TKK Dissertations in Media Technology
Espoo 2009 TKK-ME-D-1

STUDIES ON BINAURAL AND MONAURAL SIGNAL ANALYSIS —METHODS AND APPLICATIONS

Sampo Vesa

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences, for public examination and debate in Auditorium T1 at Helsinki University of Technology (Espoo, Finland) on the 4th of December, 2009, at 12 noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Media Technology

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Mediatekniikan laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Media Technology
P.O.Box 5400
FIN-02015 TKK
Finland
Tel. +358-9-470 22870
Fax. +358-9-470 25014
http://media.tkk.fi/

Available in PDF format at http://lib.tkk.fi/Diss/2009/isbn9789522482396/

ISBN 978-952-248-238-9 (print)
ISBN 978-952-248-239-6 (online)
ISSN 1797-7096 (print)
ISSN 1797-710X (online)

Yliopistopaino
Helsinki 2009

ABSTRACT

Author Sampo Vesa
Title Studies on Binaural and Monaural Signal Analysis — Methods and Applications

Sound signals can contain a lot of information about the environment and the sound sources present in it. This thesis presents novel contributions to the analysis of binaural and monaural sound signals. Some new applications are introduced in this work, but the emphasis is on analysis methods. The three main topics of the thesis are computational estimation of sound source distance, analysis of binaural room impulse responses, and applications intended for augmented reality audio.

A novel method for binaural sound source distance estimation is proposed. The method is based on learning the coherence between the sounds entering the left and right ears. Comparisons to an earlier approach are also made. It is shown that these kinds of learning methods can correctly recognize the distance of a speech sound source in most cases.

Methods for analyzing binaural room impulse responses are investigated. These methods are able to locate the early reflections in time and also to estimate their directions of arrival. This challenging problem could not be tackled completely, but this part of the work is an important step towards accurate estimation of the individual early reflections from a binaural room impulse response.

As the third part of the thesis, applications of sound signal analysis are studied. The most notable contributions are a novel eyes-free user interface controlled by finger snaps, and an investigation on the importance of features in audio surveillance.

The results of this thesis are steps towards building machines that can obtain information on the surrounding environment based on sound. In particular, the research into sound source distance estimation functions as important basic research in this area. The applications presented could be valuable in future telecommunications scenarios, such as augmented reality audio.

UDC 534.8, 621.39, 004.934, 004.85
Keywords audio signal analysis, audio signal processing, augmented reality audio, binaural signals, sound source distance, room impulse responses, reverberation time, eyes-free user interfaces, audio surveillance

TIIVISTELMÄ

Tekijä Sampo Vesa
Työn nimi Binauraalisten ja monauraalisten signaalien analyysimenetelmiä ja niiden sovelluksia

Äänisignaalit sisältävät paljon tietoa ympäristöstä ja siinä olevista äänilähteistä. Tässä väitöskirjassa esitetään uusia menetelmiä binauraalisten ja monauraalisten äänisignaalien analysointiin. Lisäksi tutkitaan sovelluksia, jotka hyödyntävät äänisignaaleista saatua informaatiota. Väitöskirjan kolme pääaihetta ovat äänilähteen etäisyyden laskennallinen estimointi, binauraalisten huonevasteiden analyysi ja lisättyyn äänitodellisuuteen liittyvät sovellukset.

Väitöskirjassa esitetään uusi binauraalinen menetelmä äänilähteen etäisyyden estimointiin. Menetelmä perustuu korvien äänisignaalien välisen koherenssin oppimiseen. Esitettyä menetelmää verrataan kirjallisuudesta löytyvään aikaisempaan menetelmään. Tulokset osoittavat, että oppimiseen perustuvilla paikannusmenetelmillä on mahdollista tunnistaa puheäänilähteen etäisyys useimmissa tapauksissa.

Väitöstyössä tutkitaan myös binauraalisten huonevasteiden analyysiä. Esitetyillä menetelmillä voidaan selvittää varhaisten heijastusten saapumisajat sekä myös estimoida niiden tulokulmat. Tätä haastavaa ongelmaa ei saatu täysin ratkaistua, mutta tämä väitöskirjan osa on tärkeä askel kohti tarkkaa yksittäisten varhaisten heijastusten estimointia binauraalisista huonevasteista.

Väitöskirjan kolmannessa osassa tutkitaan äänisignaalien analyysin sovelluksia. Tärkeimpiä tuloksia ovat uusi sormien napsutuksella ohjattava ei-visuaalinen käyttöliittymä ja äänivalvonnassa käytettävien piirteiden painoarvot.

Tämän väitöskirjan tulokset ovat askelia kohti älykkäitä kuulevia koneita. Varsinkin etäisyyden estimointiin liittyvä tutkimus on tärkeää perustutkimusta tällä alueella. Esitetyt sovellukset voivat olla hyödyllisiä tulevaisuuden matkaviestintäskenaarioissa, kuten lisätyssä äänitodellisuudessa.

UDK 534.8, 621.39, 004.934, 004.85
Avainsanat äänisignaalin analyysi, digitaalinen äänenkäsittely, lisätty äänitodellisuus, binauraaliset signaalit, äänilähteen etäisyys, huonevasteet, jälkikaiunta-aika, ei-visuaaliset käyttöliittymät, äänivalvonta

PREFACE

This research was carried out during 2005–2009 in the Telecommunications Software and Multimedia Laboratory, Helsinki University of Technology, Espoo, Finland. This work has been supported by Helsinki Graduate School on Computer Science and Engineering (Hecse), Nokia Research Center, Nokia Foundation, Tekniikan Edistämissäätiö, the Academy of Finland, project no. 119092, and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 203636.

I am grateful to my supervisor Prof. Lauri Savioja and instructor Dr. Tapio Lokki for their support, guidance, and positive attitude. Tapio's help in reading the manuscripts and providing valuable feedback has been invaluable for the publications in this thesis. I also wish to thank my support group — Prof. Matti Karjalainen and Dr. Kalle Palomäki — for feedback on some of the publications in this thesis.

I thank the pre-examiners of this thesis Assoc. Prof. Daniel Ellis and Dr. Jyri Huopaniemi for invaluable comments on the manuscript. Also thanks to Lic. Tech. Luis de Jussilainen Costa and Mr. William Martin for proofreading the manuscript.

Thanks go to my fellow co-workers from Akulab, especially Mr. Miikka Tikander and Dr. Ville Pulkki. Miikka and Ville helped with recording some of the test recordings used throughout this thesis. The collaboration and sharing of ideas in the KAMARA project (with all its different incarnations) has also been valuable during the years. Many of the ideas in this thesis stem from this project, in which I was involved between 2002 and 2008.

I also wish to thank Dr. Aki Härmä for co-authoring one of the publications in this thesis and arranging me a research visit to Philips Research in Eindhoven. I also thank Dr. Steven van de Par for guidance and collaborations during the research visit. The visit proved to be an excellent opportunity to get a taste of the research taking place in a major company, and it was also a really fun summer otherwise.

Special thanks go to my colleague Mr. Sakari Tervo for fascinating discussions on the research topics. Ms. Liisa Hirvisalo deserves thanks for taking care of practical things related to printing of this thesis. Other members of support staff of the Department of Media Technology have also provided invaluable help in practical things. My co-workers and room mates (in order of appearance, since the beginning of 2005 until early 2008) Iikka, Janne, Samuel, Raine, Janne, and Jukka — thanks for the relaxed work atmosphere! I'd also wish to thank other colleagues at the Department of Media Technology / Telecommunications Software and Multimedia Laboratory. This has been a great place to work at!

Finally, I wish to express gratitude towards my family for support during the making of this thesis.

Otaniemi, Espoo, 16th November 2009

Sampo Vesa

CONTENTS

Abstract 1

Tiivistelmä 3

Preface 5

Contents 7

List of Publications 9

List of Abbreviations 11

1 Introduction 13
- 1.1 Scope of this thesis . . . 14
- 1.2 Organization of the thesis . . . 16

2 Background 17
- 2.1 Augmented reality audio . . . 17
- 2.2 Relationship between ARA and this thesis . . . 20

3 Computational sound source distance estimation 23
- 3.1 The problem of sound source distance estimation . . . 23
- 3.2 Sound source distance perception cues . . . 24
  - Sound pressure level . . . 24
  - Spectral cues . . . 24
  - Direct-to-reverberant ratio . . . 25
  - Dynamic cues . . . 26
  - The effect of the environment . . . 26
  - Absolute and relative cues . . . 26
- 3.3 Human performance in distance estimation . . . 27
- 3.4 Related research . . . 28
- 3.5 Binaural coherence . . . 30
- 3.6 Novel contributions . . . 31

4 Analysis of binaural room impulse responses 33
- 4.1 The problem of detection and localization of early reflections 33
- 4.2 Related research on analysis of early reflections . . . 34
- 4.3 Continuous wavelet transform . . . 36
- 4.4 Related research on applications of wavelets to RIR analysis 36
- 4.5 Novel contributions . . . 37
  - Estimation of reflection arrival times from a BRIR [P3] . . . 37
  - Segmentation of reflections from a BRIR [P4] . . . 38

5 Applications to augmented reality audio 41
- 5.1 Binaural blind estimation of reverberation time for adjusting the reverberation of virtual sources . . . 42

Related research 42
Novel contributions 44
5.2 Eyes-free user interface based on finger snaps 44
Related research 44
Novel contributions 45
5.3 Features in audio surveillance 45
Related research 45
Novel contributions 46

6 Summary of Publications and Contributions of the Author 47

7 Conclusions 49

Bibliography 51

Errata 63

LIST OF PUBLICATIONS

This thesis summarizes the following articles and publications, referred to as [P1]–[P7]:

[P1] S. Vesa. Sound Source Distance Learning Based on Binaural Signals. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, pp. 271–274, New Paltz, NY, USA, October 21–24, 2007.

[P2] S. Vesa. Binaural Sound Source Distance Learning in Rooms. In *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498–1507, November 2009.

[P3] S. Vesa and T. Lokki. Detection of Room Reflections from a Binaural Room Impulse Response. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pp. 215–220, Montreal, Canada, September 18–20, 2006.

[P4] S. Vesa and T. Lokki. Segmentation and Analysis of Early Reflections from a Binaural Room Impulse Response. In L. Hirvisalo (ed.): *TKK Reports in Media Technology*, Technical Report TKK-ME-R-1, Department of Media Technology, Helsinki University of Technology, 2009.

[P5] S. Vesa and A. Härmä. Automatic Estimation of Reverberation Time from Binaural Signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 3, pp. 281–284, Philadelphia, PA, USA, March 18–23, 2005.

[P6] S. Vesa and T. Lokki. An Eyes-Free User Interface Controlled by Finger Snaps. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, pp. 262–265, Madrid, Spain, September 20–22, 2005.

[P7] S. Vesa. The Effect of Features on Clustering in Audio Surveillance. In *Proceedings of the AES 30th International Conference on Intelligent Audio Environments*, Saariselkä, Finland, March 15–17, 2007.

LIST OF ABBREVIATIONS

3-D	Three-dimensional
ARA	Augmented reality audio
BRIR	Binaural room impulse response
CC	Cross correlation
CWT	Continuous wavelet transform
DRR	Direct-to-reverberant ratio
DWT	Discrete wavelet transform
EC	Equalization cancellation
EDT	Early decay time
HRTF	Head-related transfer function
GA	Genetic algorithm
GCC	Generalized cross correlation
GMM	Gaussian mixture model
GPS	Global positioning system
IACC	Interaural cross-correlation coefficient
IC	Interaural coherence
ILD	Interaural level difference
ITD	Interaural time difference
JND	Just noticeable difference
LEF	Lateral energy fraction
MARA	Mobile augmented reality audio
ML	Maximum likelihood
MSC	Magnitude-squared coherence
PCA	Principal component analysis
RT	Reverberation time
STFT	Short-time Fourier transform
WARA	Wearable augmented reality audio
XWT	Continuous cross-wavelet transform

1 INTRODUCTION

Every day we are bombarded with sounds. Our hearing system is constantly analyzing the stream of sound entering the ear canals. Based on our hearing abilities, we can know the direction where the sound originates, how far away is its source, and what is the most likely cause of the sound. If the sound is speech, we can then attribute specific semantic meaning to series of sounds that, by themselves, do not carry any specific meaning. Even when blindfolded, we can simultaneously get some idea of the space we are in — is it a cathedral, a small closet, or are we outside on a field? When not in excessively noisy conditions, this analysis happens effortlessly. It is easy to take all of this for granted. But, when examined more closely, it becomes apparent that complex processing on several different levels by our sensory and nervous system is necessary.

When the sound has been converted to neural impulses by the auditory system, some mechanism is needed for analyzing the information received. The brain has to make some sense of the sensory information. If there are many sound sources present in the environment, how can they be separated and grouped so that it is possible to recognize their causes? The parts of the sound signal that originate from the same object in the environment have to be grouped together. This is the problem that a field called *auditory scene analysis* attempts to solve [18]. This processing is the front end that is required for the higher processing stages to get a meaningful perception of the complex sound environment.

In addition to allowing intra-species communication, it is likely that hearing in humans — and other mammals — has also developed due to the possibility of detecting impending danger by the sound that it causes [46], before our visual senses can detect the possible threat. This is possible, because sound waves bend around physical obstacles by a phenomenon called diffraction. Even if the cause of the potential threat is in the field of vision, hearing becomes useful when it is too dark for the visual senses to pick up enough information from the environment.

Recognizing the most likely cause of the sound, e.g. a lion roaring, has been vital for survival. But knowing the cause of the sound has not been enough, because whether or not and what kind of action needs to be taken when hearing a sound is dependent also on the spatial location of the sound source. The process of estimating the location where a sound originates is termed *sound source localization* [124]. Knowing the location by hearing modality first makes it possible to direct the eyes towards the source of the sound. Also, localization based on sound, without having to turn the head, permits localization of short sound events that would otherwise be missed [66]. Knowing the location by hearing modality first also makes it possible to direct the eyes towards the source of the sound.

This thesis deals with analysis of sounds using a computer. The idea is to gather information on the environment by analyzing sounds recorded by microphones present in the environment. The microphones can be stationary sensors or they can be moving along with the user of some mobile device.

One way of approaching the analysis task is by constructing computational auditory models that mimic the auditory system of humans as faithfully and accurately as possible. Often these models act as front-ends that produce such features that the higher stages of the auditory system then analyze. This thesis takes an engineering-oriented approach, where the auditory system is not accurately modeled. Instead, the focus is on solving the analysis tasks using different means — some of which can be auditorily motivated. The emphasis is on methods that are suitable for real-time processing. Many of the methods presented in this thesis have been implemented in real-time as part of the work.

1.1 Scope of this thesis

This thesis is concerned both with analysis of sound source content and its location in space, as well as properties of the space itself. The analysis is performed on sound signals recorded from the entrances of the ear canals of humans. This binaural, two-channel signal contains the same information that is available to the human auditory system. All of the publications in the thesis (with the exception of publication [P7]) are based on a scenario where a binaural sound signal recorded from the ears of the user is available for analysis. The information obtained on the three mentioned aspects of the audio environment can be useful in future telecommunication applications and hearing aids. A few applications are, therefore, also presented as a part of the thesis.

More specifically, this thesis concentrates on certain subtopics and applications of sound signal analysis. These topics are, in order of importance, the following:

- **Binaural estimation of sound source distance**
  Publications [P1] and [P2] present approaches to computational estimation of the line-of-sight distance between a sound source and a listener inside a room. One of the main contributions of this thesis is the proposal of a new machine learning approach for sound source distance estimation, and its comparison to a previously presented binaural localization method [146] in Publication [P2].

- **Analysis of binaural room impulse responses**
  Publications [P3] and [P4] present novel wavelet-based methods for the analysis of binaural room impulse responses (BRIRs). Publication [P3] presents a simple method for time-localizing reflections in a BRIR. The method is extended in publication [P4] to include segmentation of the reflections in time and frequency, making it possible to estimate the azimuth angle (the angle of arrival in the lateral plane) of the reflections.

- **Applications to augmented reality audio**
  Three different applications of sound signal analysis to augmented reality audio (ARA) are presented in this thesis (see Section 2.1 for a brief introduction to ARA). The first of these is a real-time method for blind estimation of the reverberation time (RT) based on binaural signals [P5]. The method can be used to adjust the reverberation times

Table 1.1: Classification of the publications of this thesis.

sound source location	environment properties	sound source content
distance (speech) [P1]	RT [P5]	finger snaps [P6]
distance (speech) [P2]	reflections [P3]	short audio events [P7]
azimuth (transients) [P6]	reflections [P4]	

in auralization of virtual sources within the ARA context. The second application is an eyes-free user interface that is controlled by binaurally detected and localized finger snaps [P6]. The interface is used to control the play list of music player software. The third application is audio surveillance — automatic monitoring of the environment based on sound, using pattern recognition techniques. Publication [P7] describes a study of finding the optimal set of features for audio surveillance purposes.

Another way to categorize the topics of the thesis is the type of information that is obtained from analysis of audio signals. Table 1.1 classifies the publications into the following categories:

- **Sound source location**
  The publications relating to sound source distance estimation ([P1], [P2]) are concerned on the location of the sound source, as the position of a sound source relative to a listener can be described by two angles and a distance. The eyes-free user interface ([P6]) is based on estimating the azimuth angles of finger snaps and mapping the sectors that the angles belong to into commands for the program or device that is controlled with the interface, for example an MP3 player.

- **Environment properties**
  Publication [P5] is about blind estimation of reverberation time. The reverberation time is a property of the room (environment), which characterizes the rate of sound energy decay in the room. The BRIR analysis method presented in publications [P3] and [P4] finds the times and directions of arrival of early reflections in a room. Therefore, that part of the work also falls into this category.

- **Sound source content**
  The user interface described in [P6] includes a simple method of classifying transient sounds into finger snaps and other sounds. The audio surveillance system in [P7] analyzes all sounds that deviate enough from the background sound and classifies them into different categories in an unsupervised manner. Therefore, the content of the sound, that is, its cause, is the main area of interest. Examples of these causes are doors, cars, and keyboards.

1.2 Organization of the thesis

This doctoral thesis consists of a compendium and seven publications. The next chapter describes the ARA system, which is the background from which much of the research has originated. Following that, a separate chapter is devoted to each of the three main themes, which are distance estimation, impulse response analysis, and applications. In these chapters, previous research on the area is first briefly reviewed, followed by a summary of the new contributions to the field by the author. The compendium ends with a description of the contributions of the author.

2 BACKGROUND

The research leading to this thesis originates from groundwork on augmented reality audio carried out at the Helsinki University of Technology (TKK). Three of the seven publications ([P5], [P6], and [P7]) of the thesis present applications of sound signal analysis in the context of ARA technology. The remaining publications also present methods that could be utilized in an ARA context. Therefore, a brief introduction to ARA technology is presented in this chapter.

2.1 Augmented reality audio

The main idea in augmented reality audio is to add virtual and synthetic sound sources to the natural sound environment around a person [64, 63]. The goal is to be able to preserve the experience of the surrounding sound environment, keeping it as natural as possible. In many application scenarios, the added sounds should also fit the real environment as transparently as possible. This can be accomplished by a combination of signal processing techniques and a special binaural headset, which consists of possibly insulating ear plugs that have actuators (earphones) facing the ear canal and small microphones on the other side. The terms wearable augmented reality audio (WARA) [64] and mobile augmented reality audio (MARA) [63] have been used to describe an ARA system where the user wears the headset and the system is mobile in the usage situation.

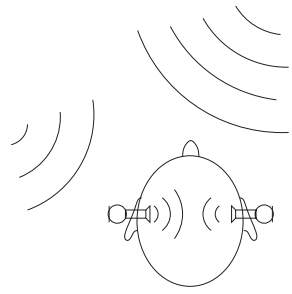

Figure 2.1: A listener in a pseudo-acoustic environment (after [64]).

Figure 2.1 illustrates a user of an ARA system in a situation where the sound of the outside world is picked up by the microphones and passed through to the ear canal side by routing it through the earphones. The user hears a representation of the surrounding environment. Because the experience may not exactly match that of the situation without the headset, the representation is called the pseudo-acoustic environment [64, 63].

In pseudo-acoustic reproduction, several factors in the headset contribute to colorations as the sound signal of the outside world is passed through the earphone to the ear canal. Depending on the type of the headset [153], there can be leakage between the headset and the skin, leakage

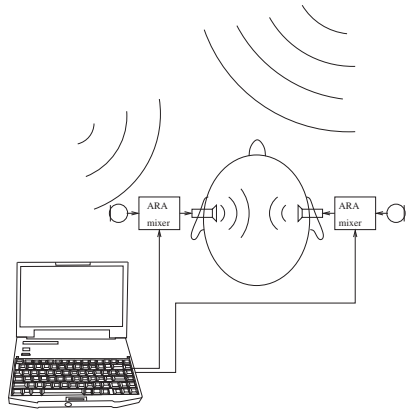

Figure 2.2: A listener in an augmented environment (after [64]).

through the headset, as well as leakage caused by tissue and bone conduction [63, 152, 153, 133, 157]. The transfer functions of the microphone and the earphone also have an effect on the signal [152]. Furthermore, when the ear canal is closed, the canal changes from an open tube to a closed one, affecting the resonances of the tube [133, 157].

In order to make the perceived pseudo-acoustic environment sound as natural as possible, special filters need to be introduced to equalize the coloration [63, 152, 153, 133, 157]. When equalizing the contribution of different types of leakage and tissue conduction, it is necessary to use equalization filters that have as little latency as possible [152, 133]. This necessitates the use of analog components to get the latency low enough. The equalization can take place in an ARA mixer [133], which also transmits the pseudo-acoustic environment from the microphones to the earphones and adds virtual sound sources to the mix (Fig. 2.2). These virtual sources can consist of, e.g., location-based advertisements, an auditory calendar, or auditory "post-it-notes" (messages played back when the user arrives at a certain location) [98, 142]. A usability study of the ARA headset combined with an ARA mixer when listening to natural sounds (the pseudo-acoustic environment) is presented in [154].

Figure 2.3 shows one of the simplest applications of ARA; binaural telepresence where the user at a remote location hears the sound environment at the location of another user. Two-way binaural communication can be similarly enabled. However, the voice of the remote end speaker will be heard at the center of the head by the local user. To mitigate this phenomenon, the voice of the far-end user can be moved away from the center while keeping the other parts of the remote binaural signal intact [97].

The entire ARA system is illustrated in Fig. 2.4. The bottom part of the diagram shows how the virtual sounds are auralized. The auralization part of the system includes a module for room simulation and binaural reproduction by using head-related transfer functions (HRTFs). In headphone listening the sound sources are often perceived to be located inside the head. This phenomenon is termed *inside-the-head locatedness* [14]. In order to make the virtual sound source sound convincing, the listener

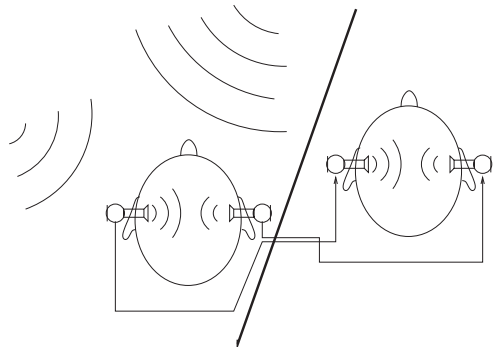

Figure 2.3: One user experiences the sound environment heard by another user (after [64]).

has to perceive the sound source being located outside the head (externalized) [63]. This can be achieved by using personalized HRTFs measured from the listener, which, unfortunately is not practical enough. Moreover, HRTFs are typically measured in free field, which makes them applicable only in anechoic situations. One possible solution for achieving externalization is to add early reflections and reverberation to the virtual sources. In a listening test using an ARA headset, a generic HRTF response measured from just one subject combined with early reflections and statistical late reverberation was found to result in adequate externalization [63].

Virtual sound sources (events) can be categorized as freely-floating or localized sources (events) [98]. The former refers to positioning the sources so that the only reference point is the head of the user, while the latter refers to localizing the events to objects in the real world. For freely-floating virtual sources to be perceived as natural, the location and orientation of the head of the user needs to be known at each moment in time. Head tracking is therefore necessary. Information from the head tracking system can be utilized to keep a sound source at a certain location (localized source) even if the user moves or turns his/her head. Moreover, many of the applications

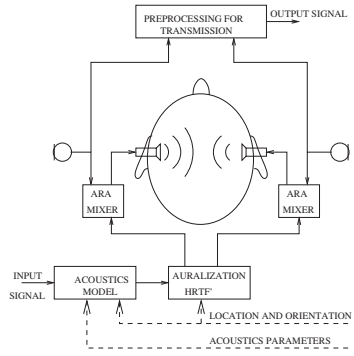

Figure 2.4: A generic diagram of an augmented reality audio system (after [64]).

of ARA technology, e.g., location-based advertisements, need the knowledge of the location of the user. Sometimes it is appropriate to know the location on a local scale, e.g., where the user is located inside a building. A global scale is appropriate when the user is in outdoor environments [63]. In outdoor use, it is possible to use the Global Positioning System (GPS) for tracking the location of the user [63, 123]. The location of the user can be also utilized in gaming applications [123].

For inside environments, acoustic head tracking [155, 156, 80] can be utilized to avoid using much extra equipment, such as magnetic trackers (see [107] for a review of tracking technologies). In acoustic head tracking as proposed in [155, 156, 80], the common sound source localization paradigm of a moving sound source and a static microphone array is reversed. The array consists of the two microphones worn by the user of the ARA headset, and the anchor sounds that are played back from one or more static loudspeakers. While the anchor sound could also be a sound source already present in the environment, e.g., a computer or an air shaft, using a known reference signal as the anchor will reduce the effect of interfering sounds [155]. The location and orientation of the head of the user can be obtained based on the cross-correlation between the received signals and the known anchor source [155, 156, 80]. By using high frequencies in the anchor sound it is possible to reduce the annoyance and sensitivity to environmental noise [80]. When using multiple anchor sounds, it is necessary to generate the anchor signals so that they do not overlap in frequency [80].

2.2 Relationship between ARA and this thesis

The research on ARA conducted at TKK has concentrated on three areas:

1. Hardware
2. Algorithms
3. Applications

Table 2.1 categorizes all publications (excluding master's theses) related to ARA research at TKK into these three classes. The publications of this thesis are also shown in order to make clear which parts of ARA research they relate to. An overview of the ARA system is presented in [64] and [63]. The research related to hardware has concentrated on the ARA headset [152, 153, 154], and the ARA mixer [133, 157]. An ARA hardware and software platform for mobile outdoor use has been described in [123]. Apart from publications included in this thesis, research on algorithms has mostly concentrated on acoustic head tracking [155, 156, 80]. One publication [49] describes a method for acquiring BRIRs for auralization purposes on-the-fly by using finger snaps made by the user as excitation. Ideas and implementations of various applications of ARA have been presented in [64, 98]. These applications include binaural telephony, speech communications with head-tracked auralization, an auditory sticker application, a calendar application, and augmented sound events such as advertisements and notifications. The problem of rendering the voice of the far-end user in binaural telephony has been specifically addressed in [97].

All of the publications of this thesis relate to the algorithms and applications of ARA. Since all of the publications deal with some kinds of analysis algorithms, they are all listed under "Algorithms" in Table 2.1. Three of the publications ([P5], [P6], [P7]) are also listed under "Applications", as they all are also concerned with a specific application, i.e. the incoming sound signal is analyzed, and the information extracted is applied in the ARA context. There are also connections to other fields such as hearing aids research, where signal processing techniques for analyzing binaural signals are also applied [61]. The following two question have been the starting points for the research described in this thesis:

1. "What information on properties of the surrounding environment, and the sound sources it contains, could be obtained by analyzing the binaural signals picked up by the microphones of an ARA system?"

2. "How could this information be applied in the ARA context?"

Publications [P1]–[P4] are related to the first question. Two of them address the problem of computational sound source distance estimation as a pattern recognition (or learning) problem ([P1], [P2]). Although information on the location of sound sources present in the environment could definitely be useful in ARA, there are a few reasons why the method presented can not directly be applied in ARA, though. The method requires a stationary listener and the system also has to be trained for each configuration, i.e. positioning of microphones and sources in each room, separately. The other two of the publications concern binaural room impulse response analysis ([P3], [P4]). This topic does not necessarily directly apply to ARA, but the inspiration for the research was the same question above.

The rest of the publications in this thesis are also concerned with the second question. Some of the information obtained from the analysis of the binaural signals can be utilized in making the virtual sources fit the surrounding real sound environment better, e.g. by blindly estimating the reverberation time from the binaural input and adjusting the RT of the auralization to match that of the space, as is described in [P5]. The binaural input can also be used as part of a user interface to the system. Publication [P6] describes one possible ARA user-interface, based on detecting and localizing finger snaps made by the user. The audio surveillance system described in publication [P7] could also be used in an ARA context.

Table 2.1: Classification of the publications related to the ARA research at Helsinki University of Technology (TKK).

Hardware	Algorithms	Applications
[64]	[64]	[64]
[63]	[155]	[97]
[152]	[156]	[98]
[153]	[80]	[**P5**]
[133]	[63]	[**P6**]
[157]	[**P5**]	[**P7**]
[154]	[**P6**]	[123]
[123]	[**P3**]	
	[**P7**]	
	[**P1**]	
	[**P2**]	
	[**P4**]	
	[49]	

3 COMPUTATIONAL SOUND SOURCE DISTANCE ESTIMATION

3.1 The problem of sound source distance estimation

Sound source distance estimation refers to estimating the line-of-sight distance, or range, between a sound source and a receiver (listener). In this thesis, only passive distance estimation based solely on the sound signals that are received, is considered. This approach can be contrasted with active approaches such as echolocation used by bats [145], which involves sending sound pulses to the environment and determining the distance based on the received signals. In the work described in this thesis, sound source distance estimation is only considered inside rooms, where there are reflections and reverberation present.

Figure 3.1 depicts the problem of sound source distance estimation in rooms. In this example there is a listener and two sound sources present. Discrete echoes from the walls of the room and statistical diffuse late reverberation are also present in this scenario. The goal is to estimate the distance to a sound source based on the received binaural or monaural signal only. Having two ears, humans naturally can utilize two sound signals to estimate sound source distance. However, distance information can also be present in monaurally recorded sounds as well.

When solving the problem of distance estimation using computers, it would be easy to add a third microphone. Since with three microphones it is possible to estimate two angles of arrival (or two time delays), the problem of distance estimation is solved as the position of the sound source can be easily revealed at the intersection defined by the two angles of arrival. Adding even more microphones makes the localization more accurate. Multi-microphone approaches utilize signal processing techniques for microphone arrays can be used for sound source localization [16]. However, in this thesis only monaural and binaural signals are considered.

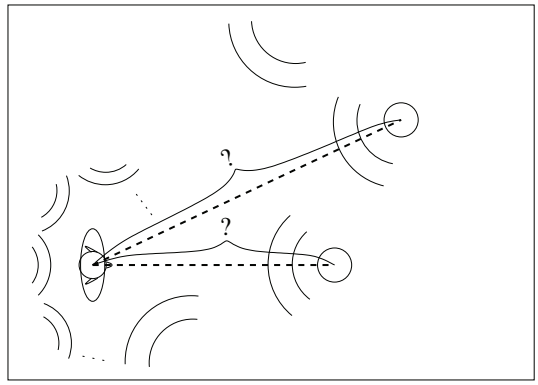

Figure 3.1: A depiction of the problem of sound source distance estimation in rooms.

3.2 Sound source distance perception cues

Human perception of sound source distance has been researched since the beginning of the last century [48]. However, systematic research only started during the 1960s (e.g. [33, 51]). A good summary of early research into distance perception can be found in the Ph.D. thesis of J. Chomyszyn [28]. Another Ph.D. thesis dealing with distance perception is the thesis of S. Nielsen [114]. A review of later research in distance perception is presented in [174], and a review that also concerns distance perception in animals can be found in [113]. Distance rendering in virtual audio and auditory displays is a topic that has gained some attention as well [143, 144, 119, 172]. There are several distance cues that are utilized by the human auditory system. These cues are combined in a flexible and adaptive manner in order to obtain the perception of auditory distance [171].

Sound pressure level

Perhaps the intuitively clearest distance cue is the sound pressure level (SPL), which for a point source obeys the so called $1/r$ law, i.e. the sound pressure drops inversely proportional to distance r. This means that for a point source in free field, the sound pressure level drops 6 dB for a doubling of the distance. Obviously, this is an idealization, as natural sound sources are rarely point sources and rarely in free field. Also, in human perception this law only holds for distant sources as interaction with the human head breaks this law at distances less than one meter [143]. Sound pressure level correlates with perceived loudness of the sound. However, the loudness of a familiar sound, such as speech, is often perceived as constant even when the distance changes [175]. This is similar to the visual size constancy effect, where the size of an object is perceived as constant while the size of the retinal image changes with distance [70]. The type of speech (whisper, conversation, shout), the perceived production level of speech (measured at one meter from the speaker), and the level of presentation (the sound pressure level at the listener) have been found to influence distance judgments as well [24].

Spectral cues

Spectral cues also affect distance perception. At distances larger than 15 m, air absorption decreases the levels of high frequencies more strongly compared to the lower frequencies [174]. In rooms the frequency-dependent absorption of the reflected sound also contributes to these spectral changes [115]. Therefore, the relative levels between high and low frequency ranges can act as a distance cue. However, it has been speculated that since low frequencies are emphasized for near-field sources, the effect of spectral cues may be the opposite at short distances [33]. This may not be the case though, because this theoretical boost of low frequencies is due to the distance dependent changes to the particle velocity at different frequencies, which the human ear is not sensitive to [174].

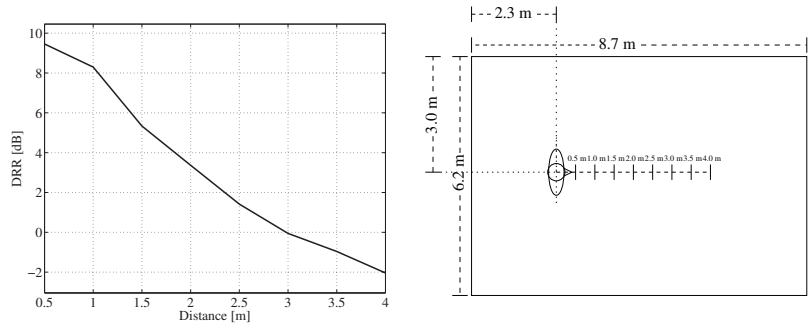

Figure 3.2: Left: Direct-to-reverberant ratio as a function of sound source distance. Right: The corresponding source and receiver configurations.

Direct-to-reverberant ratio

One of the most important distance cues is the direct-to-reverberant ratio (DRR). The DRR is defined as the ratio of the direct and reverberant sound energies of a room impulse response $h(t)$ [85, 86]

$$\mathrm{DRR} = 10\log_{10}\frac{\int_0^T h^2(\tau)\mathrm{d}\tau}{\int_T^\infty h^2(\tau)\mathrm{d}\tau} \tag{3.1}$$

where T is the duration of the direct sound, which is usually around 2–3 ms. The left panel of Fig. 3.2 shows an example of the DRR as a function of distance. The DRR is computed from the left channel of the measured real-world binaural room impulse responses corresponding to the source and receiver configurations depicted in the right panel of Fig. 3.2. It is seen that the DRR decreases monotonously as the distance to the sound source increases. Figure 3.3 shows the spectrograms of a short speech sample played back at the same source locations. The effect of reverberation can be observed in both time and frequency, as details of the spectrogram are blurred when the distance increases.

The DRR alone is not enough for estimating the distance, and other information such as the reverberation time, room volume, and source directivity are also needed. It is speculated that humans implicitly learn these additional cues based on experience [21]. It is also likely that the actual processing that computes the DRR inside the human auditory system is not necessarily based on temporal processing. Zahorik [173] has investigated the discrimination thresholds of DRR in human listeners and found evidence against temporal processing used for computing the DRR. A study by Bronkhorst [20] suggests that binaural spatial separation of the direct and reverberant sound energies may be involved in computing the DRR. A more recent study by Larsen *et al.* [85], however, suggests that spectral cues are the most discriminative cues for DRR perception. Based on the current state of knowledge, it is, therefore, not clear what kind of auditory processing is taking place to compute the DRR.

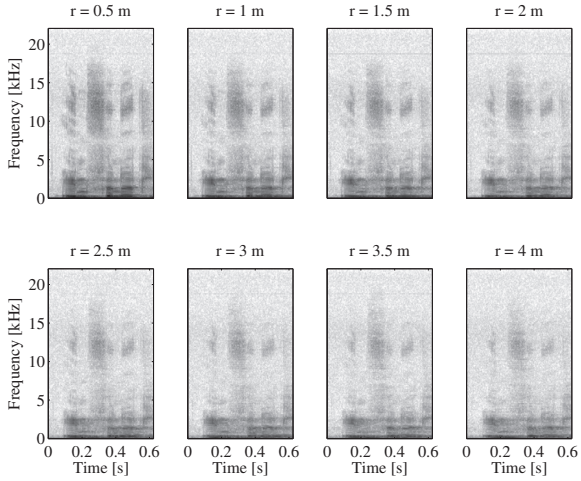

Figure 3.3: Spectrograms of a male speech sample recorded at various distances in a room with $RT = 0.6$ s.

Dynamic cues

In dynamic situations, where the source and/or the listener are moving, some additional distance cues become available. One of these cues is the acoustic tau, which specifies the time to collision with the sound source [60, 5]. The idea is that rate of change of the intensity of the sound can be used to make judgments of how far away the sound is, if the velocity of the sound source (or the listener) is constant. Another dynamic distance cue is the acoustic motion parallax, which is related to the rate of change of the azimuth angle when the listener is moving and the source is stationary [147]. The further away the source is, the slower the rate of change of the azimuth angle is.

The effect of the environment

The usefulness of distance cues depends heavily on the type of environment (free-field, room) and the range of distance considered. With human distance perception, the head has a special effect on the interaural cues at distances less than one meter. In a nutshell, the interaural time difference (ITD) stays constant while the interaural level difference (ILD) changes with distance in this range [23, 22]. The latter is thus a viable distance cue for nearby sources.

Absolute and relative cues

It is often useful to differentiate between absolute and relative auditory distance cues. The former provides information on the absolute distance to the sound source (for example, "The source is at a distance of two meters".), while the latter cues allow only for relative judgments (for example, "The source has moved to a distance twice as far away from where it was previously."). Sound pressure level (intensity) can be regarded as a relative distance cue [143], while the direct-to-reverberant ratio is considered to

provide information on absolute distance [106, 115, 21].

3.3 Human performance in distance estimation

In order to gain a perspective on the accuracy of computational sound source distance estimation, it is useful to have a look at how well humans fare in the distance estimation task. There is a tendency in human listeners to underestimate small distances and overestimate large distances [174]. A psychophysical function in the form of a compressive power function can be used to model the relationship between the true sound source distance r and the perceived distance r' as [174]

$$r' = kr^a \tag{3.2}$$

where the constant k and exponent a are parameters of the power function. Zahorik *et al.* [174] present statistics of the parameters of the power function for listening test results compiled from 21 articles related to sound source distance perception research, comprising of a total of 84 data sets obtained in various test setups and environments. Figure 3.4 shows the perceived distance modeled by Eq. (3.2) as a function of the true distance, with parameter values of k and a chosen as the mean of the parameters estimated from the 84 data sets presented in [174]. It is seen that distances below approximately 1.9 m are overestimated, while distances above that are progressively more and more underestimated. This underestimation may correspond to a phenomenon called auditory horizon, which refers to the existence of a maximum perceived distance. However, there is no direct evidence that such a saturation phenomenon exists [174].

In conclusion, human accuracy in distance estimation is clearly lower compared to estimating the angle of arrival, where the accuracy can be as low as 1° for sound sources in the front [174].

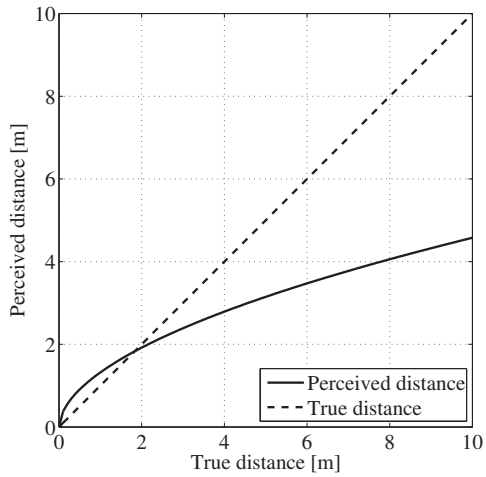

Figure 3.4: An example of modeling distance perception experiment results using a compressive power law function ($k = 1.32$, $a = 0.54$).

3.4 Related research

As a topic of research, computational distance estimation differs in an important way from estimation of the direction of arrival, i.e. azimuth and elevation angles. The difference is that actual models for computational distance estimation have not been proposed until recently (as discussed later), while models for direction localization have existed for several decades. Already in 1907, Lord Rayleigh proposed the so called duplex theory of sound localization [118]. The theory basically states that for sounds arriving from directions other than the medial plane there are differences in level (ILD) and time (ITD) between the two ears, and these differences are used by the auditory system to resolve the direction of incoming sound. A popular auditory model for calculating the ITD is the cross correlation (CC) model of Jeffress [74]. Another prominent model is the equalization cancellation (EC) model, first proposed by Durlach [41]. In addition to these pioneering models, lots of new ideas for directional localization have been proposed over the years — extensions of the CC and EC models [93, 94, 31, 17, 45, 121] and new architectures, such as [130] and [125]. Binaural localization and tracking algorithms have also been studied in, for example, [7, 27, 120, 95, 134, 136, 165, 135]. Reviews of computational binaural sound localization can be found in [149, 150]. The basics of spatial hearing are presented in [14].

In an article published in 1968, Hirsch [69] proposed that the sound source distance r could be estimated by the ratio of the time and intensity differences Δt and ΔI

$$r = \frac{2c\Delta t}{\Delta I / I_{\text{av}}} \tag{3.3}$$

where c is the speed of sound and I_{av} is the average intensity between the ears. It was later pointed out [56] that with the resolution of the human auditory system, the errors of distance estimated using the above formula would be large. Also, the method would also be very imprecise when the source is directly in front of the listener, as Δt and ΔI would be close to zero. Molino [110] presents an extension of the model where the head is modeled as a rigid sphere. However, the azimuth angle of the sound source has to be known in this extended model. Both of these models have difficulties with distant sources, as the interaural cues are virtually independent on distance for distances greater than one meter [23, 22].

Calamia [25] proposed a method for 3D localization in the nearfield ($r < 1\,\text{m}$). The azimuth, elevation, and distance (range) are estimated by comparing the interaural cues of the measured signal with those precomputed from HRTFs at different positions. The best match reveals the most likely location. With regards to the distance, the method only works at distances less than one meter, because the ILD is distance-dependent only up to a distance of approximately one meter [23, 22].

If the relative timings and amplitudes of a number of early reflections are known, it is possible to estimate the sound source distance — along with other quantities such as the DRR and the reverberation time — based solely on that information [86]. There has to be some method to accurately estimate the reflections from the signal, though. In [86] it is reported that

the distance estimates of this method suffer from underestimation and large variability when the first nine early reflections are used for the estimation.

Griesinger investigated binaural [57] and monaural [58] cues that characterize the apparent distance of sound sources. The binaural cue is based on the interaural cross-correlation coefficient (IACC), and the monaural cue is pitch coherence. The actual application of these cues to computational distance estimation is not tested, though.

A method for automatic classification of hand claps into far-field and near-field was presented by Lesser and Ellis [90]. The energy-based monaural distance cues utilized were the center of mass, slope, and energy difference of the transient and the background sound. The generalization performance between two rooms was found to be good. The presented method is applicable for transients only.

The first actual model for computational binaural sound source distance estimation was proposed in 2007 by Lu *et al.* [102, 34]. Their model is based on the acoustic tau and the motion parallax, which are estimated sequentially in a particle filtering framework. However, this model is suitable for dynamic situations only due to its sequential nature. An extended version of the model has been presented, where the DRR is added as an extra distance cue [101]. The relationship between the DRR and the distance is learned from training data using Gaussian mixture models (GMMs). Adding the DRR was reported to improve the results for simulated speech sources. Performance of the models presented has not been evaluated in real rooms yet.

Georganti *et al.* [52] have presented a method for monaural learning of the distance of a speech sound source. The proposed features measure the statistics of the LP residual and the spectrum of the signal. The learning is based on GMMs. It is reported that the method is accurate for distances up to 1.5 meters.

To conclude, the state of the art in computational distance estimation can be summarized as follows. There exists one actual computational model for sound source distance estimation [102, 101]. However, evaluation of the model with real-world data has not been presented yet. Other methods have restrictions on the ranges considered (e.g. near field only [25]) or the signal types (e.g. transients only [90]), and the accuracy (e.g. only accurate up to 1.5 m [52] or just crude classification to near/far field [90]). Further contribution is needed in finding features that depend on distance only. For example, in [101] it is reported that the way of computing the DRR used resulted in some dependence on sound source power for speech sources. Better accuracy would also be desirable. It is known that humans cannot estimate sound source distance very accurately (see Sec. 3.3), but there is no reason to settle for the accuracy of the human auditory system. The work presented in this thesis addresses some of the aforementioned problems by presenting a method that can correctly recognize the distance of a speech sound source — the most important signal class for practical applications — at distances varying from 0.5–5.0 m (possibly larger distances can also be recognized, but this has not been investigated yet).

3.5 Binaural coherence

The methods for sound source distance learning presented in publications [P1] and [P2] are based on the binaural coherence — sometimes termed the interaural coherence (IC) [45] — as the main localization cue. The binaural coherence is also utilized for different purposes in publications [P5] and [P6]. Binaural coherence is basically the strength of correlation between the left and right ear signals as a function of frequency. When there is correlation between the signals, the coherence is high. In terms of the sound field that the binaural signal corresponds to, the coherence is low when the sound field is diffuse, which usually means that there is reverberation. On the contrary, if the direct sound is strong at the listening position, the binaural coherence will be high. Therefore, the interaural coherence can be used to indicate which parts of the signal — in time and in frequency — are occupied by the direct sound, which is usually the signal of interest. At the same time, low coherence parts indicate parts of the signal occupied by reverberation, which is usually considered a nuisance to be removed. Therefore, the coherence can be useful in applications such as dereverberation [3], multi-source localization [45], and selecting signal processing strategies in hearing aids [167]. It has also been shown that the human auditory system responds to high binaural coherence with increased activity, indicating that coherence is utilized in the localization process [182].

In the work described in this thesis, the binaural coherence is estimated as the magnitude-squared coherence (MSC) between the left and right ear signals [3, 167]

$$\hat{\gamma}^2_{\mathrm{lr}}(f,t) = \frac{|\hat{G}_{\mathrm{lr}}(f,t)|^2}{\hat{G}_{\mathrm{ll}}(f,t)\hat{G}_{\mathrm{rr}}(f,t)} \tag{3.4}$$

$$\hat{G}_{\mathrm{ll}}(f,t) = \langle |X_{\mathrm{l}}(f,t)|^2 \rangle \tag{3.5}$$

$$\hat{G}_{\mathrm{rr}}(f,t) = \langle |X_{\mathrm{r}}(f,t)|^2 \rangle \tag{3.6}$$

$$\hat{G}_{\mathrm{lr}}(f,t) = \langle X^*_{\mathrm{l}}(f,t) X_{\mathrm{r}}(f,t) \rangle \tag{3.7}$$

where $\hat{G}_{\mathrm{ll}}(f,t)$ and $\hat{G}_{\mathrm{rr}}(f,t)$ are the power spectrum estimates at frequency f and time t of the left and right ear signals, respectively, $\hat{G}_{\mathrm{lr}}(f,t)$ is the cross spectrum estimate, and $X_{\mathrm{l}}(f,t)$ and $X_{\mathrm{r}}(f,t)$ are the left and right channel signal short-time Fourier transforms (STFTs), respectively. Complex conjugation is denoted by an asterisk $(*)$. While it is possible to estimate the spectra in different ways [26], in this thesis the estimation is done by time-averaging (denoted by $\langle \cdot \rangle$) using a first-order IIR

$$\langle Q(n) \rangle = \beta \cdot \langle Q(n-1) \rangle + (1-\beta) \cdot Q(n) \tag{3.8}$$

where $Q(n)$ is an arbitrary time series and β is the forgetting factor, which is more conveniently defined using the time constant

$$\beta = \exp\left(-\frac{N_{\mathrm{h}}}{\tau \cdot f_{\mathrm{s}}}\right) \tag{3.9}$$

where N_{h} is the hop size of the STFTs, τ is the desired time constant in seconds, and f_{s} is the sampling rate. By increasing the time constant, information from a longer time period is utilized in computing the coherence.

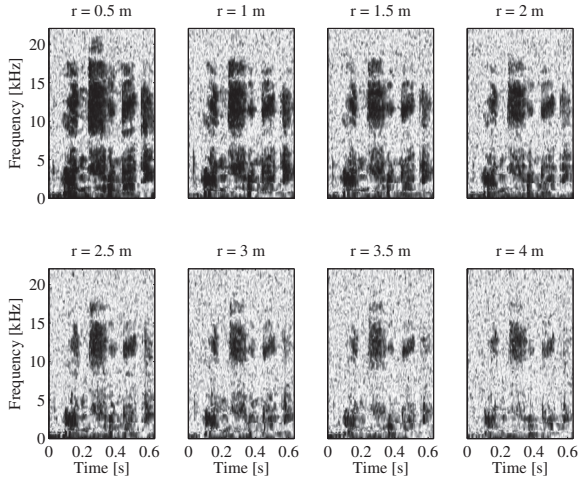

Figure 3.5: Coherence spectrograms of a male speech sample recorded at various distances in a room with $RT = 0.6$ s.

Figure 3.5 presents coherence spectrograms of the same short speech sample as in Fig. 3.3 played back at the same locations. The figures are plotted so that white corresponds to a coherence of zero and black to a coherence of one. The time constant for coherence computation was set to $\tau = 5$ ms. Compared to the corresponding spectrograms (Fig. 3.3), the effect of increasing distance is much more evident. This is seen as the decrease of the high coherence areas in the plots. In parts of the signal where noise dominates, the coherence gets random values.

3.6 Novel contributions

The approach for sound source distance estimation taken in this thesis is inspired by machine learning approaches to sound localization (e.g. [7, 120, 146]). Particularly, the approach of Smaragdis and Boufounos [146] has been the main inspiration. They presented a microphone array method for learning the position of a stationary sound source. The microphone array can have as few as two sensors. Because in [146] it is reported that when using two sensors (binaural dummy head) the method has difficulties with sound source positions that share the same azimuth angle but have different distances, it was decided to investigate how distance could be learned with similar methods when only two microphones (binaural input signal) are available.

Two articles in this thesis deal with binaural sound source distance learning. Publication [P1] is the first publication on the topic. It proposes a novel method for learning the distance of a sound source, based on the binaural coherence (see Sec. 3.5) as a distance cue. White noise is used as the training signal, while the distances of speech sounds are recognized. A stationary listener (a dummy head) is situated inside a room, and a stationary sound source (a loudspeaker) is also present. In the learning phase,

white noise is played back from the loudspeaker and the time average of the coherence of the binaural signal recorded using the dummy head is taken. This mean coherence is termed coherence profile, as it captures the mean shape that the coherence takes at a certain position in the room. Interestingly, it was found that the variance of the coherence is a distracting cue and could not be used in the classification. In the testing phase, the coherence of the speech signal is computed. The classification system is based on energy-weighted maximum likelihood, which basically compares the coherence of the received binaural signal with the coherence profiles obtained in the training phase, giving more weight to time-frequency elements that have high energy. The method is evaluated using a dummy head and a loudspeaker as the listener and the sound sources, respectively. It is shown that the method can correctly recognize the distance when the sound source is in the front of the listener.

Publication [P2] delves deeper into sound source distance learning, expanding the research of [P1] with some modifications to the method and by having a larger data set for evaluating the method. By increasing the time constant of the coherence calculation (see Sec. 3.5), good performance is obtained also when the sound source is at a side direction from the listener. The proposed method is compared with a slightly modified version of the method presented in [146]. It is shown that both methods can recognize the distance of a speech sound source correctly in most cases. The investigations also reveal that the coherence exhibits clear changes as a function of sound source distance (see Fig. 3 in [P2]). Compared to the interaural time and level difference features employed of a previous approach [146], the coherence also behaves more smoothly along the frequency axis. Possibly due to these reasons, the proposed method is able to generalize better compared to the baseline method [146], when there is a slight mismatch between the sound source locations in training and testing (see Sec. V-D in [P2]). It is revealed that the main drawback of the methods is the lack of generalization when there is a great mismatch between the spatial configuration of the source and the microphones in the training and testing conditions.

The work described in publications [P1] and [P2] presents important novel contributions to the field of computational sound source distance estimation and also to acoustic source localization in general. It is shown that the distance of a speech sound source can be recognized from binaural signals using a simple classification approach. This enables using only two microphones for localizing a speech source, while in traditional microphone array techniques [16] there are typically more than only two microphones. Also, in these kinds of learning approaches there is no need for microphone calibration and the locations of the microphones do not have to be known a priori [146]. The publication [P2] also sheds light on the behavior of the interaural coherence and other interaural parameters when the source and receiver positions change inside a room. The limits of machine learning approaches for sound source localization are investigated, gaining knowledge applicable in further research on the subject. The investigation of the behavior of coherence as a function of distance also contributes to basic research in room acoustics.

4 ANALYSIS OF BINAURAL ROOM IMPULSE RESPONSES

4.1 The problem of detection and localization of early reflections

A binaural room impulse response characterizes the acoustic behavior of a room from the source to a listener. A BRIR is the signal that is received at the ears of a listener after sending a perfect impulse into the room. There are several factors that shape the sound that a sound source emits into a room. The sound is spread into the space according to a frequency-dependent directivity pattern, which is a property of the sound source. Part of the sound travels directly to the listener if there are no obstacles for line-of-sight propagation. Some of the sound energy gets reflected from the surfaces present in the room. These surfaces include walls, floor, ceiling, and other objects such as furniture, that may be present in the room. Some of this reflected sound then reaches the listener, and some of it continues reflecting from the surfaces. These discrete reflections that occur for some time after sending sound into the room are called early reflections. In addition to specular reflections where the reflection angle is the same as the angle of the incoming sound, diffraction also happens at the surfaces, causing the sound to scatter in other directions as well. At some point in time there will be so many overlapping reflections that one cannot speak of individual reflections any more, and it is best to treat the received sound as statistical late reverberation. The effect of the listener is also included in a binaural room impulse response, as the head and torso of the listener also shape the signal entering the ears.

Figure 4.1 shows a simplified version of a monaural (single-channel) room impulse response. The direct sound arrives first. Then, early reflections arrive, followed by late reverberation. In this thesis, two different

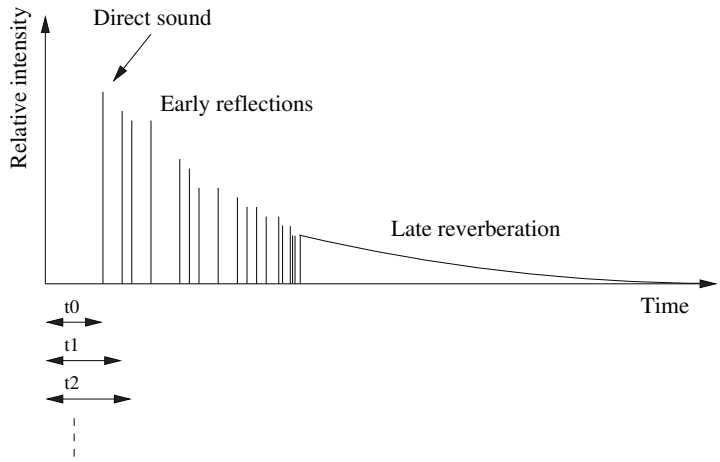

Figure 4.1: A simplified room impulse response composed of direct sound, early reflections, and late reverberation.

sub-problems related to analysis of binaural room impulse responses are considered:

1. Estimating of the arrival times (detection) of early reflections ([P3], [P4]).
2. Estimating the angles of arrival (localization) of early reflections ([P3]).

The first problem refers to finding the arrival times of the early reflections. The arrival time corresponds to the total time that the sound has traveled in the air, since sending the sound into the room. Fig. 4.1 depicts the times of arrival of the early direct sound (t_0) and the two first early reflections (t_1,t_2). The second problem is about finding the direction of arrival of each of the early reflections that were detected as part of the solution for the first problem. In this thesis, only the estimation of the azimuth angle, i.e. the angle of arrival in the lateral plane, is considered. Estimation of the angle of arrival is made possible by having two channels, which provides interaural localization cues for determining the direction.

4.2 Related research on analysis of early reflections

Analysis of room impulse responses gives insights into room acoustics [84]. It has been widely regarded that the lateral reflections — reflections that come from the sides relative to the listener — contribute to the quality of the acoustics of a hall [8]. By using a microphone with both omnidirectional and figure-of-eight directivity characteristics, a quantity called lateral energy fraction (LEF) can be computed [9]. The LEF characterizes the amount of acoustic energy coming from the sides relative to the total incoming energy. A related quantity, the interaural cross-correlation coefficient (IACC) is also related to spatial impression [9, 67]. It can be calculated from binaural room impulse responses measured using small microphones situated at the entrances of the ear canals of a real head or a dummy head.

However, both LEF and IACC do not give detailed information about the individual early reflections in the BRIR. The arrival times and directions of arrival of the reflections could be useful for constructing new measures of the quality of acoustics and also for gaining insight into the details of the acoustic behavior of rooms. One example of the latter is listening to room impulse responses in slow motion so that the relative arrival times and directions of early reflections can be heard [96]. According to the knowledge of the author, not much research exists on the topic of estimating the arrival times and directions of individual reflections, based on any kind of room impulse response.

Gover *et al.* [55] used a 32-microphone spherical array to capture directional room impulse responses. From these impulse responses, the reflections could be localized in both time and frequency. The arrival times and directions were found to match with knowledge of the room geometry.

Park and Rafaely [122] used a 98-element spherical scanning microphone array to obtain directional impulse responses in concert halls. Their method was reported to correctly localize the direct sound and some initial

early reflections — the first order reflections and a few second order reflections. Later, Rafaely [127] proposed a method which employs a microphone on a rotating boom to sample the response on an imaginary sphere, based on which the directional response can be calculated. This method also identifies the direct sound and the first order early reflections correctly.

Rigelsford [132] used a volumetric microphone array for acoustic imaging. 64 microphones were placed in pseudo-random positions inside a spherical volume. The method was able to localize multiple sound sources and therefore might be suitable for analyzing early reflections, even though this application is not investigated in [132].

O'Donovan *et al.* [116] used a spherical microphone array for acoustic imaging so that the acoustic image could be overlaid on a video of the space. This techniques allows to see visually exactly where reflections originate from as a time progresses after the arrival of the direct sound [117].

Yet another method based on microphone arrays is the one presented by Roper and Collins [137, 138]. They used a combination of a circular array (24 microphones) and a line array (23 microphones) for finding the elevation and azimuth angles of early reflections, respectively. The detection of reflections was based on emitting a special chirp signal and then detecting the arrival times based on matched filtering, which is practically implemented using the matching pursuit algorithm. The direction of arrival is then estimated using beamforming. The algorithm was able to identify all first order reflections and many of the second order reflections.

Defrance [39] applied matching pursuit to identify the arrival times of reflections in a monaural room impulse response. The source signal was a pistol shot, which was used as the atom in the matching pursuit. Therefore, the method cannot be used with impulse responses measured in the standard way, for example, with sweep [112] or the maximum length sequence [131]. The arrival time distributions were investigated and a way of measuring the mixing time (see [68]) was presented.

Kuster [83] investigated the estimation of room volume from a room impulse response. In the article, a method for detecting the arrival times of reflections from a monaural room impulse response is presented. This method is used in publication [P4] as a baseline method for detecting early reflections in time.

In conclusion, the state of the art in analysis of early reflections can be summarized as follows. There are various methods which require an array consisting of several microphones [55, 122, 132, 137, 138]. Some methods are based on a special recording system such as a rotating boom [127] or a spherical microphone array [116, 117], or they require the impulse response to be measured using a special signal such as a pistol shot [39]. A monaural method for identifying the arrival times of early reflections has been presented [83]. According to the knowledge of the author, methods that can estimate the arrival times and directions of early reflections based on binaural room impulse responses have not been presented earlier. The work presented in this thesis addresses this gap in knowledge.

4.3 Continuous wavelet transform

Publications [P3] and [P4] present novel wavelet-based methods for analysis of binaural room impulse responses. The methods are largely based on the continuous wavelet transform (CWT), which is a time-frequency transform that allows a finer time resolution at higher frequencies compared to the lower ones [32]. Also, the frequency resolution is logarithmic, which corresponds to human hearing that also has a logarithmic frequency resolution [111].

The CWT is defined for a signal $x(t)$ as [99, 32]

$$W_{\mathrm{x}}(u,s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-u}{s} \right) \mathrm{d}t \tag{4.1}$$

where u is translation, s is scale (sometimes called dilation), t is time, ψ is the mother wavelet, and $*$ denotes the complex conjugation. Two different mother wavelets are used in publication [P3]: the Morlet and Paul wavelets. In publication [P4], only the former is used. The Morlet wavelet is defined as [158]

$$\psi(t) = \pi^{-1/4} e^{j\omega_0 t} e^{-t^2/2} \tag{4.2}$$

where t is time and ω_0 is the oscillating frequency. Both are non-dimensional. The Paul wavelet is defined as [158]

$$\psi(t) = \frac{2^m j^m m!}{\sqrt{\pi(2m)!}} (1 - jt)^{-(m+1)} \tag{4.3}$$

where m is the order of the wavelet. The continuous cross-wavelet transform (XWT) is used for detecting ([P3]) and segmenting ([P4]) early reflections. The XWT is defined as [59]

$$|W_{\mathrm{LR}}(t,s)| = |W_{\mathrm{L}}(t,s) W_{\mathrm{R}}^*(t,s)| \tag{4.4}$$

Figure 4.2 plots the two wavelet functions used in publications [P3] and [P4]. The Paul wavelet is more localized in time compared to the Morlet wavelet. Therefore the time resolution of the Paul wavelet is better than that of the Morlet wavelet, while with the frequency resolution the situation is the other way around. Figure 4.3 shows a measured binaural room impulse response and the XWT computed from it. The strongest early reflections can easily be seen as dark areas in the XWT (top panel), as well as in the time-domain signals (two bottom panels). Since the wavelet functions are well localized in time, they are suitable for the analysis of early reflections and other transient-like signals.

4.4 Related research on applications of wavelets to RIR analysis

Since the analysis method presented in publication [P4] is based on wavelets, previous applications of wavelets to audio and especially acoustic impulse response analysis are briefly summarized. The continuous wavelet transform has been used previously in applications such as audio restoration [168], additive synthesis [12], and analysis of intermodulation effects [13].

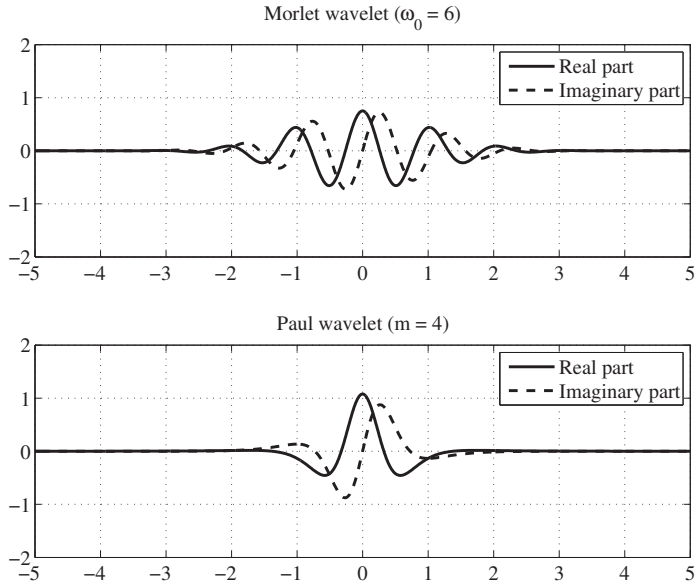

Figure 4.2: The complex Morlet and Paul wavelets.

Very few applications of wavelets to room impulse response analysis can be found from the literature related to room acoustics. Schönle *et al.* [139] present a wavelet-based method for sub-band decomposition and re-synthesis of room impulse responses. Their method is based on the discrete wavelet transform (DWT), though. Loutridis [99] used the CWT for analysis of room and loudspeaker impulse responses. The CWT is used for separating modal components and estimating decay and reverberation times. Lee [88, 89] applied the CWT for accurate determination of the reverberation time. It was demonstrated that the reverberation time can be estimated more accurately by using the CWT instead of a standard third-octave bandpass filterbank.

4.5 Novel contributions

Estimation of reflection arrival times from a BRIR [P3]

A new method for detecting the arrival times of early reflections from binaural room impulse responses is presented in publication [P3]. The method is based on the cross-wavelet transform (see Sec. 4.3 and [59]) between the left and right ear signals. The method is tested with both measured and modeled responses. The estimated arrival times are compared to the ground truth obtained from a shoebox room model, which is constructed for real rooms using the image source method [2, 15]. The average errors between the arrival times of the reflections in the model and the arrival times obtained using the proposed method were calculated. It is shown that the proposed method can locate the first order reflections with a mean error between 0.24–0.30 ms with the measured responses. Including the second order reflections increases the average errors to between 0.32–0.63 ms.

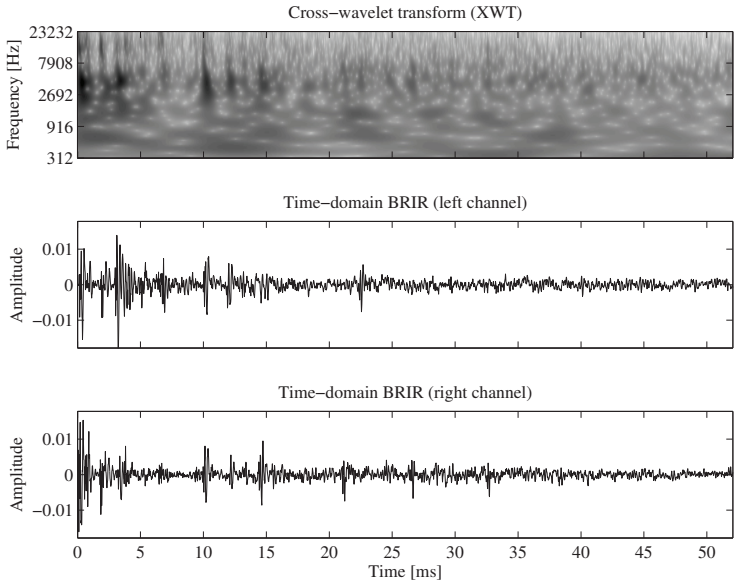

Figure 4.3: The cross-wavelet transform computed from a binaural room impulse response.

With modeled reflections the error is between 0.10–0.35 ms for first and second order reflections. The responses are analyzed only up to 30 ms from the direct sound.

Segmentation of reflections from a BRIR [P4]

In a follow-up report, publication [P4], the method proposed in publication [P3] is extended by segmenting the early reflections based on the XWT using the watershed segmentation algorithm [54] adapted from image processing. Based on the segmented areas (early reflections) of the CWTs of the left and right ear signals, the interaural cues (ILDs and ITDs) are computed and matched with cues computed from HRTFs (the KEMAR [50] and CIPIC [1] HRTF sets) measured at different azimuths and elevations. The best match gives the azimuth and elevation of each reflection. The accuracy of the method is investigated by computing the time and angle errors of the nearest reflections compared to the ground truth values obtained from an image source model. If the nearest reflection is too far (more than ± 1 ms), it is considered as a missed reflection in the evaluation. A comparison to a method consisting of the reflection detector of Kuster [83] and azimuth localization by mapping maximum cross-correlation lag to the angle is made. The idea of resynthesizing the responses with stretched time scales, in order to produce slow-motion versions of the responses for auralization [96], is also investigated.

It is shown that from the studied responses (two measured and two simulated), the times of arrival of the early reflections up to 30 ms are estimated with average errors below 0.4 ms and 0.3 ms for the proposed and baseline methods, respectively. The errors in the estimates for the azimuth

angles of the reflections are in the range 30°–54° for both of the methods. Even though the angle estimates of the proposed method are not very accurate, the proposed method can be useful for slow-motion auralization of measured responses where estimates of the angles are not needed. It should also be noted that the angles of reflections overlapping in time and frequency are very difficult to recover, and such reflections were present in the studied responses, which also contributed to the average error being large when estimating the azimuth angles.

5 APPLICATIONS TO AUGMENTED REALITY AUDIO

This part of the thesis explores how real-time sound signal analysis can be applied in the context of augmented reality audio. Various possibilities for utilizing the information obtained from analysis of the binaural input provided by the augmented reality headset described in Chapter 2 are explored. Common to all the applications described in this chapter is continuous monitoring of the surrounding sound environment and detection of discrete short-time sound events from the sound stream. These events are then analyzed in various ways and the information is utilized either immediately ([P5], [P6]) or later in off-line analysis ([P7]).

Figure 5.1 presents a common framework for the publications described in this chapter. First, the sound of the environment is picked up by a microphone (or two microphones), amplified, and converted to digital form. Then, the sound signal representation is transformed by computing features from the signal. This is followed by segmentation of the signal to discrete audio events. These events are analyzed in real-time and the results of the analysis are then passed to some software application, such as the virtual sound renderer in publication [P5], or music player software in publication [P6].

The first application ([P5]) makes the virtual sound sources fit the surrounding sound environment better, by estimating the reverberation time of the surrounding environment and then adjusting the RT of the virtual sound sources to match that. The second application ([P6]) is a simple user interface based on localizing finger snaps made by the user. This user interface then controls, for instance, the play list of a music player application. The third application ([P7]) is audio surveillance, related to which the sub-topic of the importance of different low-level audio features for unsupervised classification of transient audio events is investigated.

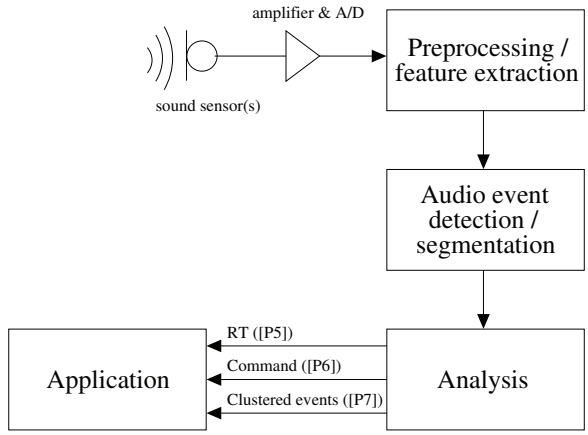

Figure 5.1: The common framework of sound signal analysis applied to ARA in publications [P5], [P6], and [P7].

5.1 Binaural blind estimation of reverberation time for adjusting the reverberation of virtual sources

Related research

Traditionally, the reverberation time of a room is obtained by first measuring the room impulse response using some method. A standard manner for measuring the RT can be found in [148] (also see [47]). The reverberation time can be obtained by inspecting the rate of decay of the energy in an impulse response. Because the energy time curve itself is typically very noisy, backward integration [140] is often applied to obtain a smoother curve, and the RT is then estimated from the slope of a line fitted to the curve. However, there are applications where the measurement of an impulse response is impractical or impossible, but it would still be useful to get an estimate of the RT based on passively received sound signals. Methods that estimate the RT without any prior information about the source signal or the room are called blind RT estimation methods.

One of the first attempts at blind estimation of reverberation time is presented by Hansen [62]. The method makes an assumption on the nature of the signal, so that the method is intended for music signals played back in a reverberant space. Information on the reverberation is extracted from the envelope of the autocorrelation function by employing the Schroeder method [140]. Some agreement with the true RT values is reported.

Couvreur *et al.* [35, 36] present a blind RT estimation method, which is based on a distortion model of the reverberated speech. Given a model for the anechoic speech, the RT can be then estimated from reverberant observations based on maximum likelihood (ML).

Neural networks have been applied in blind RT estimation by Cox *et al.* [38]. They trained a multilayer feedforward network using speech samples convolved with impulse responses having different reverberation times. Another network is used for refining the estimates.

For use in conjunction with their dereverberation algorithm, Lebart *et al.* [87] present a method which is based on detecting the parts of a sound signal containing free decays, and then estimating the RT on those signal parts using linear regression.

Baskind and Warusfel [11] proposed a method for blind estimation of the RT based on locating decaying segments and applying Schroeder integration [140] to the segmented parts. The method utilizes binaural responses by taking the average over the RT estimates of the left and right channels.

Baskind and de Cheveigne [10] presented a pitch-synchronous method for estimating the RT from binaurally recorded reverberant music. The short-time coherence (see Sec. 3.5) is used to find the times for the beginnings and ends of decay analysis. The Schroeder method and linear regression is used for the estimation of RT. Estimation errors were not calculated, but most estimates were found to be between the true early decay time (EDT) and the true reverberation time.

Ratnam *et al.* [129] modeled the late reverberation as exponentially decaying Gaussian noise, and employed a ML approach for estimating the reverberation time. No segmentation is used in the approach, but the

ML estimation was performed by using a sliding window and the final RT estimate was based on the statistics of the running estimates. The method shows good correlation with the true RT values. A computationally efficient real-time version of the algorithm has also been presented [128].

Vieira [163] presented a method that estimates the RT by segmenting the parts of free decay and applying a least squares fit to the logarithmic decay curve obtained by the Schroeder method. Later, the method was improved by using a nonlinear regression approach [170] for blind RT estimation [164].

Wu and Wang [169] derived a monotonic relationship between pitch strength and reverberation time. Their method works only for reverberation times up to 0.6 s.

Zhang *et al.* [179, 178] proposed a blind RT estimation method that can handle noisy situations (e.g. occupied rooms). Blind source separation (BSS) is combined with adaptive noise cancellation to preprocess the signal and remove the noisy disturbance prior to estimating the RT. The denoised signal is segmented to find suitable sections of free decay and the RTs are estimated on the segments using a maximum likelihood procedure with a model proposed by Ratnam *et al.* [129].

Kendrick *et al.* [82] present a method where the method of Ratnam *et al.* [129] is modified by modeling the reverberation with multiple decaying exponentials instead of just one. This is a more realistic assumption in many rooms and allows for flexibility and estimation of the early decay time as well. In a manner similar to [179, 178], the estimation is only carried out on suitable segments with decay. The method proposed in [82] has been compared with the envelope spectrum-based method [91] to estimate the reverberation time and other monaural acoustic parameters [81]. It was found that the enveloped spectrum method is more accurate for the EDT and the maximum likelihood method for the RT for speech and music in real and simulated rooms.

Falk *et al.* [100] have used the auditory modulation spectrum for blind estimation of the RT. A consistency measure between GMMs trained with clean speech and the reverberant test samples was calculated, and the result was mapped to RT.

Wen *et al.* [166] estimated the RT blindly based on distributions of frequency-dependent signal decay rates. The decay rates are calculated from STFTs of speech signals, and certain properties of the distributions of the decay rates are mapped onto the RT.

In conclusion, approaches for blind RT estimation can be roughly categorized into methods that are based on segmenting the input signal in order to find signal segments suitable for analysis (e.g. [62, 38, 87, 11, 10, 163, 164, 82]), and to methods that estimate the RT continuously without segmentation (e.g. [129, 128, 35, 36]). The approach in this thesis falls into the former category. While it lacks the sophistication of the current state-of-the art approaches (e.g. [82, 166]), it presents a relatively straightforward method for estimating the RT for real-time applications. The method is geared towards adjusting the reverberation time of virtual sound source in real-time in the ARA context.

Novel contributions

The method for blind estimation of reverberation time presented in publication [P5] is a real-time method intended for adjusting the reverberation of virtual sound sources in an augmented reality audio context (see Sec. 2.1). The method is based on finding transients from the incoming sound signal and then performing the backward integration method of Schroeder [140] on the free decays following each transient. Linear regression is applied to each segment to obtain an estimate of RT. By inspecting the statistics of the estimates, a final estimate for the RT is obtained.

The upper limit of integration in the Schroeder method is sought by finding the point where the noise floor starts. The starting point of the decay curve, on the other hand, is decided by inspecting the short-time coherence between the left and right ear signals and excluding the part with high coherence, which consists of the direct sound and the first early reflections. Special care is also taken to find the limits of line fitting to the decay curve obtained with the Schroeder method. The starting point is fixed to the time where the decay falls to -5 dB below the maximum of the curve, while the end point is varied from -5 dB to -35 dB and a line is fitted to that range of the decay curve. The RT is obtained from the slope of the line whose end point results in the largest correlation coefficient.

A real-time version of the algorithm running on a Linux workstation was implemented in C++ within the Mustajuuri real-time audio processing framework [73]. The main use of the algorithm in the ARA context is adjusting the RT of virtual sound sources to match that of the environment around the user. In practice, this means that the current estimate of the RT is set as the RT of a reverberation algorithm. In informal listening experiments it was found that matching the RT to that of the surrounding space increases the naturalness of the virtual sound sources, as they now have subjectively similar reverberation characteristics. The just noticeable difference (JND) of RT perception for speech has been found to be between 3.3% and 9.6% [79]. However, in latest studies the JND in human perception of RT from music signals has found to be between 20% and 30% [105]. The estimates obtained using the proposed algorithm are seen to fall within ± 100 ms from the true value when the true RT is 0.8 s (see Fig. 4 in publication [P5]). This range is outside the JND according to [79], but the JND studies do not answer the question of how much the RT can deviate in auralization so that the virtual sound source will still be perceived as matching the surrounding space in terms of the reverberation. The tolerance for deviations from the correct RT could be investigated as a future study.

5.2 Eyes-free user interface based on finger snaps

Related research

User interfaces that do not have visual feedback can be useful in mobile applications, such as when walking or driving a car, where having to concentrate on the visual interface would be distracting [19, 180]. Space constraints of small devices such as personal music players make it difficult to apply visual interfaces to such devices [19, 180]. Visually impaired people also benefit from eyes-free user interfaces, for obvious reasons [180]. The re-

search question addressed in this thesis is the input to such an interface. Head and hand gestures detected by sensors have been used as inputs to such interfaces [19, 104, 78]. A touchpad sensor has also been employed as an input [180]. In this thesis, only transient sounds as input are considered. Speech interfaces are beyond the scope of this work.

Li *et al.* [92] used hand claps and finger snaps as an input in virtual environment applications. An autoregressive model and wavelet coefficients are used as the features fed into a feed-forward back-propagation neural network for classification of the snaps and claps. The system is designed to be used in conjunction with camera-based tracking so that the input sounds trigger events in a manner similar to mouse clicks. An example is given of using the interface in a multimedia kiosk.

Scott and Dragovic [141] constructed a 3-D used interface based on audio input. Loud transient sounds, such as finger snaps and hand claps, are detected and localized in 3-D space using six microphones. The detection is based on simple amplitude thresholding and no classification of the signal content is made. Certain volumes of the space can be defined as buttons, so that a transient sound localized within the volumes creates a command to the user interface. Controlling an MP3 player is given as an example. It is reported that novice users easily understand how to use this user interface.

Jylhä and Erkut [75, 76, 77] have proposed a method for detecting and recognizing different types of hand claps and devised several applications with interaction based on hand claps. These applications include a virtual audience application, music tempo controller, and a sampler.

Novel contributions

In publication [P6], a novel eyes-free user interface with binaural audio input is presented. The method is based on the binaural input provided by the microphones in the ARA headset (see Sec. 2.1). First, transients are detected based on the short-time coherence between the signals (see Sec. 3.5). This is followed by calculating the cross-correlation from the transients and converting the lag of the maximum to azimuth angle. The azimuth angle plane in front of the user is divided into three sectors which correspond to three commands given to the software. As an example, controlling a music player software is presented. Evaluation of the method shows that the interface functions correctly in a quiet office environment.

5.3 Features in audio surveillance

Related research

Automatic audio surveillance monitors the environment based on sound. Many activities of interest make some kinds of sounds, which makes the audio modality a good choice for surveillance — or a good complement to visual modality. Sound diffracts around obstacles, which gives audio surveillance a benefit that video-based surveillance does not have, as line-of-sight between the sensor and the object of interest is not required. Different applications of audio surveillance have been presented in recent years, including an automated audio diary [42], automated airplane sound level measurement [4], surveillance of a living environment [71, 72], elevator surveillance

[126], monitoring of patients [160], scream and gunshot detection [161], and automated verbal aggression detection [162]. Closely related is also the research on environmental sound recognition [53, 159, 37, 40, 6, 29] and auditory context recognition [30, 43, 103, 44, 177].

The topic related to audio surveillance in this thesis is, however, investigating the importance of different features for classification in an audio surveillance application. Härmä *et al.* [71] chose ten features for audio surveillance from a larger set of features based on analysis of their covariances. Mitrovic *et al.* [109] determined the usefulness of a large set of audio features (including MPEG-7 descriptors) for recognizing environmental sounds, based on the factor loading matrix and entropy measures. Defreville *et al.* [40] sought optimal features for urban sound source classification using an automatic feature extraction system based on genetic programming [181]. Valenzise *et al.* [161] use a combination of filter and wrapper approaches for feature selection for a gunshot and scream classifier. Chu *et al.* [29] used matching pursuit for extracting features that are effective in recognizing environmental sounds.

Novel contributions

Publication [P7] presents an investigation into the significance of different features in an audio surveillance application. Short-time, transient-like audio events are collected using a real-time implementation of a foreground/background segmentation algorithm [71]. The events that deviate enough from the background are stored for later off-line analysis. In this analysis, the factor loadings [65] of principal component analysis (PCA) of the features is examined to assess the correlations between individual features. The audio events are clustered then using the K-means [151] and self-tuning spectral clustering [176] algorithms. Based on the manual labeling of the events, it is possible to assess the quality of the clustering with the goal that audio events with a particular label should be in one cluster only. A genetic algorithm (GA) is used to find weighting for the features that maximizes the clustering quality. The idea of using GAs for finding weights for features is taken from [108]. The main contribution of the paper is the insight gained into the importance of individual features based on these weights. It is shown that delta mel-frequency cepstral coefficients and variance features are important when clustering transient acoustic events.

6 SUMMARY OF PUBLICATIONS AND CONTRIBUTIONS OF THE AUTHOR

Publication [P1]

This paper presents a novel method for estimating the distance of a (speech) sound source based on the analysis of a binaural signal. The magnitude-squared coherence is used as the main distance cue in the learning system. The method is able to identify the correct source-to-receiver distance when the source is in front of the listener.

The present author is the sole author of this publication.

Publication [P2]

This is an extended version of the publication [P1]. The method presented in publication [P1] is further investigated and more scenarios are tested to get a clear idea of the limits of the approach. Comparisons to a previous approach [146] are also made. The algorithms are tested in two different positions in two rooms with different reverberation characteristics. It is shown that, in most cases, both methods can identify the distance correctly on a grid with 0.5 m spacing at source azimuth angles 0°, 60°, 90°, and 180°. The proposed method also generalizes to some extent, when the sound source is moved slightly off the training positions.

The present author is the sole author of this publication.

Publication [P3]

This paper presents a novel analysis method for binaural room impulse responses. The method utilizes cross-wavelet transform for time-localizing the individual room reflections from a measured BRIR. The results show that the Paul wavelet has potential in this application due to its better time resolution compared to the Morlet wavelet.

The present author has written 95% of this publication. The present author is solely responsible for all other parts of the work.

Publication [P4]

This paper is a continuation of publication [P3], which dealt with a novel wavelet-based impulse response analysis method. An updated version of the analysis method of publication [P3] is used for extracting the individual room reflections, analyzing their directions-of-arrival, and possibly re-synthesizing a slowed-down version of the measured response in order to hear how the room behaves acoustically. The proposed method is compared to a previous approach. Both methods are shown to find the times of arrival accurately, while estimating the azimuth angle is shown to be difficult, especially when many of the reflections come from the sides.

The present author has written 95% of this publication. The analyzed impulse responses were measured and simulated by Dr. Tapio Lokki. The present author is solely responsible for all other parts of the work.

Publication [P5]

This article describes a novel method for blind estimation of the reverberation time of a room. By analyzing a continuous arbitrary signal recorded in a room, the method iteratively calculates an estimate of the room reverberation time. An important new contribution is the use of variable limits in line fitting when calculating the slope of sound energy decay. This new method increases the accuracy of the reverberation time estimates.

The present author has written 95% of this publication. The present author is solely responsible for all other parts of the work.

Publication [P6]

A novel user interface is presented in this article. Continuous binaural signals recorded from the ears of a user are analyzed. Transient sounds are detected and passed for further analysis in which the azimuth angle for each transient is calculated. The azimuth plane is divided into three sectors ("left", "center", and "right") which translate into one of three commands. A command is executed if the transient is classified as a finger snap and a few consistency checks are passed. The method is also evaluated: 90% of the finger snaps were correctly localized by the algorithm.

The present author has written 95% of this publication. The present author is solely responsible for all other parts of the work.

Publication [P7]

The choice of features affecting the performance of a clustering-based unsupervised audio surveillance system is investigated in this paper. The results show that, in order to differentiate between acoustically similar transient event classes, delta mel-frequency cepstral coefficients and variance features are important.

The present author is the sole author of this publication.

7 CONCLUSIONS

This thesis consisted of different methods to analyze binaural signals (with the exception of publication [P7]). Three different areas were covered in the thesis:

1. Computational sound source distance estimation in rooms (publications [P1] and [P2]).

2. Analysis of early reflections from binaural room impulse responses (publications [P3] and [P4]).

3. Real-time applications of sound signal analysis with specific emphasis on augmented reality audio (publications [P5], [P6], and [P7]).

For the area of computational sound source distance estimation, the thesis contributed a new binaural learning method that can recognize the distance between a speech source and a listener correctly in most cases. Publication [P1] introduced the method, and publication [P2] investigated it further. A comparison to an earlier approach of sound source position learning [146] was also made. It was shown that both methods have their strengths and weaknesses. The baseline method made slightly less errors in recognition and could handle cases where strong reflections compared to the direct sound are present. The proposed method seems to generalize better when there is a small mismatch between training and testing data. The distance recognition capabilities of both methods rely on the effects that the early reflections and late reverberation have on the statistics of binaural cues. This differs from learning the sound source direction, where there is a clearer relationship between the interaural cues themselves and the direction of arrival — at least on the horizontal plane (azimuth angle). As a part of [P2], this difference was investigated.

The main challenge in distance estimation is the difficulty of separating the effects of distance on the sound signal from effects caused by other reasons. It is likely that both high and low level auditory processes are involved in distance perception. In order to successfully mimic these processes computationally, the processes should be understood first. The work presented in this thesis is an important step towards a true 3-D localization system, while also acting as basic research towards a more complete understanding of the challenges in computational distance estimation.

Novel analysis methods for binaural room impulse responses were proposed in publications [P3] and [P4]. These methods allow detecting the times of arrival ([P3]) and also the directions ([P4]) of early reflections present in an impulse response. To the best knowledge of the author, estimating the times and directions of arrival of reflections from a measured BRIR (as is done in [P4]) have not been attempted before. The problem is especially challenging, because there are only two microphone signals available and the diffraction effects of the head make the interpretation of the interaural cues difficult. Also, the presence of noise and other artifacts

in a measured response cause uncertainties in the estimation. Although the results in [P4] are not excellent, they contribute as pioneering attempts to tackle the problem of analyzing individual reflections from binaural responses. Also, the cross-wavelet transform is introduced as a potential tool in the analysis of room acoustics.

The binaural method for blind estimation of reverberation time presented in publication [P5] contributed new ways of setting the integration limits for the Schroeder method. Also, no other reports of actual real-time implementations of blind RT estimation methods were available at the time (in 2004), and not even later, to the best knowledge of the author. An algorithm intended for real-time use has been presented [128], but only an offline implementation is described. The method of publication [P5] was mostly intended for use in adjusting the virtual sources in augmented reality audio (see Section 2.1). It would have most value in that kind of an application. However, since several new methods for blind RT estimation [169, 179, 178, 82, 100, 166] have been proposed since publication [P5], this knowledge should be utilized when designing a real-time algorithm for an ARA context.

A novel eyes-free user interface was presented in publication [P6], showing an example of a sound-based input for augmented reality audio. Although similar interfaces based on transient sounds (hand claps, finger snaps) have been presented, this is the still the only binaural ARA-oriented method found in the relevant literature. By increasing the noise robustness and recognition accuracy of the method, there might be great potential in user interfaces with inputs like this. A great advantage is that there is no need to include special input devices, if the two microphones would be included anyway for ARA purposes. These kinds of interfaces with non-speech are useful in many portable applications, as they do not require any specialized devices and do not require tactile or visual interaction with the devices.

A third application presented in this thesis is audio surveillance. However, the actual novel contribution is a detailed investigation into the importance of different features in unsupervised clustering of short sound events. The method used for quantifying the importance of the features based the goodness of clustering might be useful in other clustering applications as well.

This thesis presented novel methods for the analysis of binaural and monaural audio signals. Most notably, the problem of computational sound source distance learning was investigated. Novel methods for analysis of BRIRs were presented. On the applications side, most importantly a novel user interface with non-speech audio input was introduced. The different research topics shared many similarities, especially on the side of methods used. Most notably, the coherence between left and right ear signals[1] was applied in several publications. Novel uses for coherence were proposed in these publications. The contributions of the thesis will be useful in the area of sound signal analysis and its applications.

[1] Publications [P3] and [P4] used the magnitude of the cross-wavelet transform, which also measures the strength of correlation between two signals.

BIBLIOGRAPHY

[1] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, pages 99–102, New Paltz, NY, USA, October 2001.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[3] J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *The Journal of the Acoustical Society of America*, 62(4):912–915, October 1977.

[4] T. C. Andringa, P. W. J. van Hengel, R. Muchall, and M. M. Nillesen. Aircraft sound level measurements in residential areas using sound source separation. In *Proceedings of the 33rd International Congress and Exposition on Noise Control Engineering*, Prague, Czech Republic, August 2004.

[5] D. H. Ashmead, D. L. Davis, and A. Northington. Contribution of listeners' approaching motion to auditory distance perception. *Journal of experimental psychology: Human perception and performance*, 21(2):239–256, April 1995.

[6] J.-J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–891, August 2007.

[7] J. Backman and M. Karjalainen. Modelling of human directional and spatial hearing using neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1993)*, volume 1, pages 125–128, Minneapolis, MN, USA, April 1993.

[8] M. Barron. The subjective effects of first reflections in concert halls-the need for lateral reflections. *Journal of Sound and Vibration*, 15(4):475–494, 1971.

[9] M. Barron. Measured early lateral energy fractions in concert halls and opera houses. *Journal of Sound and Vibration*, 232(1):79–100, April 2000.

[10] A. Baskind and A. de Cheveigne. Pitch-tracking of reverberant sounds, application to spatial description of sound scenes. In *Proceedings of the AES 24th International Conference on Multichannel Audio: The New Reality*, Banff, Canada, June 2003.

[11] A. Baskind and O. Warusfel. Methods for blind computational estimation of perceptual attributes of room acoustics. In *Proceedings of the AES 22nd International Conference on Virtual, Synthetic, and Entertainment Audio*, pages 402–411, Espoo, Finland, June 2002.

[12] J. R. Beltrán and F. Beltrán. Additive synthesis based on the continuous wavelet transform: a sinusoidal plus transient model. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 2003.

[13] J. R. Beltrán, J. P. de León, and E. Estopiñán. Intermodulation effects analysis using complex bandpass filterbanks. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, pages 149–154, Madrid, Spain, September 2005.

[14] J. Blauert. *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, October 1996.

[15] J. Borish. Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1836, June 1984.

[16] M. Brandstein and D. Ward. *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[17] J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model structure. *The Journal of the Acoustical Society of America*, 110(2):1074–1088, August 2001.

[18] A. S. Bregman. *Auditory scene analysis*. MIT press, Cambridge, Massachusetts, 1990.

[19] S. A. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker. Multimodal 'eyes-free' interaction techniques for wearable devices. In *Proceedings of ACM CHI 2003*, pages 463–480, Fort Lauderdale, FL, USA, April 2003.

[20] A. W. Bronkhorst. Modeling auditory distance perception in rooms. In *Proceedings of Forum Acusticum*, Sevilla, Spain, September 2002.

[21] A. W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517–520, February 1999.

[22] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz. Auditory localization of nearby sources. II. Localization of a broadband source. *The Journal of the Acoustical Society of America*, 106(4):1956–1968, October 1999.

[23] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, September 1999.

[24] D. S. Brungart and K. R. Scott. 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, July 2001.

[25] P. Calamia. Three-dimensional localization of a close-range acoustic source using binaural cues. Master's thesis, University of Texas at Austin, 1998.

[26] G. C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, February 1987.

[27] W. Chau and R. O. Duda. Combined monaural and binaural localization of sound sources. In *Proceedings of the 29th Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1281–1285, Oct/Nov 1995.

[28] J. Chomyszyn. *Distance of sound in reverberant fields*. PhD thesis, Department of Music, Stanford University, Palo Alto, CA, USA, August 1995. Available at http://ccrma.stanford.edu/STANM/stanms/stanm94/. Visited 8th April, 2009.

[29] S. Chu, S. Narayanan, and C.-C. Jay Kuo. Environmental sound recognition using MP-based features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 1–4, Las Vegas, NV, USA, March/April 2008.

[30] B. Clarkson and A. Pentland. Extracting context from environmental audio. In *Proceedings of the IEEE International Symposium on Wearable Computers (ISWC 1998)*, pages 154–155, Washington, DC, USA, October 1998.

[31] R. K. Clifton. Breakdown of echo suppression in the precedence effect. *The Journal of the Acoustical Society of America*, 82(5):1834–1835, November 1987.

[32] A. Cohen and J. Kovačević. Wavelets: The mathematical background. *Proceedings of the IEEE*, 84(4):514–522, April 1996.

[33] P. D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60:302–315, 1963.

[34] M. Cooke, Y.-C. Lu, Y. Lu, and R. Horaud. Active hearing, active speaking. In *Proceedings of the International Symposium on Auditory and Audiological Research*, Helsingør, Denmark, August 2007.

[35] L. Couvreur and C. Couvreur. Robust automatic speech recognition in reverberant environments by model selection. In *Proceedings of the International Workshop on Hands-Free Speech Communication (HSC2001)*, pages 147–150, Kyoto, Japan, April 2001.

[36] L. Couvreur, C. Ris, and C. Couvreur. Model-based blind estimation of reverberation time: Application to robust asr in reverberant environments. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-2001)*, volume 1, pages 2635–2638, Aalborg, Denmark, September 2001.

[37] M. Cowling and R. Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, 2003.

[38] T. J. Cox, F. F. Li, and P. Darlington. Extracting room reverberation time from speech using artificial neural networks. *Journal of the Audio Engineering Society*, 49(4):219–230, April 2001.

[39] G. Defrance, L. Daudet, and J.-D. Polack. Detecting arrivals within room impulse responses using matching pursuit. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pages 297–300, Espoo, Finland, September 2008.

[40] B. Defréville, P. Roy, C. Rosin, and F. Pachet. Automatic recognition of urban sound sources. In *Proceedings of the AES 120th International Convention*, Paris, France, May 2006. Paper no. 6827.

[41] N. I. Durlach. Binaural signal detection — Equalization and cancellation theory. *Foundations of modern auditory theory*, 2:371–462, 1972.

[42] D. P. W. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the 1st ACM Workshop on Continuous Archiving and Recording of Personal Experiences (CARPE 2004)*, pages 39–47, New York, NY, USA, October 2004.

[43] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context awareness — acoustic modeling and perceptual evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 5, pages 529–532, Hong-Kong, China, April 2003.

[44] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329, January 2006.

[45] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116(5):3075–3089, November 2004.

[46] R. R. Fay and A. N. Popper. Evolution of hearing in vertebrates: the inner ears and processing. *Hearing Research*, 149(1–2):1–10, 2000.

[47] A. Gade. Performance and assembly hall acoustics. In T. Rossing, editor, *Springer Handbook of Acoustics*. Springer, 2007.

[48] E. A. Gamble. Minor studies from the psychological laboratory of Wellesley College, I. Intensity as criterion in estimating the distance in sounds. *Psychological Review*, 16:416–426, 1909.

[49] H. Gamper and T. Lokki. Instant hrirs for auditory events in augmented reality audio. In *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing (IWPASH)*, Miyagi Zao Royal Hotel, Japan, November 2009.

[50] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical report, MIT Media Lab Perceptual Computing, 1994. Available at http://sound.media.mit.edu/resources/KEMAR.html. Visited April 1, 2009.

[51] M. B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1):47–53, January 1969.

[52] E. Georganti, T. May, S. van de Par, A. Härmä, and J. Mourjopoulos. Single channel sound source distance estimation based on statistical and source specific features. In *Proceedings of the 126th Convention of the Audio Engineering Society*, Munich, Germany, May 2009. Paper no. 7689.

[53] R. S. Goldhor. Recognition of environmental sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1993)*, volume 1, pages 149–152, Minneapolis, MN, USA, April 1993.

[54] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2001.

[55] B. N. Gover, J. G. Ryan, and M. R. Stinson. Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array. *The Journal of the Acoustical Society of America*, 116(4):2138–2148, October 2004.

[56] D. C. Greene. Comments on "perception of the range of a sound source of unknown strength". *The Journal of the Acoustical Society of America*, 44(2):634–634, 1968.

[57] D. Griesinger. Measurement of acoustic properties through syllabic analysis of binaural speech. In *Proceedings of the International Conference on Acoustics (ICA2004)*, volume 1, pages 29–32, Kyoto, Japan, April 2004.

[58] D. Griesinger. Pitch coherence as a measure of apparent distance in performance spaces and muddiness in sound recordings. In *Proceedings of the AES 121st International Convention*, San Francisco, CA, USA, October 2006. Paper no. 6917.

[59] A. Grinsted, J. C. Moore, and S. Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, 11(5/6):561–566, 2004.

[60] R. Guski. Acoustic tau: an easy analogue to visual tau? *Ecological Psychology*, 4(3):189–197, 1992.

[61] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005(18):2915–2929, October 2005.

[62] M. Hansen. A method for calculating reverberation time from musical signals. Technical Report 60, The Acoustics Laboratory, Technical University of Denmark, Building 352, DK-2800 Lynbgy, 1995.

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

[64] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa. Techniques and applications of wearable augmented reality audio. In *Proceedings of the AES 114th International Convention*, Amsterdam, the Netherlands, March 2003. Paper no. 5768.

[65] H. Harman. *Modern Factor Analysis*. University of Chicago Press, 2nd edition, 1967.

[66] R. S. Heffner and H. E. Heffner. Evolution of sound localization in mammals. In *The Evolutionary Biology of Hearing*, pages 691–715. Springer-Verlag, New York, NY, USA, 1992.

[67] T. Hidaka, L. L. Beranek, and T. Okano. Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls. *The Journal of the Acoustical Society of America*, 98(2):988–1007, August 1995.

[68] T. Hidaka, Y. Yamada, and T. Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *The Journal of the Acoustical Society of America*, 122(1):326–332, July 2007.

[69] H. R. Hirsch. Perception of the range of a sound source of unknown strength. *The Journal of the Acoustical Society of America*, 43(2):373–374, 1968.

[70] A. H. Holway and E. G. Boring. Determinants of apparent visual size with distance variant. *The American Journal of Psychology*, 54(1):21–37, January 1941.

[71] A. Härmä, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005.

[72] A. Härmä, J. Skowronek, and M. F. McKinney. Acoustic monitoring of activity patterns in office, street and garden environments. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2005)*, pages 237–240, Wageningen, The Netherlands, August 2005.

[73] T. Ilmonen. Mustajuuri — an application and toolkit for interactive audio processing. In *Proceedings of the The Seventh International Conference on Auditory Display (ICAD 2001)*, pages 284–285, Espoo, Finland, July/August 2001.

[74] L. A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41:35–39, 1948.

[75] A. Jylhä and C. Erkut. Inferring the hand configuration from hand clapping sounds. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-08)*, pages 301–304, Espoo, Finland, September 2008.

[76] A. Jylhä and C. Erkut. Sonic interactions with hand clap sounds. In *Proceedings of Audio Mostly*, pages 93–100, Piteå, Sweden, October 2008.

[77] A. Jylhä and C. Erkut. A hand clap interface for sonic interaction with the computer. In *CHI EA '09: Proceedings of the 27th international conference*

extended abstracts on Human factors in computing systems, pages 3175–3180, New York, NY, USA, April 2009. ACM.

[78] R. Kajastila and T. Lokki. A gesture-based and eyes-free control method for mobile devices. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, Boston, MA, USA, April 2009.

[79] M. Karjalainen and H. Järveläinen. More about this reverberation science: Perceptually good late reverberation. In *Proceedings of the AES 111th International Convention*, New York, NY, USA, September 2001. Paper no. 5415.

[80] M. Karjalainen, M. Tikander, and A. Härmä. Head-tracking and subject positioning using binaural headset microphones and common modulation anchor sources. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, volume 4, pages 101–104, Montreal, Canada, May 2004.

[81] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers. Monaural room acoustic parameters from music and speech. *The Journal of the Acoustical Society of America*, 124(1):278–287, July 2008.

[82] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, and J. A. Chambers. Blind estimation of reverberation parameters for non-diffuse rooms. *Acta Acustica united with Acustica*, 93:760–770, September/October 2007.

[83] M. Kuster. Reliability of estimating the room volume from a single room impulse response. *The Journal of the Acoustical Society of America*, 124(2):982–993, August 2008.

[84] H. Kuttruff. *Room acoustics*. Taylor & Francis, 2000.

[85] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng. On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, 124(1):450–461, 2008.

[86] E. Larsen, C. D. Schmitz, C. R. Lansing, W. D. O'Brien, B. C. Wheeler, and A. S. Feng. Acoustic scene analysis using estimated impulse responses. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 725–729, Pacific Grove, CA, USA, November 2003.

[87] K. Lebart, J.-M. Boucher, and P. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acustica/Acta Acustica*, 87(3):359–366, 2001.

[88] S.-K. Lee. An acoustic decay measurement based on time-frequency analysis using wavelet transform. *Journal of Sound and Vibration*, 252(1):141–153, 2002.

[89] S.-K. Lee and M.-S. Lee. Reverberation time measurement for an acoustic room with low value of BT by utilizing wavelet transform. *Journal of Sound and Vibration*, 275(3-5):1101 – 1112, 2004.

[90] N. Lesser and D. Ellis. Clap detection and discrimination for rhythm therapy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, volume 3, pages 37–40, Philadelphia, PA, USA, March 2005.

[91] F. F. Li and T. J. Cox. Speech transmission index from running speech: A neural network approach. *The Journal of the Acoustical Society of America*, 113(4):1999–2008, April 2003.

[92] Y. Li, C. Groenegress, J. Denzinger, W. Strauss, and M. Fleischmann. An acoustic interface for triggering actions in virtual environments. In Jizhou Sun and Zhigeng Pan, editors, *Proceedings of SPIE (Fourth International Conference on Virtual Reality and Its Applications in Industry)*, volume 5444, pages 246–251. SPIE, 2004.

[93] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *The Journal of the Acoustical Society of America*, 80(6):1608–1622, December 1986.

[94] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front. *The Journal of the Acoustical Society of America*, 80(6):1623–1630, December 1986.

[95] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng. Localization of multiple sound sources with two microphones. *The Journal of the Acoustical Society of America*, 108(4):1888–1905, October 2000.

[96] T. Lokki. Auralization of simulated impulse responses in slow motion. In *Proceedings of the AES 118th Convention*, Barcelona, Spain, May 2005. Paper no. 6500.

[97] T. Lokki, H. Nironen, S. Vesa, L. Savioja, and A. Härmä. Problem of far-end user's voice in binaural telephony. In *Proceedings of the 18th International Congress on Acoustics (ICA2004)*, volume 2, pages 1001–1004, Kyoto, Japan, April 2004. Invited paper.

[98] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen. Application scenarios of wearable and mobile augmented reality audio. In *Proceedings of the AES 116th International Convention*, Berlin, Germany, May 2004. Paper no. 6026.

[99] S. J. Loutridis. Decomposition of impulse responses using complex wavelets. *Journal of the Audio Engineering Society*, 53(9):796–811, September 2005.

[100] Y.-C. Lu, H. Christensen, and M. Cooke. Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH2007)*, pages 514–517, Antwerp, Belgium, August 2007.

[101] Y.-C. Lu and M. Cooke. Auditory distance perception based on direct-to-reverberant energy ratio. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC 2008)*, Seattle, WA, USA, September 2008.

[102] Y.-C. Lu, M. Cooke, and H. Christensen. Active binaural distance estimation for dynamic sources. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH2007)*, pages 574–577, Antwerp, Belgium, August 2007.

[103] R. G. Malkin and A. Waibel. Classifying user environment for mobile applications using linear autoencoding of ambient audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, volume 5, pages 509–512, Philadelphia, PA, USA, March 2005.

[104] G. Marentakis and S. A. Brewster. A study on gestural interaction with a 3d audio display. In *Proceedings of MobileHCI2004*, volume 3160, pages 180–191, Glasgow, Scotland, 2004.

[105] Z. Meng, F. Zhao, and M. He. The just noticeable difference of noise length and reverberation perception. In *Proceedings of International Symposium on Communications and Information Technologies (ISCIT 2006)*, pages 418–421, Bangkok, Thailand, September/October 2006.

[106] D.H. Mershon and J.N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8(3):311–322, 1979.

[107] K. Meyer, H. L. Applewhite, and F. A. Biocca. A survey of position trackers. *Presence: Teleoperators and Virtual Environments*, 1(2):173–200, 1992.

[108] B. Minaei-Bidgoli and W. F. Punch. Using genetic algorithms for data mining optimization in an educational web-based system. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2003)*, Chicago, IL, USA, July 2003.

[109] D. Mitrovic, M. Zeppelzauer, and H. Eidenberger. Analysis of the data quality of audio features of environmental sounds. *Journal of Universal Knowledge Management*, 1(1):4–17, 2006.

[110] J. Molino. Perceiving the range of a sound source when the direction is known. *The Journal of the Acoustical Society of America*, 53(5):1301–1304, May 1973.

[111] B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, 2003.

[112] S. Müller and P. Massarani. Transfer-function measurement with sweeps. *Journal of the Audio Engineering Society*, 49(6):443–471, June 2001.

[113] M. Naguib and R. H. Wiley. Review: Estimating the distance to a source of sound: Mechanisms and adaptations for long-range communication. *Animal Behaviour*, 62(5):825–837, November 2001.

[114] S. Nielsen. *Distance perception in hearing*. PhD thesis, Aalborg University, Aalborg, Denmark, May 1991.

[115] S. Nielsen. Auditory distance perception in different rooms. *Journal of the Audio Engineering Society*, 41(10):755–770, October 1993.

[116] A. O'Donovan, R. Duraiswami, and N. A. Gumerov. Real time capture of audio images and their use with video. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, pages 10–13, New Paltz, NY, USA, October 2007.

[117] A. O'Donovan, R. Duraiswami, and D. Zotkin. Imaging concert hall acoustics using visual and audio cameras. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 5284–5287, Las Vegas, NV, USA, March/April 2008.

[118] J. W. Strutt (Third Baron of Rayleigh). On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

[119] L. Ottaviani, F. Fontana, and D. Rocchesso. Recognition of distance cues from a virtual spatialization model. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, pages 187–190, Hamburg, Germany, September 2002.

[120] K. Palomäki, V. Pulkki, and M. Karjalainen. Neural network approach to analyze spatial sound. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, pages 233–245, Rovaniemi, Finland, March 1999.

[121] M. Park, P. A. Nelson, and Y. Kim. An auditory process model for sound localization. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2005)*, pages 122–125, New Paltz, NY, USA, October 2005.

[122] M. Park and B. Rafaely. Sound-field analysis by plane-wave decomposition using spherical microphone array. *The Journal of the Acoustical Society of America*, 118(5):3094–3103, November 2005.

[123] M. Peltola, T. Lokki, and L. Savioja. Augmented reality audio for location-based games. In *Proceedings of the AES 35th International Conference on Audio for Games*, London, UK, February 2009.

[124] A. N. Popper and R. R. Fay, editors. *Sound source localization*. Springer New York, 2005.

[125] V. Pulkki and T. Hirvonen. Computational count-comparison models for itd and ild decoding. In *Proceedings of the 19th International Congress on Acoustics (ICA2007)*, Madrid, Spain, September 2007.

[126] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2005)*, pages 158–161, New Paltz, NY, October 2005.

[127] B. Rafaely, I. Balmages, and L. Eger. High-resolution plane-wave decomposition in an auditorium using a dual-radius scanning spherical microphone array. *The Journal of the Acoustical Society of America*, 122(5):2661–2668, November 2007.

[128] R. Ratnam, D. L. Jones, and W. D. O'Brien Jr. Fast algorithms for blind estimation of reverberation time. *IEEE Signal Processing Letters*, 11(6):537–540, 2004.

[129] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing, and A. S. Feng. Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892, November 2003.

[130] M. C. Reed and J. J. Blum. A model for the computation and encoding of azimuthal information by the lateral superior olive. *The Journal of the Acoustical Society of America*, 88(3):1442–1453, 1990.

[131] D. D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *Journal of the Audio Engineering Society*, 37(6):419–444, June 1989.

[132] J.M. Rigelsford and A. Tennant. Acoustic imaging using a volumetric array. *Applied Acoustics*, 67(7):680–688, 2006.

[133] V. Riikonen, M. Tikander, and M. Karjalainen. An augmented reality audio mixer and equalizer. In *Proceedings of the AES 124th International Convention*, Amsterdam, The Netherlands, May 2008. Paper no. 7372.

[134] N. Roman and D. Wang. Binaural tracking of multiple moving sources. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 5, pages 149–152, Hong Kong, May 2003.

[135] N. Roman and D. Wang. Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):728–739, May 2008.

[136] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, October 2003.

[137] S. Roper and T. Collins. The localisation of a sound source in a reverberant room using arrays of microphones. In *Proceedings of the AES 31st International Conference: New Directions in High Resolution Audio*, London, United Kingdom, June 2007. Paper number 28.

[138] S. Roper and T. Collins. A sound sources and reflections localization method for reverberant rooms using arrays of microphones. In *Proceedings of the AES 32nd International Conference: DSP for Loudspeakers*, Hillerød, Denmark, September 2007. Paper number 16.

[139] M. Schönle, N. Fliege, and U. Zölzer. Parametric approximation of room impulse responses based on wavelet decomposition. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 68–71, New Paltz, NY, USA, October 1993.

[140] M. R. Schroeder. A new method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, March 1965.

[141] J. Scott and B. Dragovic. Accurate low-cost location sensing. In *Pervasive Computing*, chapter 1, pages 1–18. Springer-Verlag Berlin / Heidelberg, 2005.

[142] L. Seppänen. Development of an audible sticker application and a video-based tracking system. Master's thesis, Helsinki University of Technology, 2008.

[143] B. G. Shinn-Cunningham. Distance cues for virtual auditory space. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pages 227–230, Sydney, Australia, December 2000.

[144] B. G. Shinn-Cunningham. Localizing sound in rooms. In *Proceedings of the ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoustic Rendering for Virtual Environments*, pages 17–22, Snowbird, Utah, USA, May 2001.

[145] J. A. Simmons, M. B. Fenton, and M. J. O'Farrell. Echolocation and pursuit of prey by bats. *Science*, 203(4375):16–21, January 1979.

[146] P. Smaragdis and P. Boufounos. Position and trajectory learning for microphone arrays. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):358–368, 2007.

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

[148] ISO Standard. 3382. Acoustics–Measurement of the reverberation time of rooms with reference to other acoustical parameters. *International Standards Organization*, 1997.

[149] R. M. Stern, G. J. Brown, and D. Wang. Binaural sound localization. In D. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, chapter 5. WILEY—IEEE Press, October 2006.

[150] R. M. Stern and C. Trahiotis. Models of binaural interaction. In B. C. J. Moore, editor, *Hearing*, chapter 10. Academic Press, 1995.

[151] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[152] M. Tikander. Sound quality of an augmented reality audio headset. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, pages 178–181, Madrid, Spain, September 2005.

[153] M. Tikander. Modeling the attenuation of a loosely-fit insert headphone for augmented reality audio. In *Proceedings of the AES 30th International Conference on Intelligent Audio Environments*, Saariselkä, Finland, March 2007.

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

[155] M. Tikander, A. Härmä, and M. Karjalainen. Binaural positioning system for wearable augmented reality audio. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2003)*, pages 153–156, New Paltz, New York, USA, October 2003.

[156] M. Tikander, A. Härmä, and M. Karjalainen. Acoustic positioning and head tracking based on binaural signals. In *Proceedings of the AES 116th International Convention*, Berlin, Germany, May 2004. Paper no. 6124.

[157] M. Tikander, M. Karjalainen, and V. Riikonen. An augmented reality audio headset. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pages 181–184, Espoo, Finland, September 2008.

[158] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

[159] Y. Toyoda, J. Huang amd S. Ding, and Y. Liu. Environmental sound recognition by multilayered neural networks. In *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04)*, pages 123–127, Beijing, China, September 2004.

[160] M. Vacher, D. Istrate, J.-F. Serignat, and N. Gac. Detection and speech/sound segmentation in a smart room environment. In *Proceedings of the 3rd International Conference on Speech Technology and Human-Computer Dialogue*, pages 37–48, Cluj, Romania, May 2005.

[161] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, pages 21–26, London, UK, September 2007.

[162] P. W. J. Van Hengel and T. C. Andringa. Verbal aggression detection in complex social environments. In *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, pages 15–20, London, UK, September 2007.

[163] J. Vieira. Automatic estimation of reverberation time. In *Proceedings of the AES 116th International Convention*, Berlin, Germany, May 2004. Paper no. 6107.

[164] J. Vieira. Estimation of reverberation time without test signals. In *Proceedings of the AES 118th International Convention*, Barcelona, Spain, May 2005. Paper no. 6499.

[165] H. Viste and G. Evangelista. Binaural source localization. In *7th International Conference on Digital Audio Effects (DAFx-04)*, pages 145–150, Naples, Italy, October 2004.

[166] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor. Blind estimation of reverberation time based on the distribution of signal decay rates. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 329–332, Las Vegas, NV, USA, March/April 2008.

[167] T. Wittkopp. *Two-channel noise reduction algorithms motivated by models of binaural interaction*. PhD thesis, Carl von Ossietzky University Oldenburg, March 2001.

[168] P. J. Wolfe and S. J. Godsill. Audio signal processing using complex wavelets. In *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, March 2003. Paper no. 5829.

[169] M. Wu and D. Wang. A pitch-based method for the estimation of short reverberation time. *Acta Acustica united with Acustica*, 92(2):337–339, March/April 2006.

[170] N. Xiang. Evaluation of reverberation times using a nonlinear regression approach. *The Journal of the Acoustical Society of America*, 98(4):2112–2121, 1995.

[171] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846, April 2002.

[172] P. Zahorik. Auditory display of sound source distance. In *Proceedings of the 8th International Conference on Auditory Display (ICAD 2002)*, Kyoto, Japan, July 2002.

[173] P. Zahorik. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117, November 2002.

[174] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420, May/June 2005.

[175] P. Zahorik and F. L. Wightman. Loudness constancy with varying sound source distance. *Nature neuroscience*, 4:78–83, 2001.

[176] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, Cambridge, MA, 2005.

[177] Z. Zeng, X. Li, X. Ma, and Q. Ji. Adaptive context recognition based on audio signal. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008)*, pages 1–4, Tampa, FL, USA, December 2008.

[178] Y. Zhang, J. A. Chambers, P. Kendrick, T. J. Cox, and F. F. Li. Acoustic parameter extraction from occupied rooms utilizing blind source separation. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 1208–1215. Springer Berlin / Heidelberg, 2006.

[179] Y. Zhang, J. A. Chambers, F. F. Li, P. Kendrick, and Cox T. J. Blind estimation of reverberation time in occupied rooms. In *Proceedings of the European Signal Processing Conference (EUSIPCO 2006)*, Florence, Italy, September 2006.

[180] S. Zhao, P. Dragicevic, M. Chignell, R. Balakrishnan, and P. Baudisch. earPod: Eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of ACM CHI 2007*, pages 1395–1404, San Jose, CA, USA, April/May 2007.

[181] A. Zils and F. Pachet. Automatic extraction of music descriptors from acoustic signals using EDS. In *Proceedings of the 116th AES Convention*, Berlin, Germany, May 2004. Paper no. 6127.

[182] U. Zimmer and E. Macaluso. High binaural coherence determines successful sound localization and increased activity in posterior auditory areas. *Neuron*, 47(6):893–905, September 2005.