal. As the leakage depends on the fit of the headphones, adaptive isolation estimation may be utilized to control the equalization (Liski, 2016).

Tikander et al. (2008) presented an augmented reality audio (ARA) mixer, which performed equalization and amplification of the microphone signals and mixed these with other sounds that were presented with insert headphones. Over the years, the ARA mixer has been developed further by the inclusion of a USB audio interface and user-adjustable hear-through level (Albrecht et al., 2011), digital signal processing (Rämö and Välimäki, 2012), and software control of the hear-through level (Publication III).

The main advantage of microphone hear-through is the possibility to adjust how well the environment is heard, which can be useful both for comfort and for hearing protection purposes. With this technology, it is also possible to modify the perceived acoustic environment to some extent.

Microphone hear-through is, however, also associated with disadvantages. Even if equalization is applied, some colouration of sounds from the

surrounding acoustic environment is likely to remain. Some deterioration of localization cues is also to be expected (Møller et al., 2015). Microphone hear-through also introduces some noise, and wind noise is a further problem (Tikander, 2009; Albrecht et al., 2011; Miura et al., 2013). Through careful design and choice of components, these effects can be minimized. Compared with acoustic hear-through, microphone hear-through also requires additional battery-powered hardware for signal processing.

One source of discomfort with microphone hear-through using insert headphones is the occlusion effect (Tikander, 2009; Albrecht et al., 2011). Due to this effect, low-frequency bone-conducted sound can be amplified up to 30 dB or more when an earplug is shallowly inserted in the ear canal (Stenfelt and Reinfeldt, 2007). To mitigate this problem, some form of occlusion cancellation could be applied, e.g., through analog feedback (Mejia et al., 2008) or with the help of adaptive filtering (Borges et al., 2013; Sunohara et al., 2014).

Orientation tracking

Härmä et al. (2003) make the distinction between "localized acoustic events connected to real-world objects" and "freely-floating acoustic events." To present freely-floating acoustic events, no orientation tracking is needed. Instead, the virtual sound sources are tied to a head-centric frame of reference. The sound sources thus rotate around the user as the user rotates his or her head, and move as the user moves. The location of the sources relative to the user can thus be used to present information. For example, the azimuth of the source might indicate time. Alternatively, spatial separation of sources might be used to improve the clarity of communication.

Localized acoustic events, on the other hand, are tied to real-world objects or locations. To be able to present these in a mobile application, the position of the user relative to these objects or locations must be known. In addition, the orientation of the head of the user must be known, so that virtual sound sources can be presented at stable locations regardless of how the user turns his or her head.

At the time of writing, head orientation tracking is yet to be made available for large-scale commercial applications. Research projects typically use separate and often expensive head tracking devices, but ideally, orientation tracking should be integrated into the headphones for commercial

applications. Such products, e.g., the Jabra Intelligent Headset[2] and the Bragi Dash,[3] are currently beginning to emerge on the market. Smartphones also contain sensors (magnetometers, gyroscopes, and accelerometers) for orientation tracking. These can be utilized for head tracking, e.g., by attaching the smartphone to the headphones (Zwinderman et al., 2011), but such a method is unlikely to gain widespread adoption.

If head tracking is not feasible, tracking the orientation of a smartphone or other mobile device held in the hand or carried in a pocket may suffice for some applications. Heller et al. (2014) compared tracking of the head, the body, and a hand-held device for audio augmented reality applications. Nine participants were asked to walk to different virtual sound sources in a large hall and they answered a presence questionnaire after each of the three conditions. Head tracking received slightly, but not significantly, better ratings than device and body orientation tracking. Based on the results, Heller et al. recommended the use of head tracking when natural behaviour and quick movement to nearby sound sources is of importance. Device tracking might on the other hand suffice when the focus is on serendipitous discovery or the distances to the sound sources are large.

Romigh et al. (2015) showed that a head-tracked virtual auditory display can provide localization performance comparable with that of real sound sources, if sufficient attention is payed to the different aspects of implementing such a display. In addition to the primary function of providing virtual sound sources with a fixed direction or location in the real world, head tracking provides other benefits, as mentioned earlier: head tracking aids in resolving front-back confusion (Begault et al., 2001) and under some conditions it may improve externalization (Brimijoin et al. (2013) reported improved externalization, while Begault et al. did not).

One important aspect in the implementation of head tracking is the latency between rotation of the listener's head and the corresponding change in the presentation of the virtual sound sources. Brungart et al. (2005) note that excessive latency can lead to a decrease in localization performance as well as a loss of realism, possibly associated with increased fatigue or annoyance. In their experiments, Brungart et al. found that the effects of latency are imperceptible to the average listener if the latency is smaller than 80 ms. In the presence of a co-located low-latency sound source, which may be the case in augmented reality, the latency detection threshold is reduced by approximately 25 ms. Sandvad (1996) and

[2] http://intelligentheadset.com
[3] http://www.bragi.com

Lindau (2009) reported similar thresholds, while higher thresholds have been observed under other conditions (Wenzel, 2001).

2.2.3 Applications

The advent of mobile audio augmented reality technology makes a large range of different applications possible. Albrecht et al. (2011) divide these applications into two categories. The first category is so called pure augmented reality applications, which follow a strict definition of augmented reality. These applications augment the real world with virtual objects that are somehow related to the real world. Adhering to the definition of augmented reality by Azuma (1997), these objects should have a location in the three dimensions of the real world.

The second category contains applications that take advantage of augmented reality technology to some extent, but do not comply with a strict definition of augmented reality. This includes applications presenting virtual objects that are superimposed on our perception of the real world, but are tied to a head-centric rather than a world-centric frame of reference (i.e., these are freely-floating acoustic events, as defined by Härmä et al. (2003)) or lack relation to the surrounding world. In this case, AR technology might only be used as a means to display information without hindering simultaneous perception of the real world. Alternatively, the real world around us can be seen as a suitable medium for presenting virtual objects, rather than limiting these to, e.g., a computer monitor (for visual applications) or locations between the two earpieces of a set of headphones (for audio applications).

Another type of applications that belong to the second category are those that modify our perception of the real world rather than add to it. Such modifications can be achieved through the use of microphone hear-through technology. Storek et al. (2015) call this warped acoustic reality, suggesting the application of effects such as reverberation, pitch shift, and overdrive to the hear-through microphone signals. Other effects can be used to achieve different forms of "super hearing." For example, sound from certain directions may be amplified or attenuated through the use of beamforming (Härmä et al., 2004), or all sounds can simply be amplified.

While the second category of applications does not represent pure augmented reality, the value of an application does not lie in whether it is augmented reality or not. Augmented reality should be seen as a means rather than a goal.

Examples of both categories of applications include guidance, exploration, and navigation (Zimmermann and Lorenz, 2008; Blum et al., 2012; Vazquez-Alvarez et al., 2012, 2016; Publication II; Publication III); games and entertainment (Moustakas et al., 2009; Paterson et al., 2010); and communication (Härmä et al., 2003; Gamper and Lokki, 2010; Publication IV). Audio augmented reality can naturally also be used in combination with visual augmented reality, which leads to an even wider range of possibilities, including medical applications, manufacturing, and repair (Azuma, 1997).

Audio augmented reality applications are closely linked to other types of virtual auditory display applications that present spatial information to the user. Shinn-Cunningham (1998) mentions virtual auditory displays for air traffic controllers and pilots as two examples of such applications. One reason for using auditory displays in these cases is if the visual channel is overloaded. In other cases, there might not even be a possibility to use a visual display, e.g., if the user is blind. As in audio augmented reality, in many of these applications it is important or even absolutely necessary for the user to hear the surrounding acoustic environment in addition to the information presented with the auditory display.

3. Distance Presentation

According to the strict definition of augmented reality by Azuma (1997), audio augmented reality should present virtual sound sources that have a specific location in the real world. To the user, these sound sources appear to have a direction (azimuth and elevation) and a distance. The presentation of sounds in different directions using headphones has been extensively studied. Auditory distance perception has also been studied, albeit to a lesser extent (Kolarik et al., 2015). Previous studies have, however, largely focused on how distances are perceived, rather than on how virtual sound sources can be practically presented so that they are perceived at different distances.

To present virtual sound sources at different distances, one option is to utilize BRIRs measured with different distances between the source and the receiver. In mobile AR applications, this approach would require such measurements to be performed not only at a multitude of different distances, but also in different directions and in different acoustic environments. Another approach is to create the BRIRs through simulations. While interactive room acoustics modelling can be performed in real time on desktop computers and clusters thereof (Pelzer et al., 2014), the heavy processing power requirements make it a poor choice for current mobile devices and applications. Room acoustics modelling can of course also be performed in advance to create a BRIR database representing different distances, directions, and environments.

However, even though accurate simulations and measurements may produce realistic distance cues, this thus not guarantee that the virtual sound sources are perceived at the simulated or measured distances. In fact, the perceived distances of both real and virtual sound sources are typically underestimated when the distance to the sound source is more than a few meters and overestimated when the distance is less than one meter (Zahorik et al., 2005).

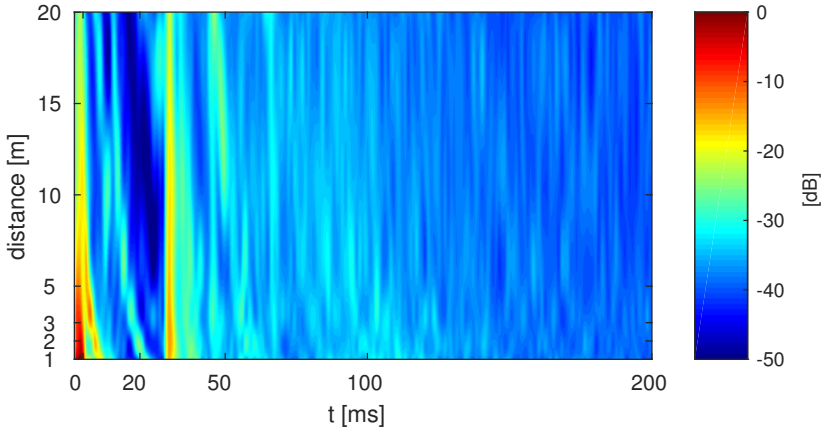


Figure 3.1. Energy of binaural room impulse responses measured at distances of 1, 2, 3, 5, 10, 15, and 20 m, with values interpolated between these distances, shown as a function of time.

Instead of constructing a large BRIR database, it might in many mobile AR applications be more feasible to apply either artificial reverberation or BRIRs measured at a single sound source distance. Appropriate use of such BRIRs both aids in the externalization of sound sources and makes them sound like they originate from the real acoustic environment around the user, rather than being "glued" onto it.

Whether artificial reverberation or measured BRIRs are used, means are needed for modifying the perceived distance of the virtual sound sources presented with these. Prior research presents several different factors that affect the perceived distance of sound sources, including sound intensity, spectrum, binaural cues, and reverberation (Zahorik et al., 2005). The direct-to-reverberant energy ratio (D/R) is often referred to as a measure of the effects that reverberation has on distance perception (Middlebrooks and Green, 1991; Nielsen, 1993; Zahorik et al., 2005; Larsen et al., 2008; Kolarik et al., 2015). The idea behind this notion lies in the fact that, in an enclosed space, late reverberation can be approximated by a diffuse sound field, and the reverberant energy in this type of field is independent of the distance to the sound source. The sound pressure of the direct sound from a point source, on the other hand, is inversely proportional to the distance. Thus, it has been hypothesized that people can utilize the relation between direct and reverberant energy to infer the distance of a sound source.

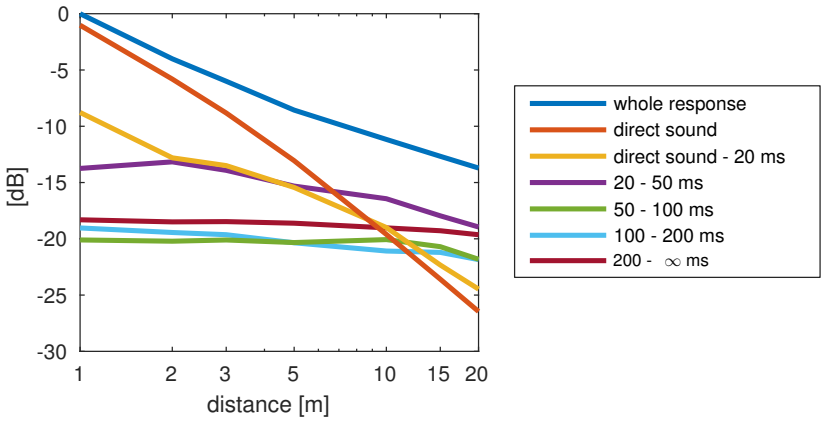


Figure 3.2. Energy of different parts of binaural room impulse responses as a function of measurement distance.

The D/R of a room impulse response $h(t)$ is calculated using the equation

$$D/R = \frac{\int_0^T h^2(t)dt}{\int_T^\infty h^2(t)dt}, \tag{3.1}$$

where T is the dividing point between the direct and the reverberant energy. T is typically 2–3 ms (Zahorik, 2002; Larsen et al., 2008), which means that all early reflections are included in the reverberant energy. However, early reflections do not, like late reverberation, form a diffuse sound field in the enclosing space, but their energy is instead highly dependent on the distance between the source and the receiver, as well as their locations relative to reflective surfaces in the space. Figure 3.1 displays the energy of room impulse responses measured at different distances as a function of time, while Fig. 3.2 displays the energy of different parts of the impulse responses as a function of distance. The impulse responses are from Aula Carolina in the Aachen Impulse Response Database (Jeub et al., 2009, 2010), and the energy of the two channels of each BRIR was combined through binaural loudness summation (Sivonen and Ellermeier, 2006). As can be seen from the figures, the energy of early reflections up to approximately 50 ms is, in this space, dependent on the distance, and these early reflections thus do not constitute a part of the diffuse reverberant field.

Rather than being perceptually integrated with late reverberation, early reflections are known to fuse with the direct sound. Blauert (1997, p. 224) notes that such reflections can make the auditory event louder, increase

its extent in space, and change its timbre. Bradley et al. (2003) demonstrated that increasing the early reflection energy (in this case the first 50 ms of the room impulse response) improves speech intelligibility as much as increasing the energy of the direct sound by the same amount. The importance of early reflections is also known in the field of concert hall acoustics. Lokki et al. (2011) note that early reflections can, due to the precedence effect, perceptually fuse to the direct sound for more than 30 ms after the arrival of the direct sound. An early-to-late energy ratio (E/L) including the first 80 ms of the room impulse response in the early energy is a common measure of the clarity of music, while a dividing point of 50 ms is typically chosen for speech (Soulodre and Bradley, 1995). This measure of clarity is calculated using the equation

$$C_T = \frac{\int_0^T h^2(t)dt}{\int_T^\infty h^2(t)dt}, \tag{3.2}$$

where $h(t)$ is the room impulse response and T is the dividing point between early and late energy. Note that C_T is calculated the same way as the D/R (Eq. (3.1)), only with a different dividing point between early and late energy. Bradley and Soulodre summarize the difference in the spatial impression of concert halls caused by early and late reflections:

> As Haas and others have shown, sound arriving shortly after the direct sound is integrated or temporally and spatially fused with the direct sound. Thus increasing levels of early lateral reflections increase the apparent level of the direct sound and cause a slight ambiguity in its perceived location. These two effects contribute to the resulting increase in [apparent source width]. Later arriving sound is not integrated or temporally and spatially fused with the direct sound, and leads to more spatially distributed effects that appear to envelop the listener. (Bradley and Soulodre, 1995)

Since early reflections are perceptually fused with the direct sound, rather than with the late reverberation, one might ask if this type of integration also is relevant when it comes to how reverberation affects our perception of distance. Figure 3.3 shows different early-to-late energy ratios, including the D/R, as well as the total energy of the Aula Carolina room impulse responses as a function of distance (again with binaural loudness summation applied). These all depend on the distance to the sound source and could therefore all potentially be used as cues for

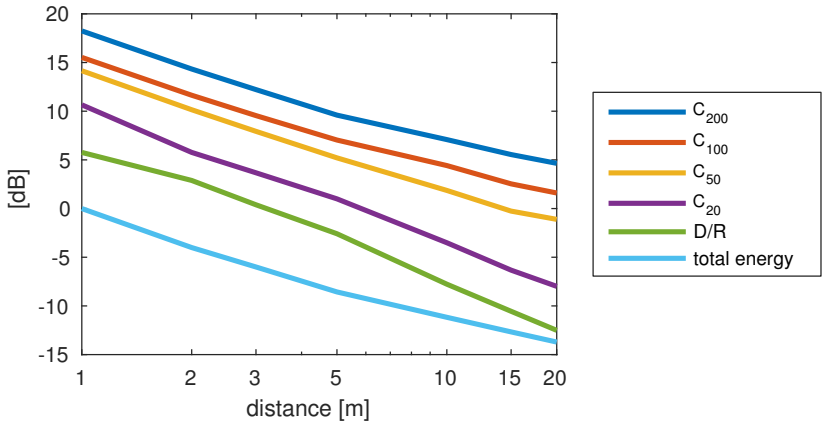


Figure 3.3. Early-to-late energy ratios and total energy of binaural room impulse responses as a function of distance.

estimating the distance. However, little research has been done to investigate this detail. Bronkhorst and Houtgast (1999) asked this question and found that an early-to-late energy ratio (that they called a modified direct-to-reverberant energy ratio) including the first 6 ms of the room impulse response in the early energy best explained the results of their experiments. Thus, in their model only the very first reflections are integrated with the direct sound. While these results are interesting, they may not be applicable to conditions differing from those of the experiments. In particular, only small spaces were studied (with only 1, 3, 9, 27, 81, or 800 simulated reflections included, depending on the condition), only noise bursts were used as stimuli, and intensity differences were eliminated as a possible distance cue.

Publication I studies the effect reverberation has on distance perception, providing practical methods for modifying the perceived distance of virtual sound sources, as summarized in Section 3.1. In addition to these, some augmented reality applications might need the means to present a wide range of different distances. For example, an application might present different points of interest (POIs) surrounding a user walking around in a city. However, if the distances to the POIs are large, say hundreds of meters, these cannot be presented with naturalistic auditory distance cues, since most sound sources would not be audible at such distances, especially in a noisy urban environment. Even if such cues could be utilized, listeners tend to underestimate the distances to far-away sound sources (Zahorik et al., 2005; Fluitt et al., 2014; Kolarik et al., 2015).

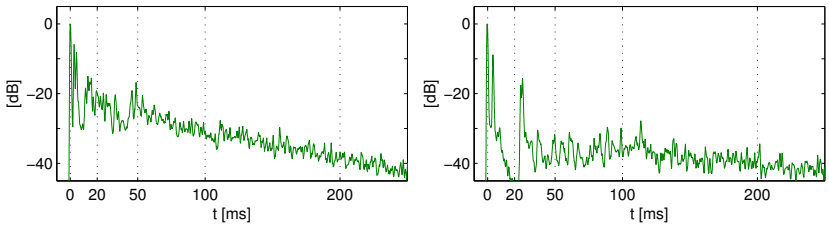
There are different options available if we need to present distant POIs using sound. One obvious approach is to simply present the distance to the sound source in, e.g., meters using speech. If speech cannot be used, or the length of the sound should be limited, different acoustic parameters of the presented sound can instead be used for conveying distance information. These parameters may or may not be based on auditory distance cues. Liljedahl and Lindberg (2011) studied modifications of pitch, reverberation (D/R), low-pass filtering, and combinations of these, and found that low-pass filtering and reverberation produced the most reliable distance cues. Liljedahl and Lindberg also recommend using multiple distance cues simultaneously.

Talbot and Cowan (2009) tested pitch encoding, beat rate encoding, and so called ecological distance encoding (utilizing intensity and air absorption, i.e., low-pass filtering cues) with blind participants. Based on the results, they suggest the use of ecological distance encoding and as a second alternative beat rate encoding. Talbot and Cowan, as well as Liljedahl and Lindberg, found pitch to be an unreliable distance cue, as some participants intuitively associated higher pitch with shorter distances, while others associated it with longer distances. Distance information can also be given using separate sounds, e.g. earcons, spearcons, or short pulses (Hussain et al., 2014). Such an approach does, however, require training for users to know the mapping between the different sounds and distances.

The studies of Talbot and Cowan, and Liljedahl and Lindberg suggest using auditory distance cues rather than other acoustic parameters for conveying distance. Since auditory distance cues associated with a sound naturally produce an auditory event that is perceived at a certain distance, which usually is underestimated compared with the actual distance of a corresponding real sound source, the question remains if and how we can present larger distances using these cues. Publication II, summarized in Section 3.2, studies how people can learn to map available auditory distance cues, which naturally produce relatively short distance percepts, to longer distances in the real world.

3.1 Modification of binaural room impulse responses

Publication I investigates how reverberation affects the perceived distance of sound sources. However, the focus is not on providing insight



(a) Space I (b) Space II

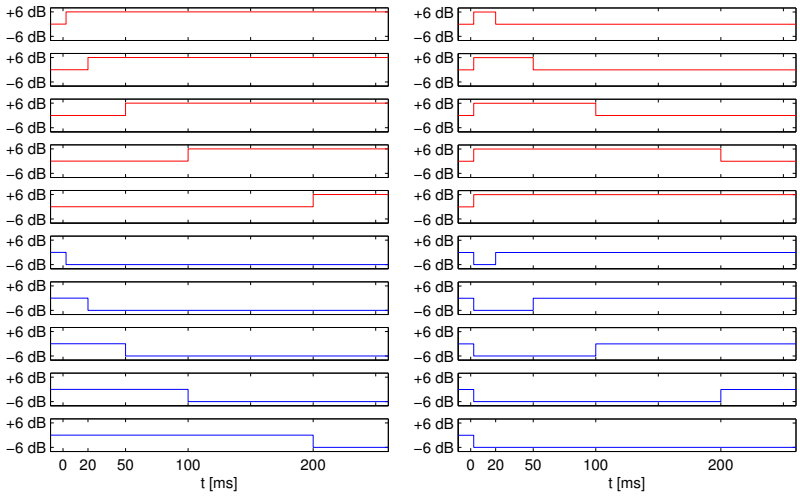
Figure 3.4. Energy envelopes of the binaural room impulse responses. The responses of the right (ipsilateral) ear are shown.

into the mechanisms with which the human auditory system interprets reverberation cues. Instead, the study approaches the problem from a practical point of view: how can reverberation be modified in a practical application to affect the perceived distance effectively? As many real-time applications might utilize either measured or artificial binaural room impulse responses, the effect of modifying the temporal envelope of these is examined.

3.1.1 Methods

In the user study reported in Publication I, participants were, using headphones, presented with anechoic speech samples convolved with one of two BRIRs from the Aachen Impulse Response Database. The BRIRs, depicted in Fig. 3.4, represented a moderately reverberant and a highly reverberant space. Impulse responses measured with a source azimuth of 90° were chosen, since impulse responses measured at an azimuth of 0° have been shown to provide poor externalization (Begault and Wenzel, 1993; Catic et al., 2015). The measurement distance was 3 meters. Modifications were made to these BRIRs, amplifying or attenuating different portions of the temporal envelopes. The two sets of modifications performed are illustrated in Fig. 3.5.

In the listening tests, 24 participants were asked to judge the relative perceived distances of the virtual speech sources in different samples, with the interface shown in Fig. 3.6. Each condition was associated with one of the two BRIRs, a speech sample spoken by either a male or a female voice, and one of the two modification sets, resulting in a total of eight conditions. In addition to the modifications illustrated in Fig. 3.5,



(a) Modification set A (b) Modification set B

Figure 3.5. Modifications made to the binaural room impulse responses. Modifications starting immediately after the direct sound (at $t = t_d = 2.4$ ms) included the earliest reflections.

each condition also included the unmodified BRIR, the whole BRIR amplified 6 dB, and the whole BRIR attenuated 6 dB, resulting in a total of 13 samples per condition.

Estimation of relative distances was chosen in favour of absolute distance estimation. If asked to estimate the perceived distance in meters, participants might try to actively deduce where the (virtual) sound source is, rather than rely solely on the perceived location of the auditory event. In auditory distance perception research, it is generally the distance of the auditory event that is of interest, rather than the estimated distance of the sound source (Blauert, 1997, p. 46, 116–117, 120–121, 123–124). Nevertheless, in some other cases, as in Publication II, a distance estimate based not only on the distance of the auditory event, but also on other factors, is called for.

3.1.2 Results

Figure 3.7 displays the medians of the distance estimates for modification set A with the female speech sample, together with their 95% bootstrap confidence intervals. Before calculating the medians, Z-scores were calculated separately for each participant's answers under each condition. This was done to compensate for the differing use of the answering scale among participants. The results for the male speech samples are not shown here,

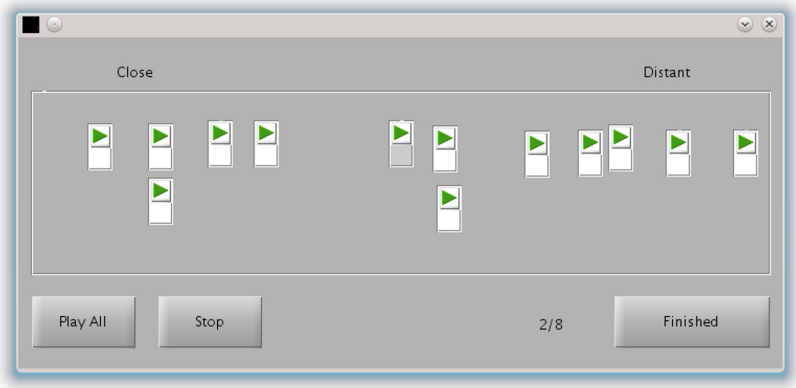
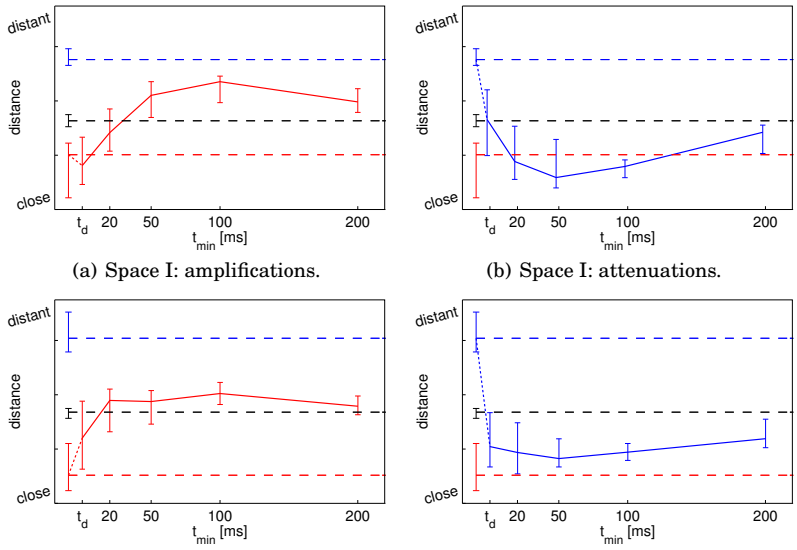


Figure 3.6. A condensed depiction of the interface used in the listening tests. Participants could listen to individual samples and move these around so that their locations on the horizontal axis represented their relative perceived distances.

as they largely follow the same pattern as those for the female speech samples.

The presentation of the distance estimates in Fig. 3.7 differs slightly from the one in Publication I. Here, amplifications and attenuations are shown in different figures. Amplification refers to samples where parts of the BRIRs are amplified compared with the unmodified reference sample (see Fig. 3.5). Correspondingly, attenuation refers to samples with parts of the BRIRs attenuated.

The E/L modifications shown in Fig. 3.5(a) amplify or attenuate the late energy. E/L modifications are, however, typically done by amplifying or attenuating the early energy, which is the case when the distance to a sound source changes in the real world. To emphasize this relationship, the samples with E/L modifications are in Fig. 3.7 connected to the appropriate sample where the whole impulse response is amplified or attenuated, which can be seen as a natural starting point for the displayed set of modifications. For example, in Fig. 3.7(a) the first red sample from the left has no E/L modifications. Compared with this sample, the next sample ($t_{min} = t_d$) represents a 6 dB attenuation of the direct sound. The next sample ($t_{min} = 20$ ms) represents a 6 dB attenuation of the first 20 ms of the impulse response. The next sample ($t_{min} = 50$ ms) represents a 6 dB attenuation of the first 50 ms of the impulse response, and so on. These samples thus represent a series of E/L modifications, where the early energy is attenuated while the late reverberant energy remains constant. The modifications in Fig. 3.7(a), as well as Fig. 3.7(c), could



(a) Space I: amplifications.

(b) Space I: attenuations.

(c) Space II: amplifications.

(d) Space II: attenuations.

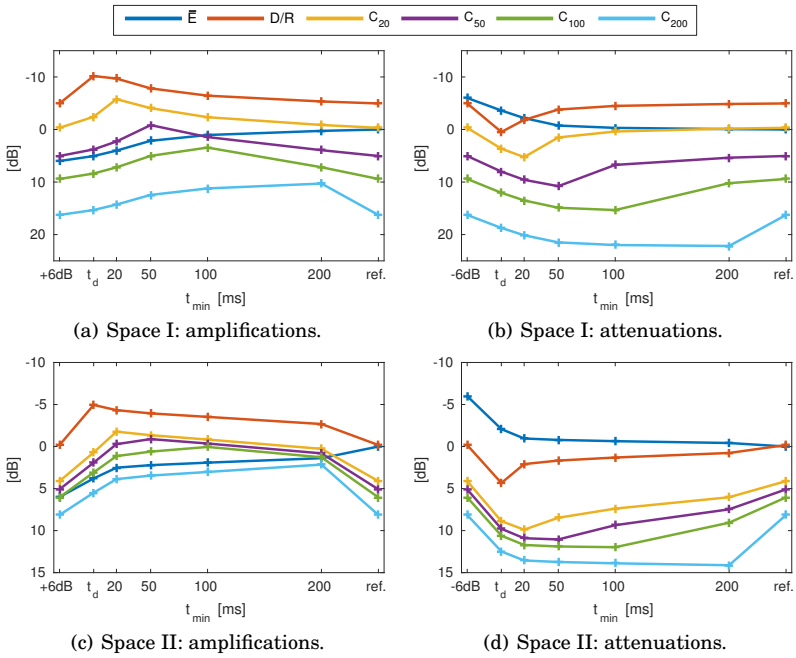
Figure 3.7. Distance estimates for modification set A with the female speech sample. The medians and their 95% confidence intervals are shown. The dashed line in the middle represents the unmodified sample. The lower dashed line represents the 6 dB amplification of the whole sample, while the upper dashed line represents the 6 dB attenuation of the whole sample. The other data points are positioned on the horizontal axis at the time when the attenuation or amplification of the impulse response begins.

Table 3.1. Differences between the perceived distances of the amplifications in modification set A. p-values of a two-sided sign test are shown for female speech samples in space I.

	+6dB	$\uparrow t_d\rightarrow$	↑20ms→	↑50ms→	↑100ms→	↑200ms→	ref.
$\uparrow t_d\rightarrow$	**0.023**						
↑20ms→	**<0.001**	**<0.001**					
↑50ms→	**<0.001**	**<0.001**	**0.002**				
↑100ms→	**<0.001**	**<0.001**	**<0.001**	0.152			
↑200ms→	**<0.001**	**<0.001**	**0.007**	0.541	0.839		
ref.	**0.002**	**0.002**	0.307	**0.023**	**0.003**	**<0.001**	
-6dB	**<0.001**	**<0.001**	**0.002**	**0.002**	**0.023**	**<0.001**	**<0.001**

Table 3.2. Differences between the perceived distances of the attenuations in modification set A. p-values of a two-sided sign test are shown for female speech samples in space I.

	-6dB	$\downarrow t_d\rightarrow$	↓20ms→	↓50ms→	↓100ms→	↓200ms→	ref.
$\downarrow t_d\rightarrow$	**<0.001**						
↓20ms→	**<0.001**	**0.007**					
↓50ms→	**<0.001**	0.064	0.307				
↓100ms→	**<0.001**	**0.023**	**0.023**	0.541			
↓200ms→	**<0.001**	0.541	0.541	0.307	**0.007**		
ref.	**<0.001**	1.000	0.064	**0.007**	**0.023**	0.152	
+6dB	**<0.001**	0.152	0.839	0.541	0.541	0.541	**0.002**



(a) Space I: amplifications.

(b) Space I: attenuations.

(c) Space II: amplifications.

(d) Space II: attenuations.

Figure 3.8. Early-to-late energy ratios and energy for the different female speech samples with modification set A. The values of the energy $\bar{E}$ are relative to the energy of the unmodified sample. The horizontal axis specifies the time when the attenuation or amplification of the impulse response begins. Additionally, the unmodified sample (ref.), the wholly amplified sample (+6dB), and the wholly attenuated sample (-6dB) are shown on the horizontal axis. The vertical axis is inverted to emphasize the potential effects of the energy ratios and energy on the distance estimates in Fig. 3.7.

Table 3.3. Differences between the perceived distances of the amplifications in modification set A. p-values of a two-sided sign test are shown for female speech samples in space II.

	+6dB	$\uparrow t_d \rightarrow$	↑20ms→	↑50ms→	↑100ms→	↑200ms→	ref.
$\uparrow t_d \rightarrow$	**<0.001**						
↑20ms→	**<0.001**	**0.007**					
↑50ms→	**<0.001**	**0.002**	0.541				
↑100ms→	**<0.001**	**0.002**	0.210	0.210			
↑200ms→	**<0.001**	0.064	0.839	0.839	0.541		
ref.	**<0.001**	0.152	0.541	0.541	**0.035**	0.839	
-6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 3.4. Differences between the perceived distances of the attenuations in modification set A. p-values of a two-sided sign test are shown for female speech samples in space II.

	-6dB	$\downarrow t_d \rightarrow$	↓20ms→	↓50ms→	↓100ms→	↓200ms→	ref.
$\downarrow t_d \rightarrow$	**<0.001**						
↓20ms→	**<0.001**	1.000					
↓50ms→	**<0.001**	0.152	0.839				
↓100ms→	**<0.001**	0.678	0.307	0.832			
↓200ms→	**<0.001**	0.839	0.064	0.152	0.307		
ref.	**<0.001**	0.064	**0.023**	**<0.001**	**0.002**	**0.007**	
+6dB	**<0.001**	0.541	0.307	0.541	0.541	0.541	**<0.001**

thus be referred to as attenuations, and the modifications in Fig. 3.7(b), as well as Fig. 3.7(d), could similarly be referred to as amplifications, but the same convention as in Publication I was here chosen for the sake of consistency.

A two-sided sign test was used to check for significant differences between the distance estimates of the different samples. The p-values of these are shown in Tables 3.1, 3.2, 3.3, and 3.4 (not included in Publication I). Note that appropriate corrective procedures, e.g., the Holm–Bonferroni method (Holm, 1979), must be applied to control the family-wise error rate when drawing conclusions based on these p-values.

Fig. 3.8 shows different early-to-late energy ratios and the total energy of the different samples. These are analyzed under Section 3.1.3, but are presented here to allow easy comparison with the corresponding distance estimates in Fig. 3.7.

The results for space I (Figs. 3.7(a) and 3.7(b)) show that E/L modifications including approximately the first 50–100 ms of the impulse response in the early energy produce the largest changes in the perceived distance. Looking at the amplifications (Table 3.1) and comparing the conventional D/R modification ($t_{min} = t_d$) with the other E/L modifications ($t_{min} =$ 20, 50, 100, and 200 ms) confirms that all the E/L modifications produce a significantly larger change in the perceived distance than the D/R modification does (at a significance level of 0.05, with the Holm-Bonferroni correction applied to the results of these four sign tests). For the attenuations (Table 3.2), the E/L modification with $t_{min} =$ 20 ms produces a significantly larger change in the perceived distance than the D/R modification does, but the differences between the D/R modification and the other E/L modifications ($t_{min} =$ 50, 100, and 200 ms) are not statistically significant (again with the Holm-Bonferroni correction applied).

For space II, the results (Figs. 3.7(c) and 3.7(d), and Tables 3.3 and 3.4) are less clear. For amplifications, E/L modifications with the first 20, 50, or 100 ms of the impulse response included in the early energy produce larger perceived distances than the D/R modification (Holm-Bonferroni correction applied). For attenuations, there was no significant difference between the D/R modification and any of the E/L modifications (Holm-Bonferroni correction applied).

The distance estimates of space I with modification set B are shown and analyzed in Figs. 3.9(a) and 3.9(b), and Tables 3.5 and 3.6. Visual inspection of the results shows that these modifications exhibit the same

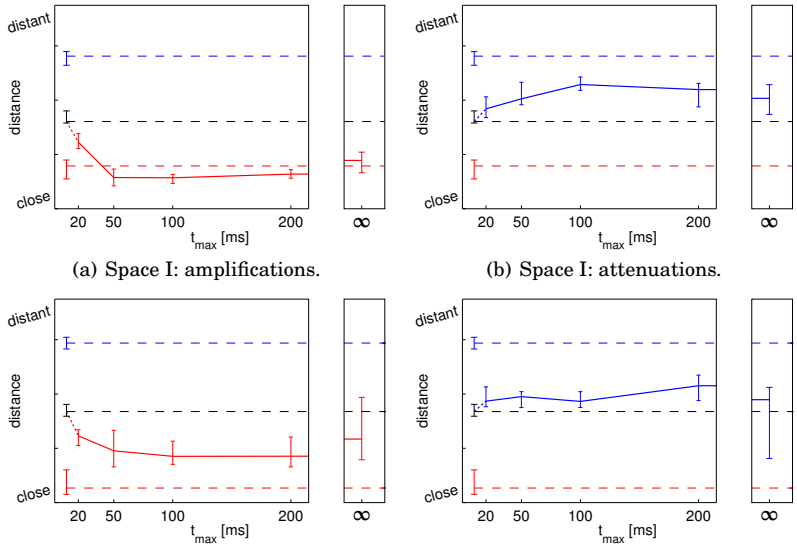
behaviour as those of modification set A: amplification of early reflections up to approximately 50–100 ms decreases the perceived distance, while attenuation of these reflections increases the perceived distance. Sign tests confirm a significant difference between the reference sample and amplifications from t_d to 20, 50, 100, 200, and ∞ ms (Holm-Bonferroni correction applied), as well as attenuations from t_d to 20, 50, 100, and 200, but not ∞ ms (Holm-Bonferroni correction applied). However, these modifications do not produce as large changes in the perceived distance as some of the modifications in set A, where the direct sound and early reflections were amplified or attenuated simultaneously.

The distances estimates of space II with modification set B are shown and analyzed in Figs. 3.9(c) and 3.9(d), and Tables 3.7 and 3.8. These distance estimates display, to some extent, the same behaviour as those of space I, but the observed changes are not as large as those for space I. Amplifying early reflections decreases the perceived distance (amplifications from t_d to 20, 50, 100, and 200, but not ∞ ms significantly decrease the perceived distance, with the Holm-Bonferroni correction applied), but attenuating these reflections has little effect on the perceived distance (no significant differences can be observed when comparing the reference sample with the attenuations from t_d to 20, 50, 100, 200, and ∞ ms, with the Holm-Bonferroni correction applied).

Fig. 3.10 shows different early-to-late energy ratios and the total energy of the different samples with modification set B, corresponding to the distance estimates displayed in Fig. 3.9. These are analyzed in Section 3.1.3.

3.1.3 Further analysis

The listening test results raise the question how much of the change in the perceived distance is due to the change in the balance between direct, early, and late energy and how much is due to the change in intensity. The results show that the amplification of early reflections reduces the perceived distance while attenuation of early reflections increases the perceived distance. Since this effect partially can be explained by the increase or decrease in intensity caused by the increase or decrease of early energy, it is not entirely clear what effect the balance between direct, early, and late energy has. In particular, if we want to explain this effect using an early-to-late energy ratio, how much of the impulse response should be included in the early energy: only the direct sound or up to 20, 50, 100, or even 200 ms?



(a) Space I: amplifications.

(b) Space I: attenuations.

(c) Space II: amplifications.

(d) Space II: attenuations.

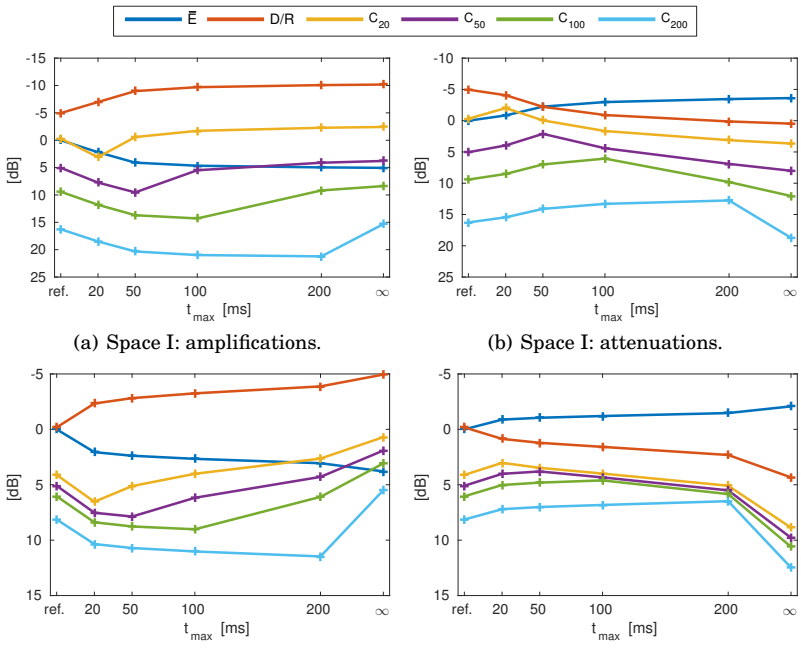
Figure 3.9. Distance estimates for modification set B with the female speech sample. The medians and their 95% confidence intervals are shown. The dashed line in the middle represents the unmodified sample. The lower dashed line represents the 6 dB amplification of the whole sample, while the upper dashed line represents the 6 dB attenuation of the whole sample. The other data points are positioned on the horizontal axis at the time when the attenuation or amplification of the impulse response ends.

Table 3.5. Differences between the perceived distances of the amplifications in modification set B. p-values of a two-sided sign test are shown for female speech samples in space I.

	ref.	$\uparrow t_d$–20ms	$\uparrow t_d$–50ms	$\uparrow t_d$–100ms	$\uparrow t_d$–200ms	$\uparrow t_d$–∞ms	+6dB
$\uparrow t_d$–20ms	**<0.001**						
$\uparrow t_d$–50ms	**<0.001**	**0.002**					
$\uparrow t_d$–100ms	**<0.001**	**<0.001**	0.678				
$\uparrow t_d$–200ms	**<0.001**	**0.011**	0.307	0.210			
$\uparrow t_d$–∞ms	**<0.001**	0.152	**0.023**	**<0.001**	**0.017**		
+6dB	**<0.001**	**0.023**	0.307	**0.023**	0.307	0.678	
-6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 3.6. Differences between the perceived distances of the attenuations in modification set B. p-values of a two-sided sign test are shown for female speech samples in space I.

	ref.	$\downarrow t_d$–20ms	$\downarrow t_d$–50ms	$\downarrow t_d$–100ms	$\downarrow t_d$–200ms	$\downarrow t_d$–∞ms	-6dB
$\downarrow t_d$–20ms	**0.007**						
$\downarrow t_d$–50ms	**<0.001**	**0.002**					
$\downarrow t_d$–100ms	**<0.001**	**0.002**	**0.035**				
$\downarrow t_d$–200ms	**0.023**	0.307	0.678	0.307			
$\downarrow t_d$–∞ms	0.064	0.541	0.839	0.064	0.307		
-6dB	**<0.001**	**<0.001**	**<0.001**	**0.002**	**<0.001**	**<0.001**	
+6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.002**	**<0.001**



(a) Space I: amplifications.

(b) Space I: attenuations.

(c) Space II: amplifications.

(d) Space II: attenuations.

Figure 3.10. Early-to-late energy ratios and energy for the different female speech samples with modification set B. The values of the energy $\bar{E}$ are relative to the energy of the unmodified sample. The horizontal axis specifies the time when the attenuation or amplification of the impulse response ends. Additionally, the unmodified sample (ref.) is shown on the horizontal axis. The vertical axis is inverted to emphasize the potential effects of the energy ratios and energy on the distance estimates in Fig. 3.9.

Table 3.7. Differences between the perceived distances of the amplifications in modification set B. p-values of a two-sided sign test are shown for female speech samples in space II.

	ref.	$\uparrow t_d$–20ms	$\uparrow t_d$–50ms	$\uparrow t_d$–100ms	$\uparrow t_d$–200ms	$\uparrow t_d$–∞ms	+6dB
$\uparrow t_d$–20ms	**<0.001**						
$\uparrow t_d$–50ms	**<0.001**	0.307					
$\uparrow t_d$–100ms	**<0.001**	0.152	0.541				
$\uparrow t_d$–200ms	**<0.001**	0.093	1.000	1.000			
$\uparrow t_d$–∞ms	0.152	1.000	0.541	0.405	1.000		
+6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	
-6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 3.8. Differences between the perceived distances of the attenuations in modification set B. p-values of a two-sided sign test are shown for female speech samples in space II.

	ref.	$\downarrow t_d$–20ms	$\downarrow t_d$–50ms	$\downarrow t_d$–100ms	$\downarrow t_d$–200ms	$\downarrow t_d$–∞ms	-6dB
$\downarrow t_d$–20ms	0.307						
$\downarrow t_d$–50ms	0.093	1.000					
$\downarrow t_d$–100ms	**0.023**	1.000	1.000				
$\downarrow t_d$–200ms	0.093	0.307	0.541	0.307			
$\downarrow t_d$–∞ms	0.839	0.307	0.541	0.678	0.064		
-6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	
+6dB	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	0.064	**<0.001**

In Publication I, using a least-squares approach, models of the influence of intensity and different early-to-late energy ratios on the perceived distance were fitted to the results, both separately and combined:

$$d_C = a \cdot \left(\frac{1}{C_T}\right)^k, \tag{3.3}$$

$$d_E = b \cdot \left(\frac{1}{\bar{E}}\right)^l, \tag{3.4}$$

and

$$d = c \cdot \left(\frac{1}{C_T}\right)^k \cdot \left(\frac{1}{\bar{E}}\right)^l. \tag{3.5}$$

Here, C_T is the early-to-late energy ratio with the dividing point T, $\bar{E}$ is the time-averaged energy, and a, b, c, k, and l are constants. Since the participants only made relative distance estimates without any absolute reference, the origin assigned to the distance estimates in the analysis was arbitrary. Thus, an additive constant would have been motivated in the models, but since it did not noticeably affect the results, it was in the end left out for the sake of simplicity. The results of the fitting indicated that intensity and E/L both contribute approximately to the same extent to the distance estimates of modification set A. However, with modification set B, intensity dominates. The fitting for space I displays a minimum error when approximately the first 100 ms (i.e., $T = 100$ ms) of the impulse response are included in the early energy of the early-to-late energy ratio (C_T). For space II, no clear minimum can be seen. Note that there was an error in Publication I in the fitting of modification set B. The error is explained and the corrected fit shown in the Errata on page 109.

While the fitting of the distance models gave some insight into how different early-to-late energy ratios could explain the results together with intensity, a more illustrative approach is here taken to examine these effects. Figure 3.8 shows different early-to-late energy ratios (including the D/R) and the total energy of the different samples in modification set A. These are shown in decibels, meaning that the multiplication of $\frac{1}{C_T}$ and $\frac{1}{\bar{E}}$ in Eq. (3.5) corresponds to addition of the appropriate E/L curve (D/R, C_{20}, C_{50}, C_{100}, or C_{200}) and the energy curve ($\bar{E}$), and the multiplicative constant c corresponds to an additive constant. Taking the logarithm of the whole equation gives us:

$$10\log_{10}(d) = 10\log_{10}(c) - k \cdot 10\log_{10}(C_T) - l \cdot 10\log_{10}(\bar{E}). \tag{3.6}$$

Although the distance estimates in Fig. 3.7 are not presented as logarithms, the shape of the curves would not change considerably if this was

done, since the range of perceived distances was small in this experiment (see Publication I, Fig. 9).

Thus, we can directly compare the contribution of energy and different early-to-late energy ratios in Fig. 3.8 with the distance estimates in Fig. 3.7 without introducing much error compared with Eq. (3.6). While the energy itself can partially explain the results for modification set A in space I (Figs. 3.7(a) and 3.7(b)), it does not account for the influence of late reverberation. Combining the effects of the energy and the D/R does not explain the results either. These effects would predict a large change in the distance estimates caused by D/R modifications ($t_{min} = t_d$), but the results do not display this type of change. Instead, a combination of energy and an early-to-late energy ratio including approximately the first 50–100 ms in the early energy results in the best match with the medians of the distance estimates.

As with space I, the energy does not alone or combined with the D/R explain the results for modification set A in space II (Figs. 3.7(c) and 3.7(d)). The best match to the distance estimates is also here provided by a combination of energy and an early-to-late energy ratio including approximately the first 50–100 ms in the early energy.

When comparing the energy and early-to-late energy ratios of the samples of modification set B (Fig. 3.10) with the corresponding perceived distances (Fig. 3.9), it can be seen that the energy alone could explain most of the results, but again not the effect of amplified or attenuated late reverberation. The D/R does not explain the results, but would produce distance estimates in the opposite direction to those observed, negating the effect of the total energy. A combined effect of energy and an early-to-late energy ratio including approximately 100 ms in the early energy provides the best explanation for the perceived distances.

3.1.4 Discussion

In the results of Publication I, a clear difference can be seen between the two investigated spaces (see Fig. 3.7, and Tables 3.1, 3.2, 3.3, and 3.4). In space I, E/L modifications with approximately 50–100 ms included in the early energy produce the largest changes in the perceived distance. In space II, E/L modifications including 20–200 ms in the early energy all produce perceived distances that do not differ significantly. A possible explanation for the difference between the two spaces can be found by examining the impulse responses in Fig. 3.4. Space I is the smaller

of the two spaces, with a large number of early reflections and thus a large amount of early reflection energy in the first 100 ms of the response. Space II, on the other hand, is a large space, where early reflections are sparse and the energy of these is relatively small compared with the direct energy and the late energy. Thus, amplifying or attenuating different parts of the early reflections has a much smaller effect on the early-to-late energy ratios in space II than in space I (as can be seen by comparing Figs. 3.8(c) and 3.8(d) with Figs. 3.8(a) and 3.8(b)) and, consequently, on the perceived distance.

The results suggest that an early-to-late energy ratio containing even up to 100 ms of early reflections in the early energy is a better estimate of how reverberation affects human distance perception than the direct-to-reverberant energy ratio. Of course, the study has its limitations. These include the choice of BRIRs, the choice of source direction, and the choice of modifications to the BRIRs. Most importantly, the experiment utilized only speech samples, which means that the results are directly applicable only to human speech. In their experiments done with noise bursts, Bronkhorst and Houtgast (1999) found that an early-to-late energy ratio (or modified direct-to-reverberant ratio, as they called it) including 6 ms in the early energy best explained the results. The differences between the results of Publication I and those of Bronkhorst and Houtgast might be related to the different types of stimuli used, but they might also be related to other differences between the two studies.

One difference lies in the binaural room impulse responses used. Bronkhorst and Houtgast used simulated BRIRs consisting of up to 800 reflections. The simulated room and source distance, the number of reflections included (the 1, 3, 9, 27, 81, or 800 strongest reflections), and the relative level of the reflections were varied. The maximum duration of these BRIRs was 110 ms and the BRIRs including only a few reflections were presumably much shorter. Thus, they did not represent natural simulations of real rooms, since they contained limited or no late reverberation. The samples with BRIRs including only 1 or 3 reflections were often perceived inside the head, and were in these cases given a perceived distance rating of 0 m. One might speculate that such distance judgements should be removed from the results, since the lack of externalization could be considered a failure to produce any perception of distance, rather than producing a zero distance. Kolarik et al. (2015) point out that external-

ization and distance perception are related but distinct phenomena and add that externalization is a prerequisite for distance perception. On the other hand, there seems to be a continuum between in-the-head localization and externalization (Hartmann and Wittenberg, 1996), which makes it difficult to distinguish between fully externalized and poorly externalized stimuli. In addition to the small number of simulated reflections in the experiments of Bronkhorst and Houtgast, the choice of a source azimuth of 0° might also have decreased the amount of successfully externalized stimuli (Begault and Wenzel, 1993; Catic et al., 2015).

While Bronkhorst and Houtgast in 1999 suggested that a modified direct-to-reverberant ratio explains how humans translate reverberation cues into distance percepts in rooms, Bronkhorst (2002) performed experiments that could not be well explained by this modified direct-to-reverberant ratio. Instead, Bronkhorst suggested a new model, in which the direct sound was calculated by including all reflections within an interaural time difference window of ± 36 µs from the direct sound. Correspondingly, all reflections outside this ITD window were deemed reverberant energy. This model could explain the listening test results of both Bronkhorst and Houtgast (1999) and Bronkhorst (2002). However, it is unlikely that such a model can explain the results of Publication I. Since the BRIRs used here were measured at a source azimuth of 90° and an elevation of 0°, the ITD window would only include reflections having both approximately the same azimuth and approximately the same elevation as the direct sound. For example, the first floor and ceiling reflections would not be included, as would be the case if the source azimuth was 0°. The ITD-based distance model is thus sensitive to changes in the source direction with respect to the listener, as well as the listener's orientation with respect to the surrounding reflective surfaces.

It is, however, also clear that an early-to-late energy ratio with up to 100 ms of early energy, which explains the results of Publication I, cannot explain the results of Bronkhorst and Houtgast (1999), where the maximum BRIR duration was 110 ms. For these results, a dividing point of 50–100 ms between early and late energy would result in an early-to-late energy ratio that is close to infinite or even infinite. This raises the question whether the dividing point might depend on the characteristics of the room (or the simulated impulse response), and how the human auditory system processes the different reflections in different rooms. In the study

by Bronkhorst and Houtgast, the simulated rooms were small (and the lengths of the impulse responses even further limited), while in Publication I the rooms were considerably larger.

Like the effect of reverberation, the mechanisms with which stimulus intensity affects auditory distance perception are not thoroughly understood. Naturally, humans do not perceive intensity as such, but the sensation of loudness depends also on spectral, temporal and spatial aspects of the sound (Skovenborg and Nielsen, 2004). Also here reverberation plays a role. If reflections are absent, the equation is much simpler, but in a reverberant space, what does the human auditory system use as a cue for distance: the intensity of the direct sound, the direct sound and early reflections, or also the late reverberation? In Publication I, the intensity calculation included both early reflections and late reverberation. However, it might be that the intensity cue contains only the direct sound and the early reflections, since early reflections are perceptually integrated with the direct sound, which is not the case for late reverberation.

The results of the listening test in Publication I do not seem to support this notion. As an example, look at the modifications with $t_{min} = 100$ ms in space II (Figs. 3.7(c) and 3.7(d)). Both the amplification and the attenuation produce a 6 dB change in C_{100}, in opposite directions. However, the amplification (Fig. 3.7(c)) produces a much smaller change in the perceived distance, with respect to the unmodified reference sample, than the attenuation (Fig. 3.7(d)) does. Assuming both that C_{100} explains the effect that reverberation has on the perceived distance and that this effect is the same for changes in C_{100} in both directions, which of course might not be true, an explanation to the observed asymmetry could be that energy in the late reverberation ($t > 100$ ms) is included in the intensity cue for distance perception. If we look at Figs. 3.8(c) and 3.8(d), we can see that amplifying the late reverberation ($t_{min} = 100$ ms) has a larger effect on the total energy, with respect to the unmodified reference sample, than the same attenuation has. This could thus explain the perceived asymmetry in Figs. 3.7(c) and 3.7(d). If the intensity cue instead included only the early energy, both the investigated amplification and attenuation would produce identical intensity cues and would thus not explain the observed asymmetry. Note, that the same type of asymmetry can be observed also for modifications other than those with $t_{min} = 100$ ms and also for distance estimates in space I (Figs. 3.7(a) and 3.7(b)). Clearly, this is a topic for further investigation.

Even though the exact mechanisms are unclear, it is evident from the results of Publication I that for controlling the perceived distance of virtual speech sources in moderately reverberant and highly reverberant spaces, E/L modifications with the first 50–100 ms of the impulse response included in the early energy are to be preferred over traditional D/R modifications. While D/R modifications have some effect, especially the effect of attenuating the direct sound in the presence of strong early reflections is small and only limited attenuation can be performed until the energy of the direct sound is negligible compared to that of the early reflections. The results may also be useful when choosing the most appropriate BRIR modifications for affecting the perceived distance while taking other criteria, such as the dynamic range or speech intelligibility (Bradley et al., 2003), into account.

3.2 Distance presentation in outdoor augmented reality

Publication II studies the presentation of distance in audio augmented reality. The aim of the study was to investigate and compare methods for presenting the distance of points of interest in an urban outdoor environment using audio. For example, a user might want to find a place to eat nearby. Using virtual auditory display techniques, a voice may present different restaurants in different directions around the user. Obviously, the voice could mention the distance to the restaurants, but the messages could be shorter if the user instead could infer the distance based on the perceived distance of the voice.

The problem with an approach based on the perceived distance of the voice is that human speech normally cannot be heard or understood at distances larger than a few dozens of meters, especially in a noisy urban environment. In Publication II, several different cues to auditory distance perception were used together to present POIs. The cues were modified to present the POIs at different distances. These modifications were done within limits that kept the presented messages intelligible. User study participants were asked to localize the POIs presented this way, and the precision and accuracy were compared with presenting the distance with a voice saying the distance in meters. Since the hypothesis was that the auditory distance cues would produce relatively short distance estimates, the experiment was performed again after training with visual feedback

to see if the participants could learn to map the available cues to the intended POI locations.

3.2.1 Methods

The user study presented in Publication II was designed to provide answers to the following main research questions:

- If we present virtual sound sources representing POIs in an urban audio augmented reality environment, where in the environment do people locate these?
- If the virtual sound sources are localized at distances shorter than those we want to present, can we easily teach people to map these perceived distances to longer distances?
- How does providing auditory distance cues compare with the alternative of saying the distance in meters?

The following secondary research questions were also considered:

- Does localization performance improve with repeated presentation of sounds compared with a single presentation?
- How does localization of completely artificial sounds (noise bursts) compare with speech?
- Can we provide additional cues that certain POIs are obstructed by buildings?

In the user study, twelve participants sat in an urban outdoor environment (see Fig. 3.11). Using headphones and head orientation tracking, they were presented with sounds representing different POIs around them. Open headphones were chosen to allow participants to hear the environment clearly. The participants were asked to choose the location on a zoomable map where they thought that each POI was located. Participants were told that the POIs were imaginary, i.e., they might or might not coincide with any real POIs, but that they always were located in some part of the buildings in the area. The participants were, however, instructed that they could, e.g., choose a location in the middle of a street, if they were unsure on which side of the street the POI was. Some of the

Figure 3.11. The location of the user study, with one participant performing the test.

POIs were not visible from the location of the participants and were presented with obstruction cues (described in more detail later in this section) in order to provide the participants with a hint that they were obstructed by other buildings. The participants were, however, not told anything regarding the visibility of the POIs. A map of the user study area and the intended locations of the POIs are shown in Fig. 3.12.

The user study consisted of four main conditions, repeated in two different versions, resulting in a total of eight conditions. The four main conditions were:

1. Speech presented once, with distance cues provided by modification of the intensity, the early-to-late energy ratio, and the type of speech. Additionally, obstruction cues were applied.

2. Speech presented repeatedly, with distance cues provided by modification of the intensity, the early-to-late energy ratio, and the type of speech. Additionally, obstruction cues were applied.

3. Speech presented once, with the distance announced in meters as the only distance cue. Obstruction cues were not applied.

4. Noise bursts presented repeatedly, with distance cues provided by modification of the intensity and the early-to-late energy ratio. Additionally, obstruction cues were applied.

The participants were first presented with all of the four conditions in random order (version a of the conditions). During each condition, all of

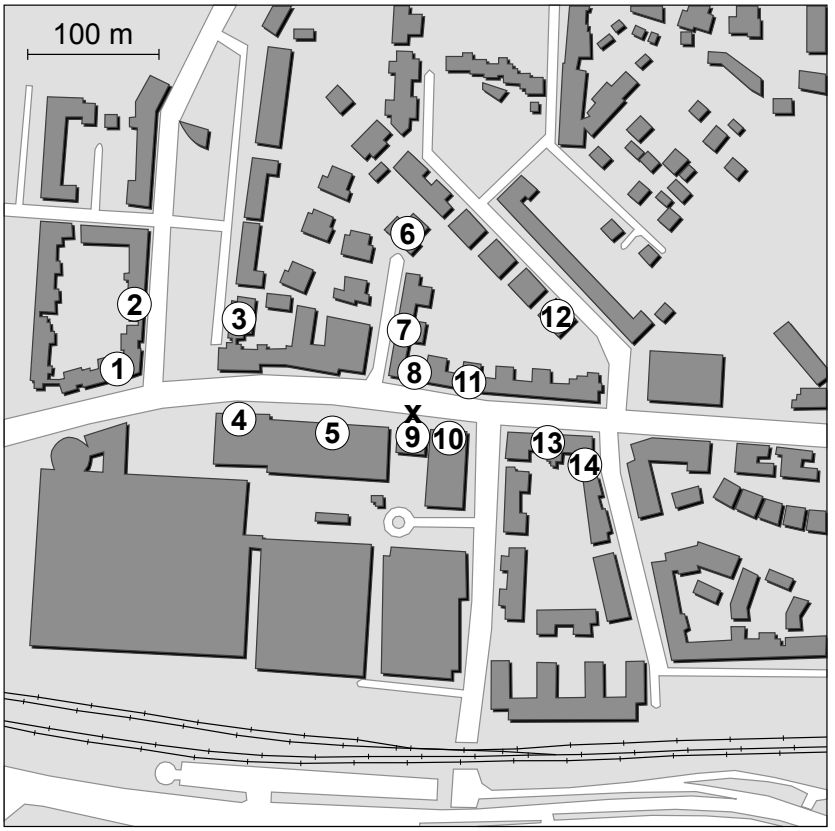


Figure 3.12. Intended locations of the POIs presented in the user study. The POIs are represented by numbers and the location of the participant by a cross.

the 14 POIs were presented in random order. After this, the same conditions were again presented with the order of the POIs and the conditions randomized once more, but this time each condition was preceded by a short training session (version *b* of the conditions). The training consisted of ten POIs that differed in location from the POIs presented during the actual tests, but covered approximately the same area. During training, the participants first heard the sound representing the POI and placed it on the map, after which the intended location of the POI was shown on the map. No feedback was, however, given during the actual tests.

The speech samples used under conditions 1 and 2 contained either a synthesized whispering, speaking, or shouting male voice. Under conditions 1 and 2, the voice said "here's a point of interest." Under condition 3, a speaking male voice said "here's a point of interest at x meters," where x was the distance to the POI rounded to the nearest 10 m. Under condition 4, the stimulus consisted of 250-ms bursts of white noise.

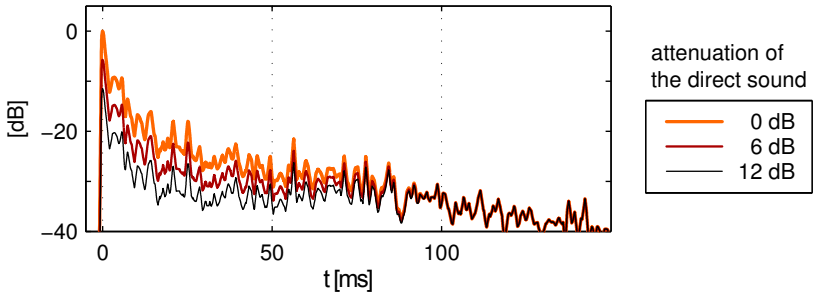


Figure 3.13. An example of modifications of the early-to-late energy ratio. The attenuation of the early energy gradually decreases from its full value at the direct sound to 0 dB at 100 ms. Note that if only the direct sound were attenuated by 12 dB or more, the first reflections would be stronger than the direct sound.

The speech and noise samples were presented using HRIRs from the CIPIC database (Algazi et al., 2001) for the direct sound and BRIRs for early reflections and late reverberation. The BRIRs were selected from a set of BRIRs measured in different outdoor environments: in the vicinity of a large building on one side, on a street with buildings on both sides, in an open field, and in a moderately dense forest. The suitability of these different BRIRs for presenting virtual speech sources in varying acoustic environments was assessed by the author through informal listening. The BRIRs measured in a forest were deemed to provide the best integration of virtual speech sources in different environments and were therefore selected for the user study. The versatility of the selected BRIRs was probably due to the fact that they contain a moderate amount of diffuse reflections from different directions. These "generic" BRIRs were chosen for the user study in preference to BRIRs measures in situ for two reasons. First, in situ measurements would have been difficult due to the noisy location of the user study. Secondly, practical mobile augmented reality applications cannot perform in situ measurements of BRIRs, but rather have to rely on more generic BRIRs. The utilized BRIRs were measured at source azimuths of 0°, 90°, 180°, and 270°, and were interpolated between these angles.

Under conditions 1, 2, and 4, distance cues were provided through early-to-late energy ratio and intensity modifications, based on the results of Publication I. E/L modifications were performed with gradual changes in the energy (and not with discrete dividing points as in Publication I), as depicted in Fig. 3.13. This corresponds quite well with the relative decrease of different parts of the early energy that can be observed in

Fig. 3.2. In addition to modifying the E/L, the intensity of the whole impulse response was attenuated slightly with distance, so that the late reverberation would not sound unnaturally loud at large distances.[1] The E/L and intensity modifications were done so that the intensity of the direct sound varied with distance to a much lesser degree than in reality, to ensure that distant sound sources remained audible.

Under conditions 1 and 2, the type of speech (whispering, speaking, or shouting) was used as an additional distance cue, as proposed by Brungart and Scott (2001). A realistic mapping between distance and type of speech would have resulted in almost all POIs being presented in a shouting voice, so a different mapping was instead chosen: POIs closer than 20 m were presented with a whispering voice, POIs between 20 and 100 m away were presented with a normal conversational voice, and POIs more than 100 m away were presented with a shouting voice. Under condition 3, the distance announced in meters was the only distance cue available and BRIRs were used only to aid in externalization.

Additionally, obstruction cues were utilized under conditions 1, 2, and 4 for POIs that were obstructed from the point of view of the participant (i.e., POIs no. 2, 3, 6, 7, 12, and 14). Because obstruction of speech sources by large buildings normally would render them inaudible, obstruction cues provided by small buildings were instead studied through measurements when preparing the user study. These measured BRIRs were too noisy to be used as such, but cues similar to those observed in these BRIRs were instead applied to the same BRIRs that were used to present unobstructed POIs: the direct sound and the early reflections were heavily attenuated and the whole room impulse response was low-pass filtered. Since the early energy was already heavily attenuated, no further E/L modifications were performed on the obstructed BRIRs; only intensity modifications were done to produce distance cues. Obstructed sound sources included the direct sound from the measured BRIRs, not from separate HRIRs as for unobstructed sound sources.

In the following analysis, differences between conditions are compared using the two-sided sign test, which makes very few assumptions about the distributions under test. The differences between unobstructed and

[1] While the late reverberation in many spaces with a limited volume can be approximated by a location-independent diffuse sound field, the level of the late reverberation in an outdoor environment depends on the distance to the sound source.

obstructed POIs, which cannot be compared pairwise, are instead compared with the two-sided Wilcoxon rank sum test. For sign tests, the Z-statistic is reported for large samples where normal approximation was used to calculate the p-value. For small samples, the number of pairs, S, where the data for the first condition are larger than for the second condition is reported together with the total number of pairs, n. For Wilcoxon rank sum tests, the rank sum test statistic W and the Z-statistic are reported.

3.2.2 Results

The interquartile ranges (IQRs) of the distance and azimuth estimates (extracted from the POI locations estimated by the participants) are shown in Figs. 3.14 and 3.15. Tables 3.9 and 3.10 show the mean IQR of the distance estimates, the mean circular variance V of the azimuth estimates, and the mean answering time for the different conditions. In order to make IQRs comparable, the IQR for each POI and condition was divided by the median distance estimate of that POI and condition.

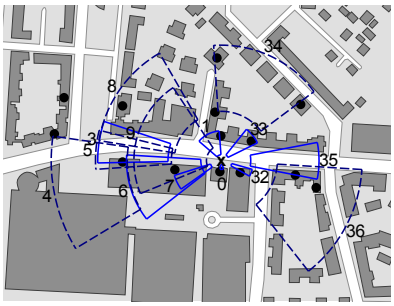
When comparing the dispersion in the distance estimates, measured by the IQR, it was significantly larger under condition 1 than under condition 3 ($S = 25, n = 28, p < 0.001$). More precise distance estimates were thus attained by giving the distance in meters than by providing auditory distance cues. The circular variance was also significantly larger under condition 1 ($S = 22, n = 28, p = 0.004$). The answering times were, however, significantly shorter under condition 1 than under condition 3 ($Z = -11.4, p < 0.001$). Thus, the increase in precision in the location estimates under condition 3 came at the expense of longer answering times.

There was no significant difference in the dispersion of the distance estimates between conditions 1 and 2 ($S = 9, n = 28, p = 0.087$). Repeated presentation of the stimuli did thus not significantly increase the precision of the distance estimates. It did, however, increase the precision of the azimuth estimates ($S = 21, n = 28, p = 0.013$). Again, the increased precision came at the expense of longer answering times ($Z = -11.8, p < 0.001$).

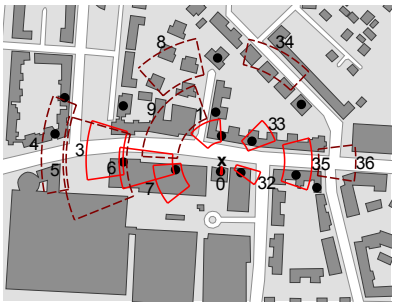
No significant difference was found between speech (condition 2) and noise bursts (condition 4) in the dispersion of the distance estimates ($S = 16, n = 28, p = 0.57$). There was, however, a significant difference in the circular variance ($S = 7, n = 28, p = 0.013$). This difference can mostly be observed among the obstructed POIs, which appear more difficult to

(a) Condition 1a: speech, distance cues, single presentation.

(b) Condition 1b: speech, distance cues, single presentation, feedback.



(c) Condition 2a: speech, distance cues, repeated presentation.



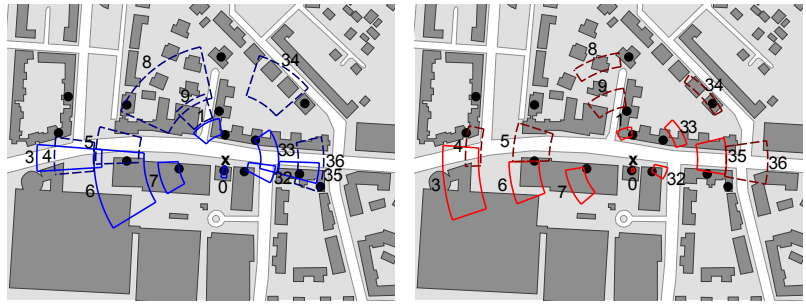
(d) Condition 2b: speech, distance cues, repeated presentation, feedback.

Figure 3.14. Interquartile ranges of the distance and azimuth estimates under conditions 1 and 2. The interquartile ranges of obstructed POIs are drawn with dashed lines. The intended POI locations are represented by black dots.

localize in azimuth when presented with noise (Figs. 3.15(c) and 3.15(d)) than when presented with speech (Figs. 3.14(c) and 3.14(d)).

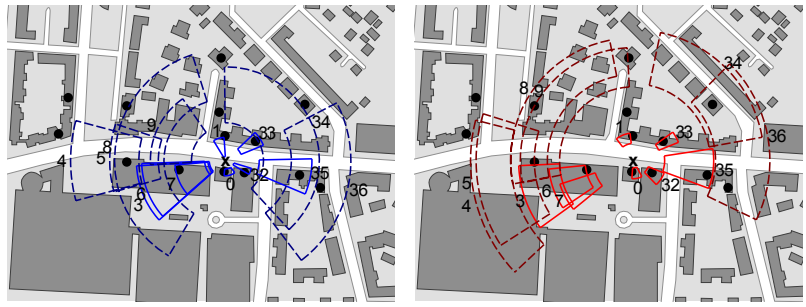
The participants made significantly shorter distance estimates before feedback was given than after training with visual feedback under conditions 1 ($Z = -3.6, p < 0.001$), 2 ($Z = -3.8, p < 0.001$), and 4 ($Z = -5.2, p < 0.001$), where auditory distance cues were provided. Under condition 3, where the distances were given in meters, training with feedback did not significantly affect the length of the distance estimates ($Z = -0.85, p = 0.40$). For all conditions as a whole, training reduced the dispersion of the distance estimates ($S = 49, n = 56, p < 0.001$), but did not affect the circular variance ($S = 29, n = 56, p = 0.89$).

Table 3.11 shows how training with feedback affected the accuracy of the distance and azimuth estimates. Under all conditions, training reduced the deviation of the distance estimates from the distances of the intended locations of the POIs. Training did not, however, significantly



(a) Condition 3a: speech, distance given in meters, single presentation.

(b) Condition 3b: speech, distance given in meters, single presentation, feedback.



(c) Condition 4a: noise, distance cues, repeated presentation.

(d) Condition 4b: noise, distance cues, repeated presentation, feedback.

Figure 3.15. Interquartile ranges of the distance and azimuth estimates under conditions 3 and 4. The interquartile ranges of obstructed POIs are drawn with dashed lines. The intended POI locations are represented by black dots.

reduce the deviation of the azimuth estimates. When comparing conditions 1 and 3, the distance estimates were closer to the intended distances when the distance was given in meters than when auditory distance cues were provided, both before ($Z = 2.5, p = 0.013$) and after feedback ($Z = 3.3, p < 0.001$).

Table 3.9. Mean interquartile range of the distance and circular variance of the azimuth estimates.

	$\overline{IQR}$		$\bar{V}$	
condition	a	b	a	b
1	0.82	0.54	0.21	0.18
2	1.18	0.53	0.16	0.11
3	0.52	0.28	0.11	0.06
4	0.98	0.57	0.21	0.24

Table 3.10. Mean answering time.

	t [s]	
condition	a	b
1	5.7	4.7
2	14.1	9.4
3	10.4	7.4
4	12.9	9.1

Table 3.11. RMS deviation from the intended locations of the POIs. Distance deviations are shown as percentages of the intended distances. The deviations without feedback given (a) and after visual feedback (b) were compared with a two-sided sign test, for which Z-statistics and p-values are shown.

	distance				azimuth			
condition	a	b	Z	p	a	b	Z	p
1	73%	54%	4.2	**<0.001**	58°	48°	1.9	0.062
2	75%	46%	5.5	**<0.001**	47°	34°	0.39	0.70
3	65%	36%	3.9	**<0.001**	37°	26°	1.9	0.054
4	53%	49%	2.1	**0.037**	51°	55°	1.5	0.14

The directions of the obstructed POIs were significantly more difficult to estimate than those of the unobstructed POIs, as measured by the circular variance ($W = 2091, Z = 5.07, p < 0.001$). On the other hand, the dispersion in the distance estimates was significantly smaller for the obstructed POIs ($W = 1187, Z = -3.10, p = 0.002$).

3.2.3 Discussion

Based on the results of the user study in Publication II, presenting POIs with a voice saying the distance in meters is a good approach if accurate and precise distance information is needed. This approach can be recommended especially if accurate distance estimates are required without prior training. The auditory distance cues utilized in the study produced less accurate and precise distance estimates than giving the distances in meters did. Of course, the difference in accuracy does not come as a surprise, since the mapping between the auditory distance cues and the intended distances of the POIs was not naturalistic. The use of auditory distance cues also resulted in less precise azimuth estimates, but this was probably mostly related to the presentation of the obstructed POIs.

When auditory distance cues were utilized, the distance estimates were shorter before than after visual feedback was given. Under all conditions, training with visual feedback resulted in more accurate distance estimates, measured by the deviation from the intended locations of the POIs. The precision of the distance estimates also improved through training. These results suggest that people can, with training, quickly learn to map auditory distance cues to distances in the surrounding.

Training did not, however, have any significant effect on the accuracy or precision of the azimuth estimates. While Shinn-Cunningham et al. (1998) showed that people can, with training, perform remapping of azimuths in situations where head movement is prohibited, the situation in

Publication II was considerably more complex. The main error in the azimuth estimates was probably due to the magnetic deviation of the head tracking. Since the deviation depended on the orientation of the head tracker, the direction from where the sound was presented also varied depending on the direction that the participant was facing. In addition, the accuracy of human azimuth estimation also varies with sound source azimuth (Blauert, 1997, p. 41). Due to both of these factors, it is presumable that participants made different azimuth estimates depending on if they had time to turn towards the sound source or not.

Based on the study, utilizing auditory distance cues for presenting the distances of POIs can be recommended if high precision or accuracy is unnecessary. Without training, distance estimates are based on the available cues and on assumptions regarding the locations of the POIs. Thus, these estimates may or may not coincide with the distances of the POIs. However, people can with training learn to map these cues to the actual distances of the POIs. While the accuracy and precision of these distance estimates still might be inferior to those obtained by presenting the distances in meters, using auditory distance cues allows the spoken messages to be shorter, thus reducing the time needed for processing the messages, as well as auditory clutter and presumably also cognitive load.

Of course, in a real application, training would presumably be a natural part of learning to use the application, rather than being provided by a separate training mode (even though this certainly is a possibility). As the user listens to the presentation of the POIs and tries to localize these in the real world, he or she will over time learn the mapping between the provided cues and real-world distances.

If the range of distances to be presented is large, presenting distances in meters might be a better approach. If auditory distance cues are utilized, the resolution for presenting shorter distances decreases as the range grows. However, if distances are presented in meters, the resolution for presenting shorter distances is independent of the range.

Some participants in the user study thought that the task was easier when distances were given in meters, while others had trouble associating these distances with distances on the map. Different approaches might thus suit different people.

Repeated presentation of the POIs resulted in more precise azimuth estimates (but not distance estimates), since more time was available, and also used, for localizing the sound source. In practical applications, a

single presentation might be appropriate at first, but there should be a possibility to repeat it if necessary. Constantly repeated presentation is appropriate when the user wants to navigate to a single POI.

No significant difference was found between speech and noise in the dispersion of the distance estimates, even though type-of-speech cues were utilized for speech. The precision of the azimuth estimates was lower for noise than for speech, but this was probably mostly due to the difficulty of estimating the azimuth of obstructed POIs presented with noise. Based on this study, no recommendations can be made regarding the type of stimuli used to present POIs. Of course, speech allows different types of information to be easily conveyed to the user.

Based on the debriefing of the participants, it became clear that the obstruction cues utilized in this study had failed in their task. While no explicit questions were asked about these cues, only one participant said that she intuitively associated them with POIs behind buildings. Other participants said that they associated these cues, e.g., with the POI being behind them or inside a nearby building. Since the POIs presented with obstruction cues had significantly different dispersion in the distance and azimuth estimates compared with the unobstructed POIs, the inclusion of such cues under conditions 1, 2, and 4 reduced the clarity of comparison of these conditions with condition 3. In hindsight, the use of obstruction cues should preferably have been studied separately.

The larger dispersion in the azimuth estimates among obstructed POIs can be explained by the fact that these were not presented with the aid of the CIPIC HRTFs, like the unobstructed POIs were. Instead, the direction was provided simply by interpolating the BRIRs measured in four different directions. This approach resulted in large dispersion especially when presenting noise samples.

Interestingly, the distances of the obstructed POIs were estimated more precisely than the distances of the unobstructed POIs. A possible explanation to this might lie in the fact that the obstructed POIs were not presented using E/L modifications. It might thus be that the availability of more distance cues for the unobstructed POIs resulted in more dispersion in the distance estimates. This could imply that more cues are not always better, but further studies are needed to confirm this notion.

Under condition 3a, the distances to the POIs were given in meters. While the participants were not aware of the fact that the POIs under the other conditions (1a, 2a, and 4a) were the same as under condition

3a, they might have deduced that these were likely to be located within approximately the same range of distances. Thus, conditions presented to a participant after condition 3a might have been biased by this deduction. Statistical analysis (shown in Publication II) showed that such an order effect might exist, but also showed the presence of other effects related to the presentation order. While presentation of condition 3a after conditions 1a, 2a, and 4a would have removed the possibility of deducing the range of distances based on this condition, it would have introduced other order effects, related to, e.g., fatigue.

It is worth noting that the distance estimates made by the participants before training with feedback already were relatively large. These distance estimates were considerably larger than the presumed perceived distances of the auditory events produced by the stimuli, and also larger than the distances of hypothetical real sound sources with the corresponding distance cues. Thus, the participants already did a mapping of these distance cues to longer distances in the real world, based on the instructions given: the sounds represent POIs located in buildings in the surrounding area. For this reason, the difference in magnitude between distance estimates made before and after feedback was relatively small. Presumably, the participants would still have been able, with training, to map these cues to distances, e.g., two times farther away than in this user study. Further studies would, however, be necessary to determine how different mismatch between the provided distance cues and the intended distances would affect the precision and accuracy of distance estimates attained after training.

3.3 Conclusions

Publication I presents practical methods for modifying the perceived distance of virtual sound sources. In particular, adjusting an early-to-late energy ratio, including the first 50–100 ms of the room impulse response in the early energy, is shown to effectively modify the perceived distance of speech sources. These methods and results are applicable in many different spatial audio applications, but they are especially important for augmented reality and other real-time applications. The publication also provides some new insight into how people interpret reverberation cues, but further studies are needed to explain the mechanisms used for this

and to determine how these cues are interpreted under different conditions.

Publication II shows that people can, with a small amount of training, map auditory distance cues to different distances in the real world around them. While the study was performed in an augmented reality setting, these results are also relevant for, e.g., virtual reality applications, where virtual sound sources are not linked to the real world, but to a virtual world surrounding the user. The results are also useful in different types of auditory displays, where auditory distance cues are used to present different distances or other types of information to the user.

4. Spatial Audio Guidance

People normally use their visual sense to navigate through the world around them, both for finding their way to another location, and for finding objects of interest. Different navigational aids can be used to help in this task. Since the visual sense already is occupied with observing the environment, navigational aids utilizing other senses would be preferable in many situations. In addition to occupying the visual sense, visual navigational aids for pedestrians, e.g., maps, often also occupy one or both hands. Using auditory interfaces instead leaves both eyes and hands free for other tasks and thereby increases safety. While visual heads-up displays also leave the hands free and may reduce distraction of the visual attention, these are limited to the field of view, while sounds can be perceived in any direction any time. Compared with tactile interfaces, auditory interfaces make it easy to convey larger amounts and different types of information.

While guidance can be given with verbal instructions, using spatial audio has been shown to increase performance and reduce cognitive load compared with verbally indicating the direction in navigation tasks (Loomis et al., 1998; Klatzky et al., 2006). Spatial audio guidance has many possible applications, including aviation (Begault et al., 1996), locating objects indoors (Sandberg et al., 2006), and locating points of interest outdoors (Publication II). One particular target group for spatial audio guidance, as well as audio guidance in general, are the visually impaired (Loomis et al., 1998; Blum et al., 2012).

One of the obvious applications of spatial audio guidance is pedestrian navigation. Holland et al. (2002) developed a spatial audio navigation interface, AudioGPS, to be used with eyes, hands, or attention otherwise engaged. To indicate the direction to the next waypoint, amplitude panning was applied to the presented musical tone. To provide more exact

feedback about the user's direction of motion compared with the direction to the next waypoint, a so called chase tone was played. While the main navigation tone was repeated at a constant pitch, the chase tone had a pitch that differed from the main tone the more, the larger the deviation from the correct direction was. When the user was walking straight towards the waypoint, the pitches of the main tone and the chase tone were the same.

Amplitude panning, i.e., utilizing interaural level differences, is a simple approach for presenting different azimuths. However, it may require further cues to help distinguish between directions in front of and behind the user. In the AudioGPS application, this was achieved by playing a sharp tone for sound sources in front of and a muffled tone for sources behind the user. Utilizing HRTFs provides more natural directional cues, and has in some cases been shown to provide better performance in navigational tasks (Larsen et al., 2013). On the other hand, HRTFs come with a considerably larger demand for processing power.

Calvo et al. compared auditory and tactile navigation displays with map-based navigation in two different studies. The auditory display used HRTFs and the tactile display a vibration belt, both having a resolution of 45°. In the first study (Calvo et al., 2013), these two displays were compared with an allocentric map showing the position of the participant and the next waypoint. No difference in completion times or navigation errors was found between the display types. While the first study took place on the walking paths of a university campus, the second study (Calvo et al., 2014) was conducted in an open field. This time, these three displays were further compared with an egocentric map and a visual arrow with the same resolution of 45°. Using the egocentric map resulted in significantly shorter completion times than using the tactile display or allocentric map. Using the auditory display resulted in significantly, but only slightly, shorter completion times than using the tactile display. Calvo et al. concluded that both auditory and tactile displays produce low mental workload and are effective means for eyes-free navigation.

4.1 Music guidance for pedestrian and cyclist navigation

While different sounds can be used specifically for navigation, one might question if people would want to listen to such continuous or repeated sounds for long periods of time. Instead, navigation applications could

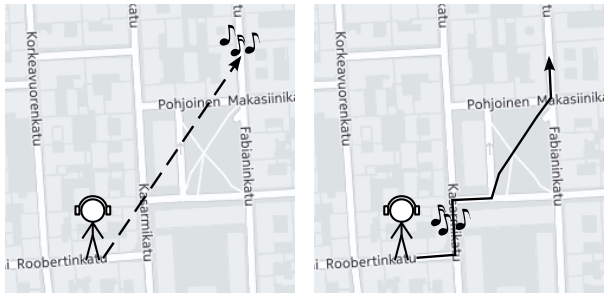
take advantage of sounds that the user otherwise would listen to, such as music.

Etter and Specht (2005), Strachan et al. (2005), and Jones et al. (2008) all proposed and tested music guidance for pedestrian navigation. To indicate the direction to the next waypoint or the destination, stereo balance adjustment or amplitude panning was applied to the music. The user studies all showed music guidance to be a promising approach. While stereo balance adjustment provides the benefit that the music is heard normally when walking in the right direction, Jones et al. pointed out that stereo effects in the music might be mistaken for navigational cues.

Eyes-free and hands-free navigational aids are beneficial for pedestrians, but the need for these is even more evident when it comes to cycling. Little research has, however, been done on cyclist navigation: Pielot et al. (2012) investigated a tactile interface for conveying direction, attaching vibration motors to the handlebars of the bicycle, whereas Zwinderman et al. (2011) briefly tested music guidance for cyclists. While Etter and Specht, Strachan et al., and Jones et al. panned the music relative to the compass heading of hand-held devices or the current direction of motion, Zwinderman et al. utilized the magnetometer of a smartphone attached to the headphones.

Previous studies on navigation with music guidance (Etter and Specht, 2005; Strachan et al., 2005; Jones et al., 2008; Zwinderman et al., 2011) provide interesting and useful insights, but only brief reports of user studies. Publication III aims to provide a more detailed and extensive look into the user experience of a music guidance application, considering, e.g., different types of guidance, normal stereo music listening vs. spatial audio, and safety issues. Compared with previous studies, this study differs by using HRTFs instead of amplitude panning or stereo balance adjustment, by investigating both pedestrian and cyclist navigation, and by utilizing head orientation tracking (also used by Zwinderman et al.).

The utilization of HRTFs and head orientation tracking for the presentation of spatial audio offers several benefits. It provides a natural representation of different directions, which presumably results in a clearer and more intuitive navigation experience. The use of HRTFs reduces front-back confusion and other directional confusion present when using stereo balance adjustment or amplitude panning. By using head tracking, such confusion can be further resolved through simple head movements (Wightman and Kistler, 1999; Begault et al., 2001).



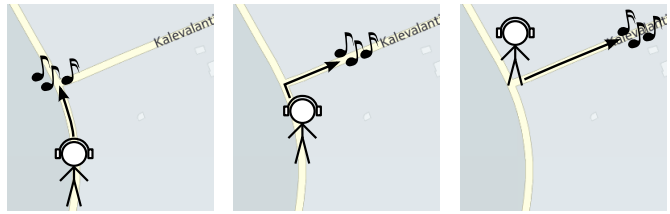
(a) Beacon guidance (b) Route guidance

Figure 4.1. Guidance types investigated in Publication III.

The largest difference to previous studies, however, lies in the investigation and comparison of both beacon guidance and route guidance. Beacon guidance is an often investigated approach for navigation, probably due to its simplicity. In beacon guidance (Fig. 4.1(a)), the direction straight towards the destination is constantly presented to the user, which means that the user must find his or her own route to the destination based on this information. It has been used, in some cases with multiple waypoints, in previous studies on music guidance (Etter and Specht, 2005; Strachan et al., 2005; Jones et al., 2008; Zwinderman et al., 2011). Route guidance (Fig. 4.1(b)), on the other hand, guides the user along a specific route to the destination. Route guidance has previously been proposed (but not investigated) by Strachan et al., who suggested that the sound source could travel at a certain distance on the route in front of the user, like a "carrot on a stick." Väänänen et al. (2014) previously studied route guidance for pedestrians with the sound of a walking horse leading the way. Figure 4.2 illustrates how the virtual sound source, or the so called audio guide, moves ahead of the user on the route to the destination. The user is thus supposed to follow the audio guide leading the way at each intersection in order to reach the destination.

4.1.1 Methods

Publication III presents three user studies that were conducted to evaluate the user experience of music guidance for both pedestrian and cyclist navigation. These user studies were designed to provide answers especially to the following research questions:



(a) (b) (c)

Figure 4.2. During route guidance, the audio guide moves along the route at a certain distance ahead of the user.

- What is the overall user experience of using spatial music guidance for pedestrian and cyclist navigation?
- How well can people perceive the direction of spatialized music and use this direction to guide them to a destination?
- What are the preferences between route and beacon guidance?
- Do people feel safe when navigating with this type of music guidance?
- Is spatialized mono music considered pleasant to listen to?
- How does the user experience differ when using hear-through headphones compared with headphones that block out sounds from the environment?

The navigation interface used in these studies was a further development of the interface evaluated by Väänänen et al. (2014), utilizing HRTFs and head orientation tracking to present a virtual music source in the direction where the user should be guided using either route or beacon guidance. The user study conditions are summarized in Table 4.1. During route guidance tasks, participants were supposed to find their way to the destination following a predefined route. Rerouting was not enabled, i.e., participants that left the intended route were always directed back to the closest point on the route. During beacon guidance tasks, participants had to find their own route to the destination. In both cases, the task was performed with the help of the music guidance alone; participants were not shown a map or told what their destination was.

The participants wore microphone hear-through headphones, a further development of the headphones presented by Albrecht et al. (2011), allowing software control of the hear-through level. The participants also carried a head orientation tracker and a computer running the navigation

Table 4.1. Summary of the user study conditions.

	user study A	user study B	user study C
method	walking	walking	cycling
location	city	city	suburbs
tasks	route guidance 1	route guidance	route guidance
	route guidance 2	beacon guidance	beacon guidance
	beacon guidance		
counterbalanced	no	yes	yes
route guidance	guidance at turns	constant guidance	constant guidance
hear-through	yes/no	yes	yes
participants	12	9 (10)	12

software in a backpack. During the tasks, the participants listened to music that they could choose from a small number of different alternatives.

In the first user study, user study A, twelve participants performed three navigation tasks in an urban environment. The first two tasks were route guidance tasks, while the third task was a beacon guidance task. The three tasks were performed on different routes, with the order being the same for all participants. The hear-through functionality of the headphones was disabled during one of the two route guidance tasks, while it was enabled during the other route guidance task (with the order counterbalanced between participants) and the beacon guidance task. During the route guidance tasks, the music was presented as normal stereo (in so called stereo mode) when the participant was supposed to continue in the current direction. The music was presented in guidance mode, as depicted in Fig. 4.2, only when the participant was near an intersection and was supposed to make a turn. The routes in user study A were planned so that the tasks would take approximately ten minutes each.

User study B employed a further developed version of the navigation interface utilized in user study A, with the music constantly presented in guidance mode during route guidance. During this user study, ten pedestrians performed a route guidance task and a beacon guidance task in an urban environment. The planned length of the tasks was 15 minutes each. Two different routes were used, and the order of the tasks as well as the order and direction of the routes was counterbalanced between participants. The hear-through functionality of the headphones was enabled during both tasks.

User study C consisted of twelve participants cycling in a suburban environment. Two tasks were again performed: one route guidance task and one beacon guidance task. The planned length of the tasks was 15–20 minutes each. The order of the tasks as well as the order and direction

of the routes was counterbalanced between participants. Also during this user study, the music was constantly presented in guidance mode during route guidance. The hear-through functionality was enabled during both tasks, but the level of hear-through was in a few cases decreased because of disturbing wind noise.

After each task, the participants answered a questionnaire and a number of interview questions, providing both quantitative and qualitative answers. Since the total number of questions was large, only the questions of most interest are summarized here. Due to technical issues, the questionnaire answers of one of the ten participants in user study B were removed from the final results.

The statistical analysis of differences in the results was done with a two-sided sign test. The number of pairs, S, where the data for the first condition are larger than for the second condition is reported together with the total number of pairs, n, and the p-value.

4.1.2 Results

Maps of the routes that the participants took are shown in the appendix of Publication III. All participants completed all tasks successfully, but in some cases the routes taken during route guidance deviated slightly from the intended routes.

In general, the participants in the user studies liked the navigation interface using music guidance, and most participants said that they would use this type of interface. Figure 4.3 displays the preference between route and beacon guidance during the three user studies. During user study A, there was a general (but not significant, $S = 2, n = 12, p = 0.065$) preference for beacon guidance. Based on the interviews, participants seemed to prefer beacon guidance since the constant switching between stereo mode and guidance mode during route guidance was perceived as disturbing. In general, participants did not find that the music being played from a mono source using HRTFs reduced the enjoyability of the music very much (Fig. 4.4). Thus, switching to stereo mode when guidance is not needed can be seen as an unnecessary and even unwanted feature. Many participants also found it confusing that the switching did not happen at the same distance before each turn due to GPS location inaccuracy. Based on these findings, the switching between modes was removed from the navigation interface in user studies B and C, where the music instead was constantly presented in guidance mode.

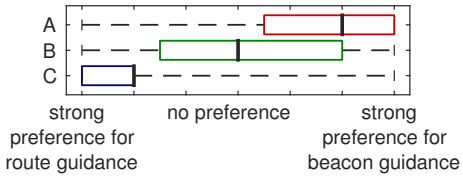


Figure 4.3. Preference between route and beacon guidance in the three user studies.

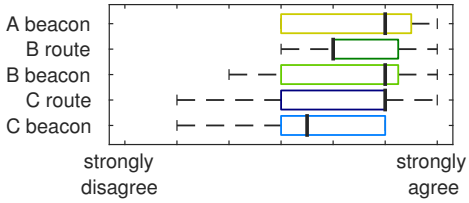


Figure 4.4. Enjoyability of the music: "I think that listening to the music was pleasant compared with normal stereo listening."

No clear preference between route and beacon guidance could be seen when participants walked in the city in user study B ($S = 4, n = 9, p = 1$). Several participants said that they would use beacon guidance if the environment was somewhat familiar and route guidance in unfamiliar environments. A few participants, however, said that they would want to use beacon guidance for exploring new areas. Some participants said that route guidance might be a better alternative when in a hurry.

When cycling in the suburban environment in user study C, participants preferred route guidance over beacon guidance ($S = 10, n = 12, p = 0.012$). The main reason for this was presumably related to the environment rather than associated with cycling. Many participants commented that it was challenging for them to find their route to the destination with beacon guidance, since they did not know if their choice of route would lead them towards the destination or perhaps only to a dead end. Most participants did, however, find a good route to the destination, but the feeling of uncertainty during beacon guidance (Fig. 4.5) made them prefer route guidance.

One of the weaknesses of route guidance was that participants did not know where they were supposed to turn next until they approached the next intersection. This problem was emphasized in these user studies, since the participants did not know what their destination was beforehand, and could thus not anticipate upcoming turns. Some participants

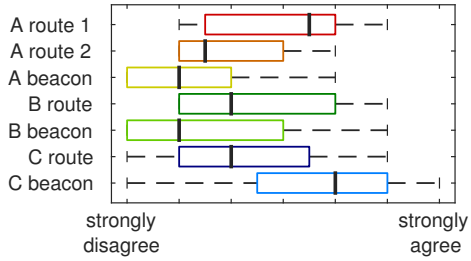


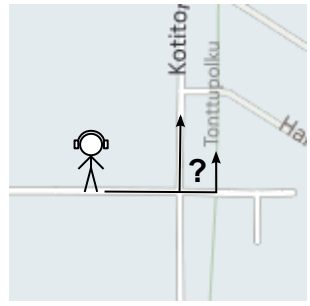
Figure 4.5. Feeling of uncertainty: "I was often uncertain whether I was going in the right direction."

commented that a good feature of beacon guidance was that they constantly knew in which direction the destination was.

With route guidance, most participants found it easy to know if they should make a turn left or right, but they generally found it difficult to distinguish between alternative paths with less than a 90-degree angle in between (Fig. 4.6(a)). Participants also found it difficult to know which of two consecutive paths to take (Fig. 4.6(b)).

In general, the participants felt safe while using the music guidance (Fig. 4.7) and did not feel that using this type of interface reduced their sense of safety compared with normal music listening when walking or cycling. A few participants did, however, notice that the guidance at intersections took some of their attention away from the traffic. On the other hand, several participants thought that it was safer to listen to the

(a) The audio guide leads the user slightly to the right, but there are two paths to the right. The user cannot perceive the direction of the guidance precisely enough to be sure which path to take.



(b) The audio guide leads the user left, but there are two paths leading left close to each other. The user is unsure which path to take.

Figure 4.6. Two problematic situations when using route guidance.

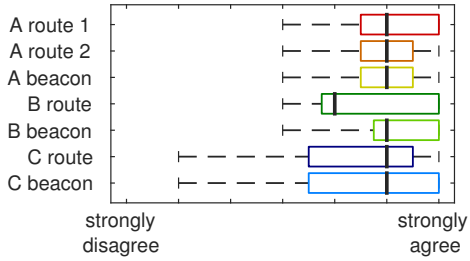


Figure 4.7. Sense of safety: "I felt safe while using the guidance."

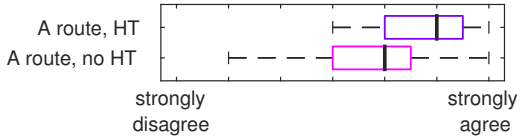


Figure 4.8. Clarity of hearing the environment during the route guidance tasks in user study A: "I could hear my surroundings clearly during the task." During one of the tasks (*HT*), the hear-through functionality of the headphones was enabled. During the other task (*no HT*), it was disabled.

guidance than to have their vision occupied by looking at a map.

In user study A, the hear-through feature of the headphones was disabled during one of the route guidance tasks and enabled during the other task. The participants did, however, not report any significant difference between these tasks in how they heard the environment ($S = 6, n = 12, p = 0.29$), as shown in Fig. 4.8. Some participants said that they normally want to hear the environment when listening to music on the go, while others prefer to block out the environment and only listen to the music.

When asked what types of notifications they would want when using music guidance, many participants mentioned that they would want to know the remaining distance to the destination. In general, speech was considered a good method for conveying such information, since it stands out from the music and is easy to interpret.

4.1.3 Discussion

The results of the user studies in Publication III indicate that both investigated guidance types, route guidance and beacon guidance, are good alternatives, but that they may be suitable for different situations. Beacon guidance is a good option when the environment offers clear routes and the user only needs to be guided in the general direction of the destination. In case it is more difficult for the user to know which routes lead

towards the destination, route guidance might be the preferable option. Alternatively, beacon guidance could be extended with automatically or manually added waypoints.

In general, participants in the user studies felt safe while navigating using the music guidance. There are, however, two potential safety concerns when using this type of navigation interface: environmental isolation and inattentional blindness (Lichenstein et al., 2012). Environmental isolation is in this case caused by the headphones blocking out or the music masking sounds that warn the user about potential dangers. Inattentional blindness is due to the cognitive distraction of interpreting auditory stimuli and manipulating electronic devices.

Different studies have found different effects of listening to music on pedestrian and cyclist behaviour. Thompson et al. (2013) found music listening to be associated with an increased likelihood among pedestrians to ignore looking both ways before crossing the street. Walker et al. (2012), on the other hand, observed an increased likelihood to look both ways among male pedestrians listening to music, while no effect was seen among female pedestrians. Nasar et al. (2008) did not find a significant difference in unsafe behaviour between pedestrians listening to music and pedestrians not using any technology. De Waard et al. (2010) found that cyclists listening to music associated this with a slightly increased safety risk, but found no effect of music listening on cycling performance. Thompson et al. and de Waard et al. found text messaging to have a negative effect on pedestrian behaviour and cycling performance, respectively. Additionally, Jensen et al. (2010) found that drivers using a car navigation system with visual output displayed more unsafe behaviour than drivers using audio output.

Based on these studies, some conclusions can be drawn. First of all, the target group for the type of navigation interface presented in Publication III is people who already listen to music while walking or cycling. While studies on pedestrian and cyclist safety indicate that there might be a safety risk associated with music listening, this risk should not increase if the music is used for navigational guidance. On the contrary, if the music is presented spatially from a single direction, this should improve the intelligibility of sounds in other directions. Using music guidance for navigation also allows pedestrians and cyclists to have their hands and especially eyes free for other tasks. This should result in less inattentional blindness than when using, e.g., a map. A few participants did, however,

report that their focus on the guidance at intersections might have taken some of their attention away from the traffic. Such effects might diminish as the user gets accustomed to the music guidance, but care must be taken when designing the guidance so that it is clear and does not change drastically at intersections.

Previous studies have suggested and to some extent also tested the use of music volume to convey the distance to the destination or the next waypoint (Etter and Specht, 2005; Strachan et al., 2005; Jones et al., 2008; Zwinderman et al., 2011; Fujimoto and Turk, 2014). While this is an intuitive approach, it has severe limitations. Most importantly, music listeners tend to adjust the volume to a pleasant level, which depends on many factors. Increasing or decreasing the volume much from this level can result in the music being uncomfortably loud or quiet. As the user can and will adjust the volume himself or herself, it can only be used as a relative cue. However, as Jones et al. point out, a gradual decrease in the volume serves as a poor hint that the user is moving away from the destination rather than towards it.

Reverberation cues (Publication I) could also be added to the music to convey the distance to the destination. Effective use of such cues might, however, reduce the enjoyability of the music. Reverberation already present in the music might also affect the interpretation of these cues. Based on the suggestions by the user study participants, the remaining distance could simply be presented using speech. The user should be able to choose how often or when the distance is presented this way.

Some of the participants in these user studies said that they prefer to block out the environment when listening to music, while others would want to hear the environment when using music guidance. Microphone hear-through technology allows the user to adjust how well the environment is heard depending on the situation and personal preferences. Such adjustment could also to some extent be performed automatically, so that noisy environments are attenuated, while more quiet environments are not. The headphones used in these user studies did not, however, either attenuate sounds from the surroundings well or fit very well in the participants' ear canals, since enabling or disabling the hear-through functionality did not significantly affect how clearly the participants could hear the environment.

4.2 Conclusions

Publication III investigates the user experience of pedestrian and cyclist navigation using spatialized music, where the music either leads the way along a route or indicates the direction straight to the destination. Both types of guidance were found to be effective and pleasant methods, but they are suitable for different environments and circumstances. While the investigated navigation interface does not represent a pure augmented reality application as defined by Azuma (1997), it does represent one example of the many other possibilities that the advent of audio augmented reality technology will bring. It also represents an example where full-fledged augmented reality might not be the best approach: presenting the audio guide at a varying distance using intensity or reverberation cues might reduce the enjoyability of the music.

5. Mobile Communication

Härmä et al. (2003) predicted that speech communication will be one of the most important applications of mobile audio augmented reality technology. Using AAR techniques, the voice of another person can be heard as if he or she were in the same room, instead of hearing it from, e.g., a telephone. In addition, thanks to hear-through techniques, other people physically in the same room can be heard unhindered.

For applications where multiple people are participating, the voices of these could be rendered in different directions. This would improve speech intelligibility in case several participants are talking at the same time (Bronkhorst, 2000) and also otherwise produce a more natural communication experience. Ultimately, with the help of visual AR, remote participants would not only be heard, but also seen in the same room.

5.1 Mobile communication with co-location detection

Imagine a teleconferencing scenario, where two persons (X and Y) at the same location are talking to a third person (Z) at a different location, as illustrated in Fig. 5.1. If the teleconferencing system sends the microphone signals of each participant to all the other participants, persons X and Y will hear the voices of each other first naturally and then delayed through the headphones, which would be perceived as an echo. Such echoes would be perceived as disturbing and have a detrimental impact on speech intelligibility (Haas, 1972). To remove these echoes, the teleconferencing system needs to know which of the participants can hear each other naturally. While this information could be input manually, some form of automatic co-location detection would be beneficial especially in mobile teleconferencing scenarios, where participants might move from one location to another.

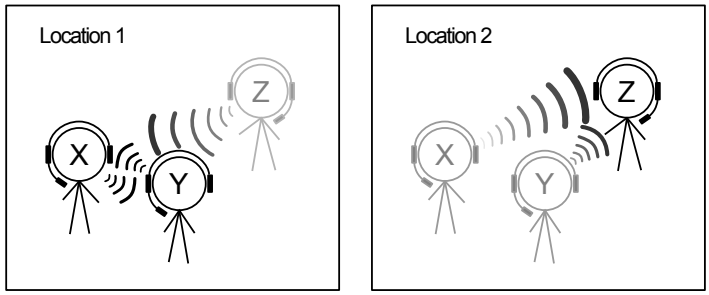


Figure 5.1. A mobile teleconferencing scenario, where two participants (X and Y) are at one location and one participant (Z) is at another location. If sound is transmitted between all participants, X and Y will hear each other's voices first naturally and then delayed through the headphones.

While the voices of remote participants can be presented in mono, this type of system could ideally be implemented as an augmented reality application. By estimating the directions of the local participants (Takanen and Karjalainen, 2010; Gamper et al., 2011), the voices of remote participants, extracted from the microphone signals (Lokki et al., 2004), can be spatially distributed and rendered in different directions. Additionally, the voices of remote participants should be integrated into the local acoustic environment by adding appropriate reverberation, e.g., through the extraction of BRIRs (Gamper and Lokki, 2009) or acoustical parameters (Vesa and Härmä, 2005).

In Publication IV, an acoustic co-location detection (ACLD) algorithm is proposed for the mentioned teleconferencing scenario. The algorithm takes the audio streams of the teleconference participants and analyzes them in real time to infer which participants are co-located. Mel-frequency cepstral coefficients (MFCCs) are extracted from short-time frames of the audio streams, and based on the correlation between the concurrent frames of two participants, the algorithm classifies the two participants as being co-located or not. The classification is done based on a threshold, with hysteresis applied to avoid back-and-forth changes near the threshold. Changes in classification are only done when voice activity is detected in at least one of the two audio streams.

5.1.1 Methods

The ACLD algorithm was evaluated both offline using recorded communication scenarios and in real time integrated into a voice-over-IP (VoIP)

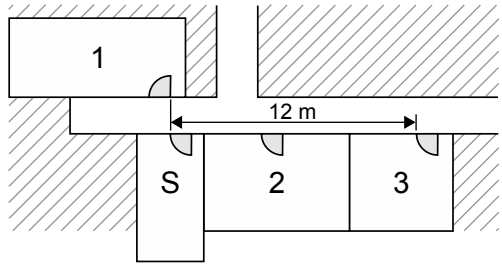


Figure 5.2. Floor plan of the rooms where the evaluation of the acoustic co-location detection algorithm took place. Recordings for the offline evaluation were done with one participant in room 1, one in room 3, and one moving between these two rooms. In the real-time evaluation, the participants moved between rooms 1 and 2, and the stairwell S.

conferencing system. For the offline evaluation as well as for the development of the algorithm, recordings of 16 different mobile communication scenarios with three participants, all wearing microphones, were made. These scenarios took place in two rooms connected by a corridor (see Fig. 5.2), with one participant in one room, a second participant in another room, and a third participant moving between these rooms. Scenarios were recorded where one, two, or all three of the participants were talking.

In the real-time evaluation, the VoIP conferencing system was tested separately by four groups consisting of three participants each. In this evaluation, the audio streams of other participants were presented in mono. The latency (between one participant speaking and another hearing it) in the system was measured to generally be just under 0.5 s. Each participant carried a netbook-type laptop computer, running the VoIP client, and wore microphone hear-through headphones (the headphones presented by Albrecht et al. (2011)).

During the evaluation, the three participants in each group moved between two rooms and a stairwell, all connected by a corridor (see Fig. 5.2), according to a predefined plan. The plan made sure that there were different situations when one, two, or three participants were in one room at the same time, and situations when the doors to the corridor were open and when they were closed. The participants were asked to keep a conversation between all three of them going.

Two test cases, A and B, were evaluated in alternating order, following the same plan. In case A, ACLD was disabled. The audio stream of each

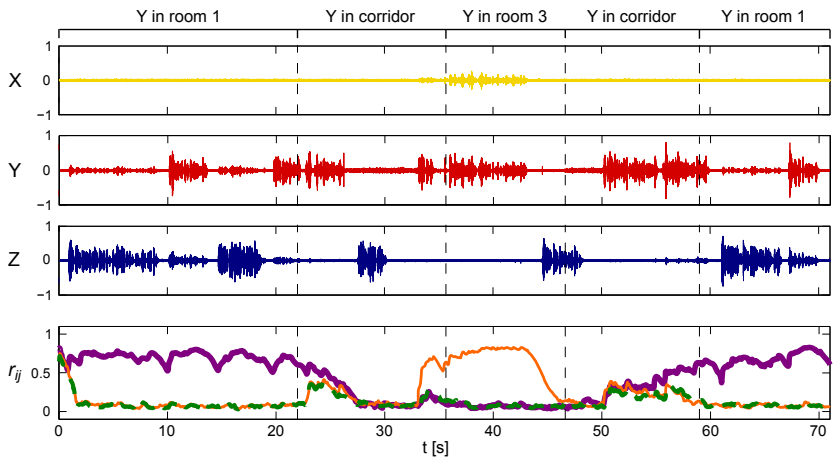


Figure 5.3. Recordings of a communication scenario for the evaluation of the ACLD algorithm. During the recording, persons Y and Z are talking. Person X is in room 3, person Z is in room 1, while person Y moves from room 1 to room 3 and back to room 1, via a corridor. The microphone signals of each person are shown at the top, while the pairwise correlation r_{ij} between these signals is shown at the bottom. The orange line represents the correlation between X and Y, the thick purple line the correlation between Y and Z, and the dashed green line the correlation between X and Z.

participant's speech was thus constantly transmitted to the other participants. The hear-through functionality of the headphones was disabled, so that the acoustic environment around the participant, including other participants talking in the same room, was attenuated by the insert-type headphones. This decision was made because it was hypothesized that attenuating the natural sound of others talking would be less disturbing than having it at an equal level with the same sound played back over the headphones after a short delay.

In case B, ACLD was enabled, and audio streams were thus not transmitted between co-located participants. The hear-through functionality of the headphones was also enabled and adjusted so that participants could hear other participants in the same room at a natural and comfortable level.

The evaluation of each case lasted about ten minutes, after which the participants answered a questionnaire. The analysis of statistical differences in the presented results was done with a two-sided sign test. The number of pairs, S, where the data for the first condition are larger than for the second condition is reported together with the total number of pairs, n, and the p-value.

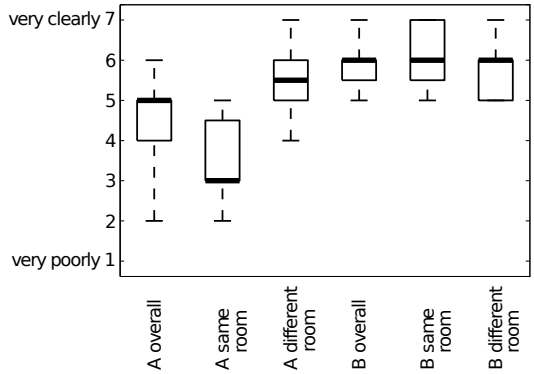


Figure 5.4. Clarity of hearing and understanding the other participants in the same room, in a different room, and overall.

5.1.2 Results

In the offline evaluation, the algorithm correctly identified co-location in 96.5% of frames containing speech, with room-level co-location as the ground truth. An example of the correlation of the MFCCs calculated from the microphone signals during one of the 16 scenarios is shown in Fig. 5.3.

Figure 5.4 shows how participants in the real-time evaluation judged the clarity of communication in different situations. Overall, communication was clearer in test case B than in case A ($S = 0, n = 12, p < 0.001$). Participants reported a significant difference in clarity when they talked to other participants in the same room ($S = 0, n = 12, p = 0.002$), but not when in different rooms ($S = 2, n = 12, p = 0.29$).

In both test cases, participants reported hearing echoes of their own or other participants' voices. In case A, 10 participants heard echoes of the voices of others, while 9 heard echoes of their own voice. In case B, 9 participants reported hearing echoes of others' voices, while 8 participants heard echoes of their own voice. One participant reported that he did not perceive any echoes in either test case. Although echoes were heard by most participants in both cases, they were perceived less annoying ($S = 11, n = 11, p < 0.001$) and deemed to affect the conversation less ($S = 10, n = 11, p = 0.012$) in test case B, as illustrated in Fig. 5.5. Participants also commented that they only seldom heard echoes in test case B, or that the echoes heard were weak.

Presumably because of this difference, participants preferred test case B over test case A ($S = 0, n = 12, p < 0.001$), as shown in Fig. 5.6. This was

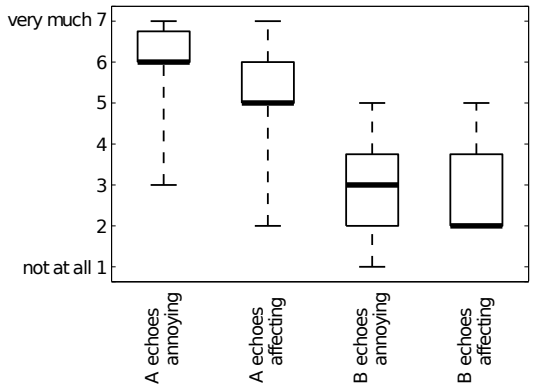


Figure 5.5. Perception of echoes: how annoying they were and how much they affected the conversation.

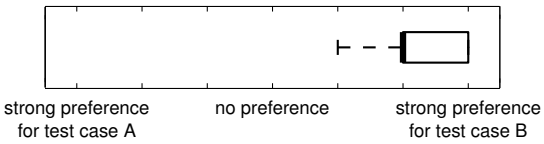


Figure 5.6. Preference between the two test cases.

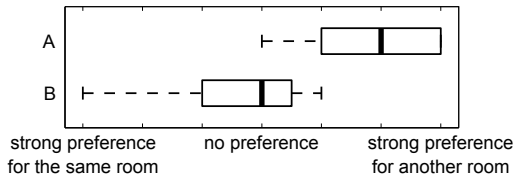


Figure 5.7. Preference for talking to participants in the same room or in different rooms in the two test cases.

associated with a preference for talking to participants in another room in test case A ($S = 0, n = 12, p = 0.002$), while there was no significant preference in test case B ($S = 4, n = 12, p = 1$), as illustrated in Fig. 5.7.

5.1.3 Discussion

The user studies showed that the proposed ACLD algorithm is a promising solution for echo-related problems in mobile communication scenarios where participants may or may not be co-located. Participants suggested that the system would be useful in situations where continuous mobile communication is needed, e.g., at a construction site.

While the proposed co-location detection algorithm was developed for

continuous mobile communication applications, it can of course be used in other applications where co-location detection is beneficial and microphones are available. Of course, the algorithm requires active sound sources in the vicinity to be able to infer co-location. The algorithm may also erroneously infer co-location if the same sound (e.g. the same radio programme) is heard at several locations. Other co-location detection algorithms should thus be used if accurate co-location inference is necessary.

For the presented communication application, the proposed ACLD algorithm offers benefits over other co-location detection methods. For example, by comparing signal strengths of mobile networks (Krumm and Hinckley, 2004; Dashti et al., 2015) one can only infer proximity between two participants, but not tell whether two participants are able to hear each other or not. Other methods rely on specialized hardware and signals (Harter et al., 2002; Mandal et al., 2005), but it is simpler and less obtrusive to utilize the speech signals readily available in a communication application.

Mutual information of voice activity can be used to infer conversation between people (Basu, 2002; Wyatt et al., 2011). Such methods equally detect conversation between co-located and non-co-located people talking to each other (Basu, 2002), and thus need to be modified in order to be useful for the current problem. As Brdiczka et al. (2005) point out, methods based on mutual information of voice activity also require a substantial length of recorded speech and are thus not suitable for real-time applications. Brdiczka et al. instead suggest using a hidden Markov model on voice activity to detect interaction groups, but also this method is useful for detecting conversation rather than co-location.

The most obvious solution to the problem at hand would be to simply calculate the cross-correlation of two raw microphone signals (Corman and Scott, 1994). One advantage of using the frame-based approach presented in Publication IV lies in the comparatively small amount of processing that is required to compensate for the varying delay between the signals. Using MFCCs, it is also possible to easily compensate for differences in the transmission paths of the signals through cepstral mean subtraction (Westphal, 1997). MFCCs are also useful for several other tasks: context recognition (Eronen et al., 2006), speaker recognition and verification (Reynolds et al., 2000), and speech recognition (Davis and Mermelstein, 1980).

5.2 Conclusions

In Publication IV, an acoustic co-location detection algorithm is presented and evaluated in a mobile communication application, where it was found to improve the clarity of communication. While the algorithm is ideally suited for this application, it might also prove useful in, e.g., the analysis of social networks (Corman and Scott, 1994; Wyatt et al., 2011; Dashti et al., 2015), augmented reality games, collaborative or social networking applications, as well as other context-sensitive applications (Krumm and Hinckley, 2004; Brdiczka et al., 2005).

While the mobile communication system with acoustic co-location detection evaluated in Publication IV is not an augmented reality application as such, it represents several steps in the direction of building a mobile augmented reality teleconferencing system, where the voices of remote participants are rendered spatially among the local participants. Further steps to develop such an application include extraction of the remote participants' voices from the microphone signals, estimation of the locations of local participants, extraction of information about the surrounding acoustic environment, and finally, integration of the voices of the remote participants in the surrounding environment. This could be extended with visual representations of the remote participants (Kato and Billinghurst, 1999; Kantonen et al., 2010).

6. Conclusions

This thesis presented four studies advancing the field of audio augmented reality, as well as related fields. The studies introduced distance presentation and acoustic co-location detection methods suitable for audio augmented reality applications. The studies also investigated spatial audio guidance and mobile communication applications, evaluating the usefulness of the proposed methods and examining different perceptual aspects of mobile audio augmented reality applications.

The effect of modifying the temporal envelope of binaural room impulse responses on the perceived distance of virtual speech sources was studied. This type of modifications serves as a practical method for controlling the perceived distance of speech sources not only in audio augmented reality applications, but also in other types of applications utilizing spatial audio. This technique was applied, together with other distance presentation methods, in another study investigating the presentation of points of interest in an outdoor augmented reality environment.

Another type of spatial audio guidance, using music to guide pedestrians and cyclists to their destination, was also studied. The study utilized hear-through headphones to allow both the virtual sound source and sounds from the surroundings to be clearly heard. Finally, hear-through headphones were utilized in a mobile teleconferencing application together with acoustic co-location detection, allowing participants to hear both co-located and remote participants clearly.

The developed distance presentation methods were not applied in the investigated music guidance application or in the mobile teleconferencing application, but they could certainly be applied in such applications. However, large modifications of the reverberation when presenting the music guidance might have a negative impact on the enjoyability of the music. In

this type of application, a better approach might be the investigated alternative method of presenting the distance in meters using speech. While this approach might seem intrusive, it is able to provide exact distance information while only temporarily interrupting the music listening.

Mobile teleconferencing, on the other hand, might in some cases benefit from the developed distance presentation methods: the voices of remote teleconference participants can be presented at different distances. These distances can be used to convey information about the participants, e.g., the physical distance to the participants. By rendering the voices of the other participants with both directions and distances representing their physical location with respect to the listener, the listener can be aware of the approximate locations of the other participants. Such information would be beneficial, e.g., when communicating with co-workers at a construction site.

The main results presented in this thesis are:

- The perceived distance of virtual speech sources can be controlled effectively by modifying the ratio of early to late energy of binaural room impulse responses. The largest effect is achieved by modifying a ratio where the first 50–100 ms of the room impulse response are included in the early energy.

- Virtual sound sources presented using auditory distance cues are intuitively perceived at relatively short distances. However, people can with a small amount of training learn to map the available cues to larger distances in the surroundings. Presenting points of interest with auditory distance cues rather than by giving the distance in meters using speech decreases the time needed for the presentation and reduces auditory clutter.

- Music is a suitable audio source for spatial audio guidance in pedestrian and cyclist navigation. Both route and beacon guidance can be implemented with music and serve as effective guidance methods without prior training.

- Acoustic co-location detection can be used to improve the clarity of communication in mobile teleconferencing applications.

During the work done for this thesis, the following future research directions in the fields of audio augmented reality and virtual auditory display have emerged:

- Further studies on the effect of reverberation on distance perception should be performed. In order to determine the generalizability of the results, the experiments presented in Publication I should be extended with stimuli other than speech, other spaces, as well as other source directions. Further experiments are also needed to determine how early reflections and late reverberation affect the intensity cue to distance perception. The study of how distance is perceived in reverberant spaces in the presence of multiple simultaneous sound sources might also provide some interesting insights.

- People can quickly learn to map auditory distance cues to distances in the surroundings. The experiment in Publication II was performed with the participants sitting in one place while feedback was provided on a map. Further studies could investigate how people would learn this type of mapping by instead being allowed to walk to the presented points of interest with continuous or repeated auditory cues provided.

- The results of Publication II raise the question whether the combination of multiple auditory distance cues improves distance estimation, or if cue combination possibly worsens distance estimation under some circumstances. Further studies are thus needed to determine if and when there are benefits from cue combination.

- The integration of virtual sound sources into the surrounding acoustic environment is a topic that should be investigated. Of interest is, e.g., how closely the virtual acoustics need to match the real acoustics and which are the relevant details to produce virtual sound sources that plausibly augment the real environment. Another important area is the extraction of information about the surrounding environment in order to synthesize or choose appropriate reverberation for presentation of the virtual sound sources.

This thesis presented different methods and applications of audio augmented reality. The presented methods and other AAR techniques are useful not only in the guidance, navigation, and communication applications investigated here, but also in other applications such as entertainment and auditory display of information. As many people already use headphones and smartphones on a daily basis, audio augmented reality applications have a huge potential to help them both in everyday tasks and, e.g., when exploring unfamiliar environments. This thesis will hopefully provide both incentives and guidelines that aid in realizing this potential.

Bibliography

ALBRECHT R., LOKKI T., AND SAVIOJA L., 2011. A mobile augmented reality audio system with binaural microphones. In *Proceedings of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*, pp. 7–11, Stockholm, Sweden, doi:10.1145/2019335.2019337.

ALGAZI V.R., DUDA R.O., THOMPSON D.M., AND AVENDANO C., 2001. The CIPIC HRTF database. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 99–102, New Paltz, NY, USA, doi:10.1109/ASPAA.2001.969552.

AZUMA R.T., 1997. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments* 6(4): 355–385, doi:10.1162/pres.1997.6.4.355.

BASU S., 2002. *Conversational scene analysis*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

BEGAULT D.R. AND WENZEL E.M., 1992. Techniques and applications for binaural sound manipulation in human-machine interfaces. *The International Journal of Aviation Psychology* 2(1): 1–22, doi:10.1207/s15327108ijap0201_1.

BEGAULT D.R. AND WENZEL E.M., 1993. Headphone localization of speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 35(2): 361–376.

BEGAULT D.R., WENZEL E.M., AND ANDERSON M.R., 2001. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society* 49(10): 904–916.

BEGAULT D.R., WENZEL E.M., SHRUM R., AND MILLER J., 1996. A virtual audio guidance and alert system for commercial aircraft operations. In *Proceedings of the 3rd International Conference on Auditory Display (ICAD)*, Palo Alto, CA, USA.

BLAUERT J., 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT Press, Cambridge, MA, USA, revised ed.

BLAUERT J., BRÜGGEN M., HARTUNG K., BRONKHORST A.W., DRULLMANN R., REYNAUD G., PELLIEUX L., KREBBER W., AND SOTTEK R., 1998. The AUDIS catalog of human HRTFs. In *Proceedings of the 16th International Congress on Acoustics (ICA)*, pp. 2901–2902, Seattle, WA, USA.

BLUM J.R., BOUCHARD M., AND COOPERSTOCK J.R., 2012. What's around me? Spatialized audio augmented reality for blind users with a smartphone. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, vol. 104 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (A. Puiatti and T. Gu, editors), pp. 49–62, doi:10.1007/978-3-642-30973-1_5.

BORGES R., COSTA M., CORDIOLI J., AND ASSUITI L., 2013. An adaptive occlusion canceller for hearing aids. In *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco.

BRADLEY J.S., SATO H., AND PICARD M., 2003. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America* 113(6): 3233–3244, doi:10.1121/1.1570439.

BRADLEY J.S. AND SOULODRE G.A., 1995. The influence of late arriving energy on spatial impression. *The Journal of the Acoustical Society of America* 97(4): 2263–2271, doi:10.1121/1.411951.

BRDICZKA O., MAISONNASSE J., AND REIGNIER P., 2005. Automatic detection of interaction groups. In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI)*, pp. 32–36, Torento, Italy, doi:10.1145/1088463.1088473.

BRIMIJOIN W.O., BOYD A.W., AND AKEROYD M.A., 2013. The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE* 8(12), doi:10.1371/journal.pone.0083068.

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

BRONKHORST A.W., 2000. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* 86(1): 117–128.

BRONKHORST A.W., 2002. Modeling auditory distance perception in rooms. In *Proceedings of EAA Forum Acusticum*, Sevilla, Spain.

BRONKHORST A.W. AND HOUTGAST T., 1999. Auditory distance perception in rooms. *Nature* 397(6719): 517–520, doi:10.1038/17374.

BRUNGART D.S., 2002. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments* 11(1): 93–106, doi:10.1162/105474602317343686.

BRUNGART D.S. AND SCOTT K.R., 2001. The effects of production and presentation level on the auditory distance perception of speech. *The Journal of the Acoustical Society of America* 110(1): 425–440, doi:10.1121/1.1379730.

BRUNGART D.S., SIMPSON B.D., AND KORDIK A.J., 2005. The detectability of headtracker latency in virtual audio displays. In *Proceedings of the 11th International Conference on Auditory Display (ICAD)*, Limerick, Ireland.

CALCAGNO E.R., ABREGÚ E.L., EQUÍA M.C., AND VERGARA R., 2012. The role of vision in auditory distance perception. *Perception* 41(2): 175–192, doi:10.1068/p7153.

CALVO A., FINOMORE V., BURNETT G., AND MCNITT T., 2013. Evaluation of a mobile application for multimodal land navigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57(1): 1997–2001, doi: 10.1177/1541931213571446.

CALVO A., FINOMORE V., MCNITT T., AND BURNETT G., 2014. Demonstration and evaluation of an eyes-free mobile navigation system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58(1): 1238–1241, doi:10.1177/1541931214581258.

CARLILE S., LEONG P., AND HYAMS S., 1997. The nature and distribution of errors in sound localization by human listeners. *Hearing Research* 114(1—2): 179–196, doi:10.1016/S0378-5955(97)00161-5.

CATIC J., SANTURETTE S., AND DAU T., 2015. The role of reverberation-related binaural cues in the externalization of speech. *The Journal of the Acoustical Society of America* 138(2): 1154–1167, doi:10.1121/1.4928132.

CORMAN S.R. AND SCOTT C.R., 1994. A synchronous digital signal processing method for detecting face-to-face organizational communication behavior. *Social Networks* 16(2): 163–179, doi:10.1016/0378-8733(94)90003-5.

DASHTI M., AMIRUDDIN ABD RAHMAN M., MAHMOUDI H., AND CLAUSSEN H., 2015. Detecting co-located mobile users. In *2015 IEEE International Conference on Communications (ICC)*, pp. 1565–1570, doi: 10.1109/ICC.2015.7248547.

DAVIS S. AND MERMELSTEIN P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* 28(4): 357–366, doi:10.1109/TASSP.1980.1163420.

DE WAARD D., SCHEPERS P., ORMEL W., AND BROOKHUIS K., 2010. Mobile phone use while cycling: incidence and effects on behaviour and safety. *Ergonomics* 53(1): 30–42, doi:10.1080/00140130903381180.

DURLACH N.I., RIGOPULOS A., PANG X.D., WOODS W.S., KULKARNI A., COLBURN H.S., AND WENZEL E.M., 1992. On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments* 1(2): 251–257, doi: 10.1162/pres.1992.1.2.251.

EDWORTHY J. AND HELLIER E., 2005. Fewer but better auditory alarms will improve patient safety. *Quality and Safety in Health Care* 14(3): 212–215, doi: 10.1136/qshc.2004.013052.

ERONEN A., PELTONEN V., TUOMI J., KLAPURI A., FAGERLUND S., SORSA T., LORHO G., AND HUOPANIEMI J., 2006. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 14(1): 321–329, doi: 10.1109/TSA.2005.854103.

ETTER R. AND SPECHT M., 2005. Melodious Walkabout: implicit navigation with contextualized personal audio contents. In *Adjunct Proceedings of the 3rd International Conference on Pervasive Computing*, pp. 43–49, Munich, Germany.

FLUITT K., MERMAGEN T., AND LETOWSKI T., 2014. Auditory distance estimation in an open space. In *Soundscape Semiotics - Localization and Categorization* (H. Glotin, editor), chap. 7, pp. 135–165, InTech, Rijeka, Croatia, doi:10.5772/56137.

FUJIMOTO E. AND TURK M., 2014. Non-visual navigation using combined audio music and haptic cues. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pp. 411–418, Istanbul, Turkey, doi:10.1145/2663204.2663243.

GAMPER H., 2013. Selection and interpolation of head-related transfer functions for rendering moving virtual sound sources. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Maynooth, Ireland.

GAMPER H. AND LOKKI T., 2009. Instant BRIR acquisition for auditory events in audio augmented reality using finger snaps. In *International Workshop on the Principles and Applications of Spatial Hearing*, Zao, Japan.

GAMPER H. AND LOKKI T., 2010. Audio augmented reality in telecommunication through virtual auditory display. In *Proceedings of the 16th International Conference on Auditory Display (ICAD)*, pp. 63–71, Washington, DC, USA.

GAMPER H., TERVO S., AND LOKKI T., 2011. Head orientation tracking using binaural headset microphones. In *Audio Engineering Society Convention 131*, New York, NY, USA, paper no. 8538.

GARDNER B. AND MARTIN K., 1994. HRTF measurements of a KEMAR dummy-head microphone. `http://sound.media.mit.edu/resources/KEMAR.html`, accessed January 18, 2016.

GÓMEZ BOLAÑOS J. AND PULKKI V., 2012. HRIR database with measured actual source direction data. In *Audio Engineering Society Convention 133*, San Francisco, CA, USA, paper no. 8759.

HAAS H., 1972. The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society* 20(2): 146–159.

HARTER A., HOPPER A., STEGGLES P., WARD A., AND WEBSTER P., 2002. The anatomy of a context-aware application. *Wireless Networks* 8: 187–197, doi:10.1023/A:1013767926256.

HARTMANN W.M. AND WITTENBERG A., 1996. On the externalization of sound images. *The Journal of the Acoustical Society of America* 99(6): 3678–3688, doi:10.1121/1.414965.

HELLER F., KRÄMER A., AND BORCHERS J., 2014. Simplifying orientation measurement for mobile audio augmented reality applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 615–624, Toronto, ON, Canada, doi:10.1145/2556288.2557021.

HLÁDEK L., LE DANTEC C.C., KOPČO N., AND SEITZ A., 2013. Ventriloquism effect and aftereffect in the distance dimension. *Proceedings of Meetings on Acoustics* 19, doi:10.1121/1.4799881.

HOLLAND S., MORSE D.R., AND GEDENRYD H., 2002. AudioGPS: spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous Computing* 6(4): 253–259, doi:10.1007/s007790200025.

HOLM S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2): 65–70.

HUSSAIN I., CHEN L., MIRZA H.T., XING K., AND CHEN G., 2014. A comparative study of sonification methods to represent distance and forward-direction in pedestrian navigation. *International Journal of Human-Computer Interaction* 30(9): 740–751, doi:10.1080/10447318.2014.925381.

HÄRMÄ A., JAKKA J., TIKANDER M., KARJALAINEN M., LOKKI T., HIIPAKKA J., AND LORHO G., 2004. Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society* 52(6): 618–639.

HÄRMÄ A., JAKKA J., TIKANDER M., KARJALAINEN M., LOKKI T., AND NIRONEN H., 2003. Techniques and applications of wearable augmented reality audio. In *Audio Engineering Society Convention 114*, Amsterdam, the Netherlands, paper no. 5768.

HÄRMÄ A., VAN DINTHER R., SVEDSTRÖM T., PARK M., AND KOPPENS J., 2012. Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface. In *Audio Engineering Society Convention 132*, Budapest, Hungary, paper no. 8644.

JENSEN B.S., SKOV M.B., AND THIRURAVICHANDRAN N., 2010. Studying driver attention and behaviour for three configurations of GPS navigation in real traffic driving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1271–1280, Atlanta, GA, USA, doi:10.1145/1753326.1753517.

JEUB M., SCHÄFER M., KRÜGER H., NELKE C., BEAUGEANT C., AND VARY P., 2010. Do we need dereverberation for hand-held telephony? In *Proceedings of the 20th International Congress on Acoustics (ICA)*, Sydney, Australia.

JEUB M., SCHÄFER M., AND VARY P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of the 16th International Conference on Digital Signal Processing*, Santorini, Greece, doi:10.1109/ICDSP.2009.5201259.

JONES M., JONES S., BRADLEY G., WARREN N., BAINBRIDGE D., AND HOLMES G., 2008. ONTRACK: dynamically adapting music playback to support navigation. *Personal and Ubiquitous Computing* 12(7): 513–525, doi:10.1007/s00779-007-0155-2.

KANTONEN T., WOODWARD C., AND KATZ N., 2010. Mixed reality in virtual world teleconferencing. In *2010 IEEE Virtual Reality Conference*, pp. 179–182, doi:10.1109/VR.2010.5444792.

KATO H. AND BILLINGHURST M., 1999. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR)*, pp. 85–94, doi:10.1109/IWAR.1999.803809.

KLATZKY R.L., MARSTON J.R., GIUDICE N.A., GOLLEDGE R.G., AND LOOMIS J.M., 2006. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied* 12(4): 223–232, doi:10.1037/1076-898X.12.4.223.

KOLARIK A.J., MOORE B.C.J., ZAHORIK P., CIRSTEA S., AND PARDHAN S., 2015. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics* 78(2): 373–395, doi:10.3758/s13414-015-1015-1.

KRUMM J. AND HINCKLEY K., 2004. The NearMe wireless proximity server. In *6th International Conference on Ubiquitous Computing (UbiComp)*, pp. 283–300, Nottingham, UK.

LARSEN C.H., LAURITSEN D.S., LARSEN J.J., PILGAARD M., AND MADSEN J.B., 2013. Differences in human audio localization performance between a HRTF- and a non-HRTF audio system. In *Proceedings of the 8th Audio Mostly Conference*, Piteå, Sweden, doi:10.1145/2544114.2544118.

LARSEN E., IYER N., LANSING C.R., AND FENG A.S., 2008. On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America* 124(1): 450–461, doi:10.1121/1.2936368.

LICHENSTEIN R., SMITH D.C., AMBROSE J.L., AND MOODY L.A., 2012. Headphone use and pedestrian injury and death in the united states: 2004–2011. *Injury Prevention* 18(5): 287–290, doi:10.1136/injuryprev-2011-040161.

LILJEDAHL M. AND LINDBERG S., 2011. Sound parameters for expressing geographic distance in a mobile navigation application. In *Proceedings of the 6th Audio Mostly Conference*, pp. 1–7, Coimbra, Portugal, doi:10.1145/2095667.2095668.

LINDAU A., 2009. The perception of system latency in dynamic binaural synthesis. In *Proceedings of the NAG/DAGA International Conference on Acoustics*, pp. 1063–1066, Rotterdam, the Netherlands.

LINDAU A. AND BRINKMANN F., 2012. Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. *Journal of the Audio Engineering Society* 60(1/2): 54–62.

LINDAU A., KOSANKE L., AND WEINZIERL S., 2010. Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses. In *Audio Engineering Society Convention 128*, London, UK.

LINDAU A. AND WEINZIERL S., 2012. Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica* 98(5): 804–810, doi:10.3813/AAA.918562.

LINDEMAN R.W., NOMA H., AND DE BARROS P.G., 2007. Hear-through and mic-through augmented reality: using bone conduction to display spatialized audio. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, Nara, Japan, doi:10.1109/ISMAR.2007.4538843.

LISKI J., 2016. *Adaptive hear-through headset*. Master's thesis, Aalto University School of Electrical Engineering, Espoo, Finland.

LITOVSKY R.Y., COLBURN H.S., YOST W.A., AND GUZMAN S.J., 1999. The precedence effect. *The Journal of the Acoustical Society of America* 106(4): 1633–1654, doi:10.1121/1.427914.

LOKKI T., NIRONEN H., VESA S., SAVIOJA L., AND HÄRMÄ A., 2004. Problem of far-end user's voice in binaural telephony. In *Proceedings of the 18th International Congress on Acoustics (ICA)*, Kyoto, Japan.

LOKKI T., PÄTYNEN J., TERVO S., SILTANEN S., AND SAVIOJA L., 2011. Engaging concert hall acoustics is made up of temporal envelope preserving reflections. *The Journal of the Acoustical Society of America* 129(6): EL223–EL228, doi:10.1121/1.3579145.

LOOMIS J.M., GOLLEDGE R.G., AND KLATZKY R.L., 1998. Navigation system for the blind: auditory display modes and guidance. *Presence: Teleoperators and Virtual Environments* 7(2): 193–203, doi:10.1162/105474698565677.

MANDAL A., LOPES C., GIVARGIS T., HAGHIGHAT A., JURDAK R., AND BALDI P., 2005. Beep: 3D indoor positioning using audible sound. In *2nd IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 348–353, Las Vegas, NV, USA, doi:10.1109/CCNC.2005.1405195.

MARIETTE N., 2007. Mitigation of binaural front-back confusions by body motion in audio augmented reality. In *Proceedings of the 13th International Conference on Auditory Display (ICAD)*, pp. 38–44, Montreal, Canada.

MARTIN A., JIN C., AND VAN SCHAIK A., 2009. Psychoacoustic evaluation of systems for delivering spatialized augmented-reality audio. *Journal of the Audio Engineering Society* 57(12): 1016–1027.

MCMULLEN K., ROGINSKA A., AND WAKEFIELD G.H., 2012. Subjective selection of head-related transfer functions (HRTF) based on spectral coloration and interaural time differences (ITD) cues. In *Audio Engineering Society Convention 133*, San Francisco, CA, USA, paper no. 8770.

MEJIA J., DILLON H., AND FISHER M., 2008. Active cancellation of occlusion: an electronic vent for hearing aids and hearing protectors. *The Journal of the Acoustical Society of America* 124(1): 235–240, doi:10.1121/1.2908279.

MIDDLEBROOKS J.C. AND GREEN D.M., 1991. Sound localization by human listeners. *Annual Review of Psychology* 42: 135–159, doi:10.1146/annurev.ps.42.020191.001031.

MIURA M., WATANABE H., KAWAI K., AND KONDO K., 2013. Intelligibility comparison of speech annotation under wind noise in AAR systems for pedestrians and cyclists using two output devices. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kaohsiung, Taiwan, doi:10.1109/APSIPA.2013.6694370.

MOUSTAKAS N., FLOROS A., AND KANELLOPOULOS N., 2009. Eidola: an interactive augmented reality audio-game prototype. In *Audio Engineering Society Convention 127*, paper no. 7872.

MØLLER A.K., CHRISTENSEN F., HOFFMANN P.F., AND HAMMERSHØI D., 2015. Directional characteristics for different in-ear recording points. In *58th International Conference of the Audio Engineering Society*, Aalborg, Denmark.

MØLLER H., JENSEN C.B., HAMMERSHØI D., AND SØRENSEN M.F., 1995. Design criteria for headphones. *Journal of the Audio Engineering Society* 43(4): 218–232.

NASAR J., HECHT P., AND WENER R., 2008. Mobile telephones, distracted attention, and pedestrian safety. *Accident Analysis & Prevention* 40(1): 69–75, doi:10.1016/j.aap.2007.04.005.

NIELSEN S.H., 1993. Auditory distance perception in different rooms. *Journal of the Audio Engineering Society* 41(10): 755–770.

PATERSON N., NALIUKA K., JENSEN S.K., CARRIGY T., HAAHR M., AND CONWAY F., 2010. Design, implementation and evaluation of audio for a location aware augmented reality game. In *Proceedings of the 3rd International Conference on Fun and Games*, pp. 149–156, Leuven, Belgium, doi: 10.1145/1823818.1823835.

PELZER S., ASPÖCK L., SCHRÖDER D., AND VORLÄNDER M., 2014. Interactive real-time simulation and auralization for modifiable rooms. *Building Acoustics* 21(1): 65–73, doi:10.1260/1351-010X.21.1.65.

PIELOT M., POPPINGA B., HEUTEN W., AND BOLL S., 2012. Tacticycle: supporting exploratory bicycle trips. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 369–378, San Francisco, CA, USA, doi:10.1145/2371574.2371631.

PLENGE G., 1974. On the differences between localization and lateralization. *The Journal of the Acoustical Society of America* 56(3): 944–951, doi: 10.1121/1.1903353.

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 Transactions on Audio, Speech, and Language Processing* 17(6): 1124–1132, doi:10.1109/TASL.2009.2020532.

REYNOLDS D., QUATIERI T., AND DUNN R., 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10(1): 19–41, doi: 10.1006/dspr.1999.0361.

ROMIGH G.D., BRUNGART D.S., AND SIMPSON B.D., 2015. Free-field localization accuracy with a head-tracked virtual auditory display. *IEEE Journal of Selected Topics in Signal Processing* 9(5): 943–954, doi: 10.1109/JSTSP.2015.2421874.

ROSSING T.D., MOORE R., AND WHEELER P., 2002. *The science of sound*. Addison Wesley, San Francisco, CA, USA, 3rd ed.

RÄMÖ J. AND VÄLIMÄKI V., 2012. Digital augmented reality audio headset. *Journal of Electrical and Computer Engineering* 2012, doi:10.1155/2012/457374, article ID 457374.

SANDBERG S., HÅKANSSON C., ELMQVIST N., TSIGAS P., AND CHEN F., 2006. Using 3D audio guidance to locate indoor static objects. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50(16): 1581–1584, doi: 10.1177/154193120605001614.

SANDVAD J., 1996. Dynamic aspects of auditory virtual environments. In *Audio Engineering Society Convention 100*, Copenhagen, Denmark, paper no. 4226.

SARTER N.B., 2006. Multimodal information presentation: design guidance and research challenges. *International Journal of Industrial Ergonomics* 36(5): 439–445, doi:10.1016/j.ergon.2006.01.007.

SCHÄRER Z. AND LINDAU A., 2009. Evaluation of equalization methods for binaural signals. In *Audio Engineering Society Convention 126*, Munich, Germany, paper no. 7721.

SCHÖNSTEIN D. AND KATZ B.F., 2012. Variability in perceptual evaluation of HRTFs. *Journal of the Audio Engineering Society* 60(10): 783–793.

SEEBER B.U. AND FASTL H., 2003. Subjective selection of non-individual head-related transfer functions. In *Proceedings of the 9th International Conference on Auditory Display (ICAD)*, Boston, MA, USA.

SHINN-CUNNINGHAM B.G., 1998. Applications of virtual auditory displays. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pp. 1105–1108, doi: 10.1109/IEMBS.1998.747064.

SHINN-CUNNINGHAM B.G., 2000. Distance cues for virtual auditory space. In *Proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia*, pp. 227–230, Sydney, Australia.

SHINN-CUNNINGHAM B.G., DURLACH N.I., AND HELD R.M., 1998. Adapting to supernormal auditory localization cues. I. Bias and resolution. *The Journal of the Acoustical Society of America* 103(6): 3656–3666, doi:10.1121/1.423088.

SHINN-CUNNINGHAM B.G., SANTARELLI S., AND KOPCO N., 2000. Tori of confusion: binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America* 107(3): 1627–1636, doi: 10.1121/1.428447.

SIVONEN V.P. AND ELLERMEIER W., 2006. Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation. *The Journal of the Acoustical Society of America* 119(5): 2965–2980, doi: 10.1121/1.2184268.

SKOVENBORG E. AND NIELSEN S.H., 2004. Evaluation of different loudness models with music and speech material. In *Audio Engineering Society Convention 117*, San Francisco, CA, USA, paper no. 6234.

SOULODRE G.A. AND BRADLEY J.S., 1995. Subjective evaluation of new room acoustic measures. *The Journal of the Acoustical Society of America* 98(1): 294–301, doi:10.1121/1.413735.

SPAGNOL S., GERONAZZO M., AND AVANZINI F., 2012. Hearing distance: a low-cost model for near-field binaural effects. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2030–2034, Bucharest, Romania.

SPAGNOL S., GERONAZZO M., ROCCHESSO D., AND AVANZINI F., 2014. Synthetic individual binaural audio delivery by pinna image processing. *International Journal of Pervasive Computing and Communications* 10(3): 239–254, doi:10.1108/IJPCC-06-2014-0035.

STENFELT S. AND REINFELDT S., 2007. A model of the occlusion effect with bone-conducted stimulation. *International Journal of Audiology* 46(10): 595–608, doi:10.1080/14992020701545880.

STOREK D., STUCHLIK J., AND RUND F., 2015. Modifications of the surrounding auditory space by augmented reality audio: introduction to warped acoustic reality. In *Proceedings of the 21st International Conference on Auditory Display (ICAD)*, pp. 225–230, Graz, Austria.

STRACHAN S., ESLAMBOLCHILAR P., MURRAY-SMITH R., HUGHES S., AND O'MODHRAIN S., 2005. GpsTunes: controlling navigation via audio feedback. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 275–278, Salzburg, Austria, doi:10.1145/1085777.1085831.

SUNOHARA M., WATANUKI K., AND TATENO M., 2014. Occlusion reduction system for hearing aids using active noise control technique. *Acoustical Science and Technology* 35(6): 318–320, doi:10.1250/ast.35.318.

TAKANEN M. AND KARJALAINEN M., 2010. Real-time tracking of speech sources using binaural audio and orientation tracking. In *40th International Conference of the Audio Engineering Society*, Tokyo, Japan.

TALBOT M. AND COWAN W., 2009. On the audio representation of distance for blind users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1839–1848, Boston, MA, USA, doi: 10.1145/1518701.1518984.

THOMPSON L.L., RIVARA F.P., AYYAGARI R.C., AND EBEL B.E., 2013. Impact of social and technological distraction on pedestrian crossing behaviour: an observational study. *Injury Prevention* 19(4): 232–237, doi:10.1136/injuryprev-2012-040601.

TIKANDER M., 2009. Usability issues in listening to natural sounds with an augmented reality audio headset. *Journal of the Audio Engineering Society* 57(6): 430–441.

TIKANDER M., KARJALAINEN M., AND RIIKONEN V., 2008. An augmented reality audio headset. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx)*, Espoo, Finland.

VAZQUEZ-ALVAREZ Y., AYLETT M.P., BREWSTER S.A., VON JUNGENFELD R., AND VIROLAINEN A., 2016. Designing interactions with multilevel auditory displays in mobile audio-augmented reality. *ACM Transactions on Computer-Human Interaction* 23(1): 3:1–3:30, doi:10.1145/2829944.

VAZQUEZ-ALVAREZ Y., OAKLEY I., AND BREWSTER S.A., 2012. Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous Computing* 16(8): 987–999, doi:10.1007/s00779-011-0459-0.

VESA S. AND HÄRMÄ A., 2005. Automatic estimation of reverberation time from binaural signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3:281–3:284, Philadelphia, PA, USA.

VÄÄNÄNEN R., VESA S., AND HÄMÄLÄINEN M., 2014. Testing the user experience of an augmented reality headset and 3D audio-guided pedestrian navigation. In *55th International Conference of the Audio Engineering Society*, Helsinki, Finland.

WALKER E.J., LANTHIER S.N., RISKO E.F., AND KINGSTONE A., 2012. The effects of personal music devices on pedestrian behaviour. *Safety Science* 50(1): 123–128, doi:10.1016/j.ssci.2011.07.011.

WARUSFEL O., 2003. Listen HRTF database. `http://recherche.ircam.fr/equipes/salles/listen/index.html`, accessed January 18, 2016.

WENZEL E.M., 1992. Localization in virtual acoustic displays. *Presence: Teleoperators and Virtual Environments* 1(1): 80–107, doi:10.1162/pres.1992.1.1.80.

WENZEL E.M., 2001. Effect of increasing system latency on localization of virtual sounds with short and long duration. In *Proceedings of the 7th International Conference on Auditory Display (ICAD)*, Espoo, Finland.

WENZEL E.M., ARRUDA M., KISTLER D.J., AND WIGHTMAN F.L., 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94(1): 111–123, doi:10.1121/1.407089.

WESTPHAL M., 1997. The use of cepstral means in conversational speech recognition. In *5th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1143–1146, Rhodes, Greece.

WIERSTORF H., GEIER M., AND SPORS S., 2011. A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *Audio Engineering Society Convention 130*, London, UK.

WIGHTMAN F.L. AND KISTLER D.J., 1999. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105(5): 2841–2853, doi:10.1121/1.426899.

WYATT D., CHOUDHURY T., BILMES J., AND KITTS J.A., 2011. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Transactions on Intelligent Systems and Technology* 2: 7:1–7:41, doi:10.1145/1889681.1889688.

ZAHORIK P., 2002. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America* 112(5): 2110–2117, doi:10.1121/1.1506692.

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(3): 409–420.

ZIMMERMANN A. AND LORENZ A., 2008. LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction* 18(5): 389–416, doi:10.1007/s11257-008-9049-x.

ZWINDERMAN M., ZAVIALOVA T., TETTEROO D., AND LEHOUCK P., 2011. Oh music, where art thou? In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 533–538, Stockholm, Sweden, doi:10.1145/2037373.2037456.

Errata

Publication I

The fitting of the distance models

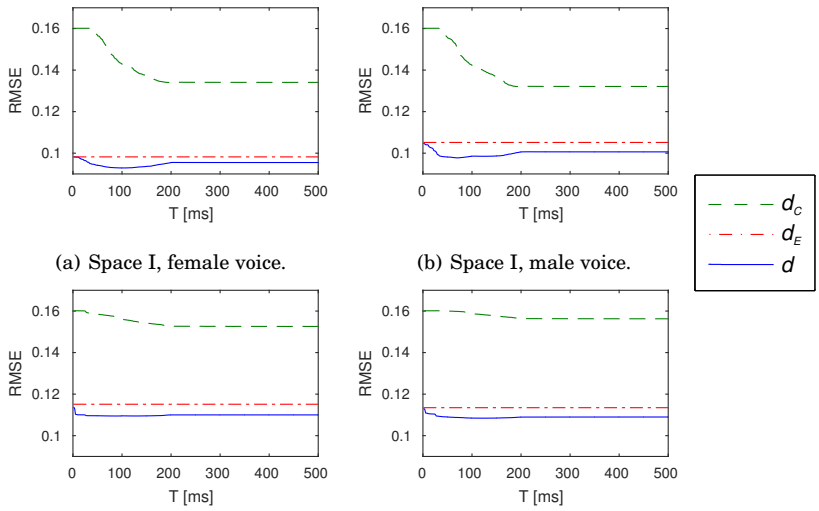
$$d_C = a \cdot \left(\frac{1}{C_T}\right)^k,$$

$$d_E = b \cdot \left(\frac{1}{\bar{E}}\right)^l,$$

and

$$d = c \cdot \left(\frac{1}{C_T}\right)^k \cdot \left(\frac{1}{\bar{E}}\right)^l$$

to the distance estimates erroneously allowed negative values of the exponents k and l. Negative exponents invert $\frac{1}{C_T}$ and $\frac{1}{\bar{E}}$, which means that increasing the early-to-late energy ratio or increasing the energy would lead to an increase in the perceived distance. The fitting of the models to the data produced positive exponents in most cases, but for the fitting of d_C for modification set B it produced negative exponents for small values of T, which provided a better fit than the perceptually motivated positive exponents. The corrected errors when fitting the models without allowing negative exponents are shown in Fig. E.1.



(a) Space I, female voice.

(b) Space I, male voice.

(c) Space II, female voice.

(d) Space II, male voice.

Figure E.1. Root-mean-square error (RMSE) when fitting d_C (dashed line), d_E (dash-dot line), and d (solid line) to the distance estimates of modification set B, shown for different values of T.

Aalto-DD 122/2016

ISBN 978-952-60-6875-6 (printed)
ISBN 978-952-60-6876-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**