Macquarie University

Doctoral Thesis

Assessing Spatial Hearing Using Laboratory-Based Real-World Environments

Author:
Dipl.-Ing. Tobias Weller

Supervisors:
Dr. Jörg M. Buchholz
Dr. Virginia Best

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy at the

Department of Linguistics, Faculty of Human Sciences
&
National Acoustic Laboratories

February 2016

Declaration of Authorship

I, Tobias Weller, declare that this thesis titled, *"Assessing Spatial Hearing using Laboratory-Based Real-World Environments"* and the work presented in it are my own. I confirm that:

- This thesis has been submitted solely to Macquarie University for consideration for the doctoral degree.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- Ethics approvals have been obtained from the Australian Hearing Human Research Ethics Committee. The signed approvals with numbers AHHREC2012-3 and AHHREC2014-23, can be found in Appendix D.

Signed: _____

Date: _____

Abstract

Spatial hearing is a crucial aspect of the way we perceive our environment. It helps us to orientate ourselves within an acoustic scene, follow a conversation in a noisy background, avoid potential dangers or simply enjoy the aesthetics of sound, e.g. in music. Hearing impaired listeners often report difficulties with spatial hearing, and current hearing devices frequently do not improve those deficits but rather disturb the cues that the auditory system uses to extract spatial information from the sound reaching our ears. To adequately predict the effect of digital hearing devices on the spatial perception of sound, tests are needed that reliably reproduce acoustic environments that listeners encounter in their daily life. This thesis investigated some of the challenges that are met when spatial hearing is assessed in laboratory-based environments which aim to recreate such realistic scenarios.

The first three studies investigated the factors affecting audibility in a reverberant multi-talker background. First, masked thresholds for the detection of a short speech stimulus were measured in normal hearing (NH) and hearing impaired (HI) listeners. Then, the effect of uncertainty about the location of the target sound source on those thresholds was examined. The results showed that audibility in the analyzed environment depended heavily on the location of the target source. Moreover, masked thresholds were generally higher in the HI group and varied less over target location than in the NH group. An analysis with an auditory detection model applied to predict the measured data suggested that this behavior is linked to the dominance of high frequency components of the target signal and their interaction with the head shadow, reverberation and hearing loss. The effect of spatial uncertainty on audibility was found to be negligible. In the fourth study, a new method to assess spatial hearing in complex environments was presented. Instead of single-source localization, the task in this study was the identification and localization of multiple concurrent talkers in a simulated reverberant environment. The results suggested that the applied method provided a reliable measure of the listeners' ability to localize the direction of the talkers within the horizontal plane, but not to measure their distance. This limitation was linked to the considered stimuli, which did not allow an unambiguous identification of the individual sources.

The results of this thesis demonstrate that it is important to control target audibility when assessing spatial hearing in complex environments. This is best done with an auditory detection model, but the model developed here needs to be further improved before it can be used to reliably predict masked thresholds for individual hearing impaired listeners. Finally, it is suggested that a multi-talker localization task similar to the one presented here may be a useful tool to analyze the effects of hearing loss and hearing devices on spatial hearing with a high ecological validity.

Acknowledgements

Writing a doctoral thesis proved to be not as easy as I thought when I first arrived in Sydney more than three years ago. I had always enjoyed working in research and was curious how life in the antipodes would be like. But there were moments when the feeling of isolation that comes with working alone on a project for several years almost got the better of me. In the end, I have never regretted making the move and this is not least thanks to the following people and institutions:

My principal supervisor Jörg, whom I have to thank for many things: Not only did he guide me in my shaky attempts to make a contribution to the scientific world, but he also offered his support outside of university life whenever I needed it. He is the person that made this project possible to begin with, and is also responsible for this project to come to a satisfactory conclusion.

My adjunct supervisor Gin who managed to infect me with her enthusiasm whenever I felt like I was hitting a dead end. She was always available for discussions and offered advice when it was needed despite being located on the other side of the globe.

Several people at NAL that helped me out in one way or another (in no specific order): Bram, Greg, John, Scott, Taegan, Els, Vivian, Jess, Katrina, Gitte and Harvey for giving me the opportunity to join NAL for the time of my PhD.

The HEARing CRC and Macquarie University, who supported my financially.

Hiranya and Collette at the Department of Linguistics and Jane and Isabelle at the HEARing CRC for their help and support with administrative matters.

All the volunteers that participated in my experiments and produced some really nice sets of data for me to analyze.

Most importantly, my sincerest gratitude goes to Natascha for her unconditional support. Thank you.

Contents

Declaration of Authorship iii

Abstract v

Acknowledgements vii

List of publications xiii

Abbreviations xv

1 Introduction 1

1.1 Aims of this thesis . . . 4

1.2 Thesis organisation . . . 4

2 Auditory masking of speech in reverberant multi-talker environments 7

2.1 Introduction . . . 8

2.2 Methods . . . 11

2.2.1 Experiment . . . 11

2.2.1.1 Subjects . . . 11

2.2.1.2 Setup . . . 11

2.2.1.3 Stimuli . . . 11

2.2.1.4 Procedure . . . 14

2.2.2 Auditory detection model . . . 14

2.2.2.1 Auditory front end . . . 15

2.2.2.2 Decision device . . . 16

2.2.2.3 Input signal generation and model implementation . . . 18

2.3 Results . . . 19

2.3.1 Experiment . . . 19

2.3.2 Model predictions . . . 20

2.4 Discussion . . . 21
2.5 Conclusions . . . 26

3 The effect of sensorineural hearing loss on auditory masking of speech in reverberant multi-talker environments 27
3.1 Introduction . . . 28
3.2 Methods . . . 29
3.2.1 Experimental methods . . . 29
3.2.1.1 Subjects . . . 29
3.2.1.2 Stimuli and setup . . . 29
3.2.1.3 Procedure . . . 31
3.2.2 Auditory detection model . . . 32
3.2.2.1 Auditory front-end . . . 32
3.2.2.2 Decision device . . . 33
3.2.2.3 Input signals . . . 33
3.3 Results . . . 34
3.4 Discussion and conclusions . . . 37

4 The effect of spatial uncertainty on speech detection in a reverberant multi-talker environment 41
4.1 Introduction . . . 42
4.2 Methods . . . 44
4.2.1 Subjects . . . 44
4.2.2 Stimuli and setup . . . 45
4.2.3 Procedure . . . 47
4.3 Results . . . 49
4.4 Discussion . . . 51
4.5 Conclusions . . . 55

5 Multi-talker localization in reverberant environments of varying complexity 57
5.1 Introduction . . . 58
5.2 Methods . . . 60
5.2.1 Subjects . . . 60
5.2.2 Setup . . . 60
5.2.3 Stimuli . . . 61
5.2.4 Procedure . . . 63
5.2.5 Analysis methods . . . 65
5.3 Results . . . 67
5.4 Discussion . . . 69

5.4.1 General observations . . . 69
5.4.2 Analysis method . . . 71
5.4.3 Front-back confusions and head movement . . . 72
5.5 Conclusions . . . 73

6 Conclusions **75**
6.1 Discussion and outlook . . . 77

A Frequency dependent regularization of a mixed-order Ambisonics encoding system using psychoacoustically motivated metrics **81**
A.1 Introduction . . . 82
A.2 Background . . . 83
A.2.1 Mixed-order Ambisonics . . . 83
A.2.2 Shape matching . . . 85
A.3 Methods . . . 86
A.3.1 Microphone array . . . 86
A.3.2 Regularization . . . 87
A.3.3 Loudspeaker array . . . 89
A.3.4 Reproduction error . . . 90
A.3.5 Audibility of microphone noise . . . 90
A.3.6 Optimization . . . 92
A.4 Results . . . 93
A.5 Discussion . . . 95

B Validation of the loudspeaker-based room auralization system (LoRA) **97**
B.1 Favrot and Buchholz (2010) . . . 97
B.2 Oreinos (2015) . . . 98

C Experimental Data **101**

D Ethics approval **105**

Bibliography **109**

List of publications

Journal articles and letters

Weller, T., Buchholz, J. M. and Best, V. (**2015**), "Auditory masking of speech in reverberant multi-talker environments", J. Acoust. Soc. Am., *submitted*.

Weller, T., Buchholz, J. M. and Best, V. (**2015**), "The effect of spatial uncertainty on speech detection in a reverberant multi-talker environment", J. Acoust. Soc. Am., *submitted*.

Weller, T., Buchholz, J. M. and Best, V. (**2015**), "Multi-talker localization in reverberant environments of varying complexity", Int. J. Aud., *submitted*.

Weller, T., Buchholz, J. M. and Best, V. (**2015**), "The effect of sensorineural hearing loss on auditory masking of speech in reverberant multi-talker environments", J. Acoust. Soc. Am. Express Letters, *in prep*.

Conference articles

Weller T., Best V. and Buchholz J. M. (**2013**). "Factors influencing target detectability in realistic listening scenarios", *21st International Conference on Acoustics ICA (Montreal, Canada)*.

Weller, T., Buchholz, J. M. and Oreinos, C. (**2014**), "Frequency dependent regularization of a mixed-order Ambisonics encoding system using psychoacoustically motivated metrics", *Audio Engineering Society 55th Conference on Spatial Audio, (Helsinki, Finland)*.

Weller, T., Buchholz, J. M. and Best, V. (**2014**), "Modelling binaural detection of speech stimuli in complex reverberant environments", *Forum Acusticum 2014, (Krakow, Poland)*.

Published abstracts

Weller, T., Buchholz, J. M. and Best, V. (**2014**), "Measuring spatial awareness in realistic scenarios", *International Hearing Aid Research Conference (IHCON), (Lake Tahoe, USA)*.

Abbreviations

3-AFC	3-**A**lternative **F**orced-**C**hoice
4-FAHL	4-**F**requency **A**verage **H**earing **L**oss
AMB	**Amb**isonics
ANL	**A**cceptable **N**oise **L**evel
ANOVA	**An**alysis **O**f **Va**riance
BRIR	**B**inaural **R**oom **I**mpulse **R**esponse
CRC	**C**ooperative **R**esearch **C**entre
CRM	**C**oordinated **R**esponse **M**easure
CV	**C**onsonant **V**owel
DI	**D**irectivity **I**ndex
DR	**D**irect-to-**R**everberant
EDT	**E**arly **D**ecay **T**ime
ERB	**E**quivalent **R**ectangular **B**andwidth
FIR	**F**inite **I**mpulse **R**esponse
GUI	**G**raphical **U**ser **I**nterface
HA	**H**earing **A**id
HATS	**H**ead **A**nd **T**orso **S**imulator
HI	**H**earing **I**mpaired
HL	**H**earing **L**oss
HOA	**H**igher-**O**rder **A**mbisonics
HRTF	**H**ead-**R**elated **T**ransfer **F**unction
IACC	**I**nter**a**ural **C**ross **C**orrelation
IID	**I**nteraural **I**ntensity **D**ifference
ILD	**I**nteraural **L**evel **D**ifference
ITD	**I**nteraural **T**ime **D**ifference
JND	**J**ust **N**oticeable **D**ifference
LoRA	**Lo**oudspeaker-based **R**oom **A**uralization

MAP	**M**inimial **A**udible **P**ressure
MOA	**M**ixed-**O**rder **A**mbisonics
NAL	**N**ational **A**coustic **L**aboratories
NH	**N**ormal **H**earing
PSD	**P**ower **S**pectral **D**ensity
RIR	**R**oom **I**mpulse **R**esponse
RMS	**R**oot **M**ean **S**quare
SMM	**S**hape **M**atching **M**ethod
SNR	**S**ignal-to-**N**oise **R**atio
SOFE	**S**imulated **O**pen-**F**ield **E**nvironment
SPL	**S**ound **P**ressure **L**evel
SSQ	**S**peech, **S**patial and **Q**ualities of hearing scale
STI	**S**peech **T**ransmission **I**ndex
VSE	**V**irtual **S**ound **E**nvironment
WNG	**W**hite **N**oise **G**ain

Chapter 1

Introduction

Spatial hearing can be described as the ability of listeners to use location-dependent information in the sound impinging on their ears to their benefit. In its simplest form, spatial hearing is equivalent to the localization of sound sources, which can have a number of purposes. In ancient times, localization was essential to avoid dangers, but even today it can be important, e.g. when we participate in traffic. Localizing sounds also contributes to a feeling of space and it helps us to pair auditory impressions with their visual counterpart, and thus making sense of our surroundings. However, spatial hearing goes beyond localization. It also enables us to use the location of a sound source to focus our attention in its direction and filter out distracting sounds from other sources. This is a major component of the so-called "cocktail-party phenomenon" formulated by Cherry (1953), which relates to the astonishing ability of human listeners to follow a conversation even in the presence of multiple interfering talkers in the immediate surroundings.

To localize sounds, the auditory system mainly relies on interaural difference cues. In the horizontal plane, localization is dominated by the interaural time difference (ITD) for frequencies below approximately 1500 Hz. Sounds coming from the front of a listener will reach both ears at the same time, but sounds coming from either side of the listener arrive earlier at one ear than at the other. For higher frequencies, however, fine structure ITD cues become ambiguous and cannot longer be used for localization. Even though envelope-ITDs are still available at higher frequencies (Blauert, 1997), these are usually less robust. Therefore, at frequencies above 1500 Hz, the auditory system mostly depends on interaural level differences (ILDs, also sometimes referred to as interaural intensity differences, IIDs) to localize sounds. At these frequencies, the head attenuates sounds reaching the ears depending on the direction of arrival. Spectral cues, introduced by reflections from the pinna, play a minor role in the localization of sounds in the horizontal plane and are only available at high frequencies ($>$ 4000 Hz). These spectral cues, however, are the main cues used by the auditory system to resolve front-back confusions and to localize sounds in the median plane, i.e. at elevated angles. Nevertheless, there is some redundancy in the auditory system and monaural cues (pinna) and binaural cues (ITDs, ILDs) are interchangeable to some extent (Butler, 1987).

When measured in quiet, hearing impaired (HI) listeners showed a reduced ability to localize sound in the horizontal plane (e.g. Durlach *et al.*, 1981; Noble *et al.*, 1994), but the difficulties to localize were mainly correlated with substantial low-frequency hearing loss (hearing loss at 500 Hz and 1000 Hz) or caused by conductive hearing loss. When measuring localization of a click train in white noise, Lorenzi *et al.* (1999a,b) also found that at least part of the tested HI listeners performed worse than their normal hearing peers. However, they also measured the detection of a click train in white noise and concluded that the reduced localization performance was most likely linked to decreased audibility of the target stimulus. In contrast, HI listeners scored considerably worse than NH listeners in a questionnaire-based study conducted by Noble *et al.* (1995), where they were asked to rate their own localization abilities in a number of questions. When HI listeners in this study were divided into two groups according to their four-frequency average hearing loss (4-FAHL, average hearing loss at 500, 1000, 2000 and 4000 Hz), the group with the more severe hearing loss (> 50 dB) also scored slightly worse than the group with the less severe hearing loss (< 50 dB). The apparent contradiction between experimental findings and the self-assessment of HI listeners raises the question as to how well tests in relatively simple and unrealistic conditions are able to reflect the situations that listeners face in their everyday life.

Unfortunately, current digital hearing aids do little to preserve spatial hearing abilities, except when decreased audibility is the cause for poorer localization performance. On the contrary, several studies measuring localization in the horizontal plane have reported that hearing aid users fitted with bilateral hearing aids performed worse with their hearing aids than without. Van den Bogaert *et al.* (2006) measured bilateral hearing aid users without their hearing aids, with their hearing aids using an omnidirectional microphone and with their hearing aids using an adaptive directional microphone. Across four different stimuli, subjects consistently localized the target most accurately without any hearing aids and least accurately with their hearing aids using an adaptive directional microphone. They attributed these findings to the placing of the microphones on the hearing aid and the distortion of spatial cues caused by different signal processing blocks of the hearing aids. Keidser *et al.* (2006) measured the effect of different hearing aid signal processing strategies on interaural difference cues and localization performance of hearing aid users. They found a small detrimental effect of noise reduction on localization performance but no significant effect of wide dynamic range compression. The results of these studies and the advent of binaural hearing aids (bilateral hearing aids with a signal link between left and right) highlight the need for test methods that can reliably predict the effects of hearing aid signal processing on spatial cues.

There are different possibilities to conduct experiments under realistic conditions. Field

tests are naturally very close to real-life situations but suffer from problems like low reliability and low reproducibility. Headphone-based virtual sound environments (VSEs) are a flexible tool to reproduce the same environment for multiple listeners, but make it difficult to incorporate natural head movements (even with elaborate head-tracking equipment). Additionally, it is not straight-forward to accurately replicate the effect of hearing aids using headphone-based VSEs, especially if all subjects in an experiment are measured with their individually fitted hearing aids. Völk and Fastl (2010) discussed some of the possible applications of headphone-based VSEs in the research of hearing aids and cochlear implants. Wiggins and Seeber (2012) used a headphone-based VSE to measure the effect of independent compression in bilateral hearing aids on the spatial perception of sound. Using a number of different stimuli spatialized using non-individualized head-related transfer functions (HRTFs), they measured different spatial attributes of sound in eleven NH listeners using a questionnaire. The results of this study showed that compression, that is applied independently to both ears, deteriorates the spatial perception of sound, mainly due to conflicts between binaural cues caused by dynamic changes of ILDs. In a study measuring localization of sound sources in realistic environments, Mueller *et al.* (2012) used individualized HRTFs and miniature speakers inside the subjects' ear canal to create different sound scenes that are relevant to real-life listening. They observed localization performance that was similar to previously reported results in simpler environments, but noted that front-back confusions in one of the sound scenes they used were increased significantly, which they attributed to the inability of their system to cope with head movements. They concluded that VSEs may be an important tool to analyze hearing device algorithms in more dynamic environments that are closer to those situations that are encountered in real life.

In contrast to headphone-based VSEs, natural head movements are possible in loudspeaker-based VSEs and it is relatively simple to test listeners using their personal hearing aids. Loudspeaker-based VSEs have been suggested for the use in hearing device research at multiple points throughout literature. Seeber *et al.* (2010) presented a loudspeaker-based environment called the "Simulated Open-Field Environment" (SOFE), which combines the spatial reproduction of sound with a video projection facility, enabling the analysis of audio-visual phenomena. Favrot and Buchholz (2010) described a framework to create sound scenes in a loudspeaker-based VSE using room acoustic simulations, and Minaar *et al.* (2010) proposed the use of loudspeaker-based VSEs in hearing aid research. Minaar *et al.* (2013) presented a feasibility study where they used microphone array recordings of different acoustic scenes and loudspeaker playback to measure hearing aid program preference in ten HI listeners. Subjects in this study used a prototype hearing aid (HA) that could be switched between two different directional microphone settings and two different noise reduction algorithms. Minaar

et al. (2013) concluded that their VSE was well suited for this kind of test and offered a more natural listening experience to the listeners than a headphone-based system. Oreinos (2015) introduced a loudspeaker-based VSE based on Mixed-Order Ambisonics (MOA) microphone array recordings and playback. He compared the benefit of a hearing aid beamformer on speech intelligibility and acceptable noise level (ANL) in 18 HI listeners measured in a real room and in a reproduction of the same room in the VSE. The results of the measurement in the VSE showed the same tendencies as the results in the real room, but the benefit of the beamformer was slightly underestimated.

These studies show that recreating challenging real-world situations using loudspeaker-based VSEs is a promising approach for hearing device research (hearing aids and cochlear implants alike), but they also show that there is still a lot of room for further work in this field. Moreover, very little research has been conducted on improving the realism of the actual test paradigms (or listener tasks) that are used within these VSEs. A more detailed description of the studies conducted by Favrot and Buchholz (2010) and Oreinos (2015) is given in appendix B.

1.1 Aims of this thesis

Since relatively little work has been done so far on the assessment of spatial hearing in realistic environments, the nature of this thesis is to some degree exploratory. It aims at

- gaining a better understanding of audibility in realistic conditions and its effect on localization,
- finding methods to best measure audibility in realistic conditions with the purpose of controlling audibility in localization experiments (e.g. through auditory modeling),
- analyzing how the characteristics of localization tasks affect audibility in realistic conditions,
- developing a new multi-source localization paradigm, highlighting the challenges and problems and finding suitable solutions for those.

1.2 Thesis organisation

This thesis is organized in six chapters. Chapter 2 deals with the control of target audibility in a complex acoustic environment. It is important to ensure target audibility

in localization tasks in order to analyze effects that are not linked to decreased sensitivity as measured by an audiogram. In this context, Chapter 2 describes an experiment measuring the detection of a short speech target in a continuous multi-talker background in a group of NH listeners. The experiment was conducted in a loudspeaker-based VSE simulating a large, reverberant room. The masked thresholds were measured from 15 different target locations across the simulated room, while the locations of the distracters remained the same throughout the entire experiment. Subsequently, a simple auditory model was adapted for the use with binaural stimuli and used to predict the results of the experiment. Finally, the model was analysed to examine the main factors influencing target audibility in the chosen environment.

Chapter 3 repeats the measurement described in Chapter 2 with a group of HI listeners with a number of different audiometric configurations. The model applied in Chapter 2 was adjusted to be able to deal with individual hearing impairment as measured in an audiogram. Based on the insights gained by the analysis of the model in Chapter 2, Chapter 3 discusses the results of the experiment with the HI group and compares them to the results of the NH group. Finally, the advantages and disadvantages of using the described model to predict masked thresholds of HI listeners are discussed.

Chapter 4 is to be understood as a continuation of the previous chapters. Here, an experiment is described that measures the effect of uncertainty about the target location on masked thresholds measured with the methods presented in Chapters 2 and 3. Since localization tasks naturally introduce this kind of spatial uncertainty it is important to examine if this has an impact on target audibility. The experiment was conducted in the same environment used in Chapters 2 and 3 and all NH and HI subjects participating in the experiment had already participated in the study described in the previous chapters. Chapter 4 concludes by discussing the outcomes of the experiment and examining the relevance of the observed effect of spatial uncertainty.

Chapter 5 explores a novel method to assess spatial hearing in a complex environment. In this method, listeners had to analyze sound scenes in a reverberant background with respect to the number of talkers present in the scene and their location in the simulated room. For the purpose of this study, head movements were allowed to make the procedure more natural. The results of the study are discussed and the effects of the applied method on the results are examined.

Finally, Chapter 6 offers some concluding remarks and discusses the contribution made in this thesis. The results are summarized and their relevance to the field are pointed out. Some suggestions are given for future research expanding on the insights gained in this work.

Part of the work done for this thesis also included the development of a microphone array for the recording of acoustic environments for the use in auditory research. Appendix A presents a study that is not explicitly connected to the main content of this thesis, in which a method is analyzed to regularize a Mixed-Order Ambisonics (MOA) encoding system based on metrics that consider perceptual effects of MOA playback in a loudspeaker-based VSE.

Chapter 2

Auditory masking of speech in reverberant multi-talker environments[1]

Auditory localization research needs to be performed in more realistic testing environments to better capture the real-world abilities of listeners and their hearing devices. However, there are significant challenges involved in controlling the audibility of relevant target signals in realistic environments. To understand the important aspects influencing target detection in more complex environments, a reverberant room with a multi-talker background was simulated and presented to the listener in a loudspeaker-based virtual sound environment. Masked thresholds of a short speech stimulus were measured adaptively for multiple target source locations in this scenario. It was found that both distance and azimuth of the target source have a strong influence on the masked threshold. Subsequently, a functional model was applied to analyze the factors influencing target detectability. The model is comprised of an auditory front-end that generates an internal representation of the stimuli in both ears, followed by a decision device combining d' information across time, frequency and both ears. The model predictions of the masked thresholds were overall in very good agreement with the experimental results. An analysis of the model processes showed that head shadow effects, signal spectrum and reverberation have a strong impact on target audibility in the given scenario.

[1]Manuscript submitted for publication to The Journal of the Acoustical Society of America.

2.1 Introduction

Auditory localization is an important aspect of the listener's ability to perceive sound, make sense of his surroundings, direct attention to a sound of interest, follow multi-talker conversations, or segregate speech in a multi-talker background (Freyman *et al.*, 1999). However, it has often been shown that hearing loss can have a detrimental effect on localization in noise (Lorenzi *et al.*, 1999a) and listeners with a highly asymmetric or a severe low frequency hearing loss even show reduced localization performance in quiet (Häusler *et al.*, 1983; Rosenhall, 1985; Noble *et al.*, 1994). Moreover, Akeroyd and Guy (2011) showed a significant reduction in localization dominance, a phenomenon that is associated with the precedence effect and expected to help localization in reverberant environments, which was highly correlated with hearing loss. Finally, Noble *et al.* (1995) and Noble and Gatehouse (2006) showed in questionnaire-based field studies that hearing impaired listeners experience localization difficulties in their daily life, which is more severe with increasing hearing loss.

A number of studies (e.g. Byrne and Noble, 1998; van den Bogaert *et al.*, 2006, 2011; Keidser *et al.*, 2006) have shown that as long as audibility is not an issue, state-of-the-art hearing aids do not improve localization performance, and that even bilateral hearing aid users performed worse in the aided condition of a localization task than in the unaided condition. The distortion of spatial cues by hearing aids can be explained by a number of different factors (e.g. Dillon, 2012): The placement of the hearing aid microphones (which in the case of behind-the-ear hearing aids for example bypasses pinna cues), independent compression between left and right ears distorting interaural level difference cues (Wiggins and Seeber, 2012), and more advanced features (e.g. adaptive beamformers, noise suppression, de-reverberation) removing or modifying interaural time and level difference cues.

Unfortunately, most localization studies apply rather artificial stimuli with a low ecological validity and their results are therefore hard to translate into real-life performance. Different loudspeaker-based virtual sound environments (VSEs) (Minaar *et al.*, 2010; Favrot and Buchholz, 2010; Seeber *et al.*, 2010) have been proposed as a means to reproduce more realistic scenarios in a controlled laboratory setting. Compared to headphone-based VSEs (e.g. Blauert, 1997) that use measured or simulated binaural room impulse responses (BRIRs), a loudspeaker-based VSE has the advantage that listeners can wear their hearing aids during testing without any additional technical effort. This makes it possible to compare normal hearing, unaided and aided hearing impaired performance in spatial hearing tasks using the same realistic, complex scenarios.

A major concern in studies that measure localization in noise is the audibility of the target stimulus, in particular in hearing impaired subjects (e.g. Le Goff *et al.*, 2013). Stimulus audibility is expected to be even more problematic in paradigms that assess localization in more complex environments. On the one hand the SNR should be in a realistic range, on the other hand the target stimulus has to be presented well above threshold. If the target stimulus is presented too close to or even at or below the masked threshold the outcome of the experiment will be skewed by the lack of audibility (e.g. Kopčo *et al.*, 2010).

In auditory localization paradigms, the target is typically presented at different locations around the listener. Additionally, many realistic scenarios will include reverberation to varying degrees and the background noise will often not be ideally diffuse as it is frequently realized in the laboratory. This means that the masked threshold, the most obvious metric of audibility, will not only depend on the type and level of both the background noise and the target but also on the target's location with respect to the listener.

Similar to auditory localization, masking related phenomena have been studied using a wide range of rather artificial stimuli and environments (see Moore (2003) for an overview). Even though Zurek *et al.* (2004) studied the effect of the amount of reverberation on auditory detection, they used a rather artificial pulsed band-filtered noise target and an ongoing broadband noise masker. Hence, to the best knowledge of the authors, there are no comprehensive studies that analyze the effect of the target location on the masked threshold in a realistic, noisy environment. Therefore, in the present study, we measured masked thresholds for a speech target at different locations in a simulated room. The masker was a mixture of multiple concurrent dialogues spread across the room. The signals were presented using a loudspeaker-based VSE.

Functional models are often established to analyze the auditory processes that are relevant in a given psychoacoustic experiment. Considering auditory detection in realistic everyday scenarios, such a model has to be able to cope with complex binaural stimuli such as speech in a reverberant multi-talker background. In addition to the analysis, a model that can successfully predict masked thresholds in such conditions would have the potential to control target audibility for localization experiments in noise, and thereby replace tedious masked threshold measurements that would otherwise be required to ensure target audibility.

Auditory modeling usually involves temporal integration of the signals. However, depending on the aspect of hearing that is being modeled, different integration constants have been applied in the past. In models predicting the results of gap detection experiments, time constants between $1 - 30\,\text{ms}$ have been applied (e.g. Shailer and Moore,

1987; Florentine *et al.*, 1999), while models predicting loudness perception or auditory detection integrate over the entire target signal duration resulting in time constants of up to $100 - 300$ ms (e.g. Zwislocki, 1969; Buus *et al.*, 1997). In an attempt to solve this contradiction, Viemeister and Wakefield (1991) proposed a model that implements time integration over multiple "looks" of short duration (≈ 5 ms).

Excitation pattern models of auditory intensity discrimination originally assumed that performance is determined by a single dominant critical band (Zwicker, 1956). However, a number of studies have shown that the auditory system integrates information over multiple critical bands in loudness perception (e.g. Fastl and Zwicker, 2007; Scharf, 1961) as well as in signal detection (e.g.,Buus *et al.*, 1986; Hant *et al.*, 1997). Florentine and Buus (1981) implemented a multi-band version of an excitation pattern model that takes into account information in all critical bands. This approach resulted in better predictions of intensity discrimination of pulsed tones than the original single-band approach.

Hant and Alwan (2003) combined the multi-look method with the multi-band approach in their model such that an optimal decision is formed based on information in multiple looks across time and frequency. They showed that their model was not only able to predict thresholds for tone complexes and sweeps of different bandwidth and duration in noise, but also of plosive bursts in speech shaped noise as well as masked discrimination thresholds of CV syllables.

In the present study we extended the multi-look, multi-band model of Hant and Alwan (2003) by a stage that combines information across both ears. This allowed the application of binaural stimuli such as those experienced by the listener in a loudspeaker-based VSE. This extended model was then used to predict the data from the speech in speech detection experiment conducted in the first part of this study as described above. Furthermore, the model was used to analyze the acoustic and auditory cues that were mainly affecting the observed detection task. It should be noted that, in contrast to other auditory models (e.g. Dau *et al.*, 1996; Breebaart *et al.*, 2001; Jepsen *et al.*, 2008), the presented model has a simple, functional structure and does not consider aspects such as non-linear processing in the cochlear, modulation filtering or more detailed binaural processing. However, it was chosen because it seems to be the simplest existing auditory detection model that rather intuitively combines temporal and spectral integration mechanisms with a signal detection stage. Hence, the model was used as a first approximation to the effective auditory processing involved in the considered auditory signal detection experiment and, in the future, may be applied to aid the design of spatial hearing tests.

2.2 Methods

2.2.1 Experiment

2.2.1.1 Subjects

Nine subjects participated in the study, 5 female and 4 male. Their age ranged from 15 to 39 years with an average of 29.3 years. All subjects had hearing thresholds of less than 20 dB HL in octave bands up to 8 kHz. Subjects were either colleagues at the National Acoustic Laboratories (NAL) or students at Macquarie University. Ethics approval was given by the Australian Hearing Human Research Ethics Committee and all subjects had given written consent prior to participating in the study.

2.2.1.2 Setup

The study was conducted in an anechoic chamber located at the National Acoustic Laboratories (NAL) in Chatswood, NSW, Australia. This anechoic chamber features an array of 41 Tannoy V loudspeakers (see Fig. 2.1). The loudspeakers are arranged in horizontal rings of different elevation covering the surface of a sphere with a radius of 1.8 m. A chair was placed in the center of the array and adjusted for every subject such that the subject's head was located exactly at the center of the sphere.

The stimuli were played back using a PC with an RME MADI soundcard. The output of the soundcard was connected to two RME M-32 D/A converters. The analog signals were then amplified by eleven 4-channel Yamaha XM4180 amplifiers.

2.2.1.3 Stimuli

The stimuli were generated using room simulation software (*Odeon*) and the LoRA toolbox for MATLAB (Favrot and Buchholz, 2010). *Odeon* (Rindel, 2000) was used to simulate a room and generate the corresponding room impulse responses (RIRs) between a set of source positions and a listener position in that room. The LoRA toolbox was then used to generate the loudspeaker signals by combining the generated RIRs with anechoic speech recordings.

A noisy, reverberant cafeteria was chosen as a good representation of a challenging everyday situation. This cafeteria was modeled as a big shoebox room (15 m × 8.5 m × 2.8 m) with reflective walls, floor and ceiling, and containing twelve tables with six chairs each. The resulting reverberation time of the modeled room was $T_{30} \approx 0.5$ s. To avoid

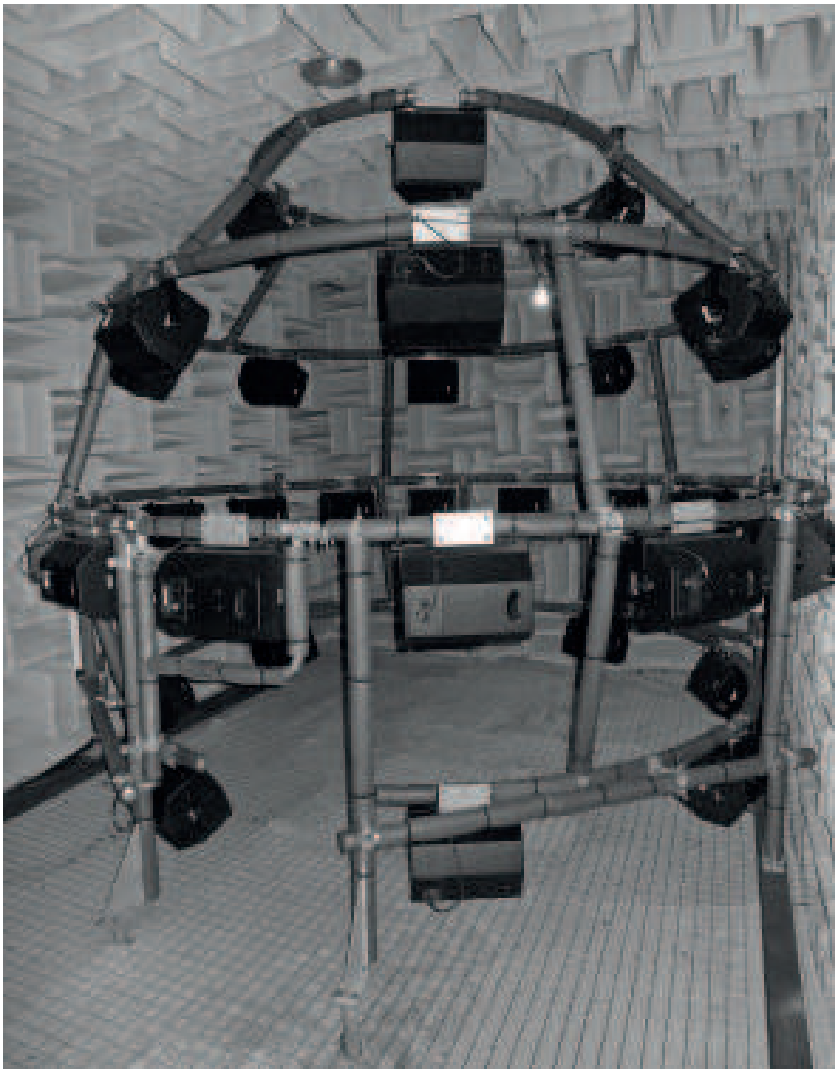


FIGURE 2.1: Photograph of the 41-channel loudspeaker array installed in an anechoic chamber at NAL

unwanted symmetry effects, the position of the listener within the room was chosen to be slightly off center on both axes. Target positions were selected equally distributed on circles of 1, 2 and 4 m around the listener. For the masker stimulus 14 positions at different tables across the rooms were chosen, forming seven pairs of talkers. A drawing of the top-down perspective of the cafeteria scene is shown in Fig. 2.2, with talkers contributing to the masker indicated by gray heads and target locations indicated by black heads.

The resulting RIRs were auralized using the LoRA toolbox for MATLAB. This toolbox divides a RIR into three different parts containing the direct sound, the specular early reflections (here described by image sources up to order 3) and the diffuse portion of the sound, i.e. the scattered energy of the early reflections as well as late reverberation. The direct sound and specular early reflections were realized according to the nearest

loudspeaker approach. In this approach the sound is assigned to the loudspeaker whose position with respect to the test subject is the most similar to the position of the sound source with respect to the listener in the simulated room. The reverberant part of the RIR was realized by multiplying directional (first order) energy envelopes with noise that is uncorrelated across loudspeakers. Details can be found in Favrot and Buchholz (2010). Finally, 41 filters are obtained for every pair of target and listener positions, mapping the respective RIR to each of the 41 loudspeakers in the array.

An anechoic recording of a female speaker saying the word "two" was chosen as the target. This number was chosen for its brief duration, broadband content, clear onset and to allow comparison with previous studies (e.g. Best *et al.*, 2011). The masker was composed of anechoic recordings of seven scripted conversations between two persons (8 female and 6 male talkers in total). The anechoic target and the masker signals were then combined with the RIRs corresponding to their positions in the simulated room and auralized as described above. This way a reverberant multi-talker masker signal was generated with a duration of about 3 minutes.

For calibration purposes the masker signal was played back by the loudspeaker array. Using a sound level meter, the average sound level of the masker at the listening position in the center of the array was adjusted to a root mean square (RMS) value of 65 dB SPL. According to predictions by *Odeon* this was the expected total sound pressure level of the described scene when each talker contributing to the masker applied a raised voice level. The time-averaged signal power of the calibrated masker was then calculated in

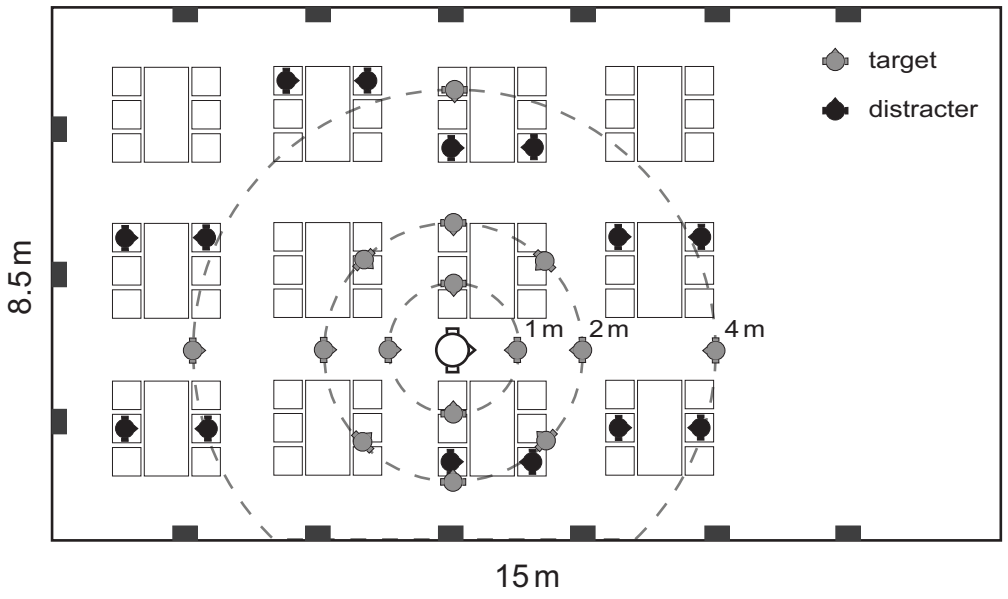


FIGURE 2.2: Top-down perspective of the cafeteria scene that was simulated. The transparent head represents the listener, gray heads indicate the target positions in 1, 2 and 4 m distance and black heads indicate the positions of sources contributing to the masker.

MATLAB and the target level adjusted accordingly for each target stimulus to obtain the desired signal-to-noise ratios (SNRs).

2.2.1.4 Procedure

In the beginning of the experiment subjects were asked to sit down on the chair in the center of the loudspeaker array. The height of the chair was adjusted such that the subject's head was in the center of the loudspeaker array. Subjects were asked to face the front loudspeaker at all times during the experiment but were not physically restrained.

Masked thresholds were measured for the 15 target locations shown in Fig. 2.2 using an adaptive 1-up 2-down staircase procedure, thus measuring the 70.7 %-correct point on the psychometric function (Levitt, 1971). Each track was started at a SNR of 5 dB and the initial step size was 8 dB. The step size was decreased to 4 and 2 dB respectively at the first two reversals in the adaptive track, and after twelve reversals the track was terminated. The masked threshold was then determined as the mean value of the SNR at the last eight reversal points. Each trial in the adaptive track was designed as a three-alternative forced-choice (3-AFC) task.

The progress of the experiment was controlled by the subjects using a graphical user interface (GUI) presented on an iPad. The GUI included a start button and one button for each interval in the 3-AFC task. After starting an adaptive track, the masker stimulus was played in a loop. In each trial, three intervals were then played with only one of the three intervals containing the target in addition to the continuous masker. The active interval was indicated to the subjects by the corresponding button on the GUI being colored. After all three intervals had been played the subjects could choose the interval in which they had detected the target stimulus by pressing the corresponding button.

2.2.2 Auditory detection model

Predicting masked thresholds of a speech target in a multi-talker background requires an auditory detection model to be able to deal with non-stationary wide-band signals. Hant and Alwan (2003) presented a modeling approach that was designed to predict the detection of wide-band, non-stationary signals (such as speech) in complex background noise. The model combines information over time and frequency, and was optimized to predict noise-masked thresholds for bandpass noise of different center frequencies, bandwidths and durations. The derived model was then shown to reliably predict thresholds

of plosive bursts in speech shaped noise as well as masked discrimination thresholds of CV syllables.

The model presented here closely follows that of Hant and Alwan (2003), however, their original approach was extended by a binaural integration stage. The model is divided into two parts, the auditory front end and the decision device. The auditory front end is a signal processor that generates an internal representation of the stimuli. Based on this internal representation the decision device computes and combines d' information across time, frequency and both ears. The desired masked threshold is defined as the SNR that yields an overall d' matching a theoretical value corresponding to the psychoacoustical task that had to be completed. Although not shown here, the implementation of the original part of the model was successfully verified by reproducing all the model predictions shown in Hant and Alwan (2003).

2.2.2.1 Auditory front end

The processing steps of the auditory front end are shown in Fig. 2.3. First the stimuli arriving at the listener's left and right ear are separately processed by an auditory filterbank. The resulting narrowband signals are half-wave rectified by squaring. After that, the signals are integrated over time, logarithmically compressed, and finally internal noise is added.

The auditory filterbank contains 31 4-th order gammatone filters (Patterson *et al.*, 1982) with center frequencies being separated by 1.019 equivalent rectangular bandwidths (ERBs). The ERB depends on the center frequency c_f (in kHz) and is given by

$$\mathrm{ERB} = 24.7 \cdot (4.37 \cdot c_f + 1). \tag{2.1}$$

Center frequencies range from 105 Hz to 7411 Hz.

The narrowband signals are processed by a temporal integrator using a flat window with raised-cosine skirts. The flat part of the window is 4 ms long and the skirts have a duration of 1 ms each. The window is applied every 5 ms resulting in an overlap of 1 ms between temporally adjacent windows. Due to the use of raised-cosine skirts and the overlap of 1 ms the stimulus is weighted equally across its duration.

After temporal integration, the signal is logarithmically compressed before internal noise is added. The internal noise is modeled as Gaussian noise with zero mean and frequency-dependent variance $\sigma^2_{int}(k)$, where k is the ERB number (Moore and Glasberg, 1983).

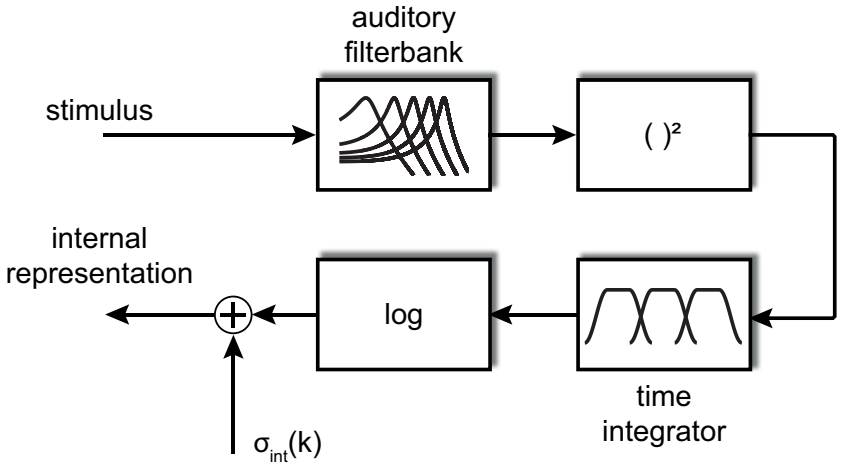


FIGURE 2.3: Schematic of the auditory front end of the detection model, applied to the signals arriving at the right and left ear. First the signal is processed by an auditory filterbank. Then, in each frequency channel the signal is squared, integrated over a short time window, and logarithmically compressed. Finally, Gaussian noise is added to each channel, representing the internal noise of the auditory system.

Hant and Alwan (2003) suggested values of the variance given by the sigmoidal fit

$$\sigma_{int}^2(k) = 16.62 + 7.88 \cdot \left(-\frac{1}{2} + \frac{1}{2} \cdot \frac{1 - e^{k-21.8}}{1 + e^{k-21.8}} \right). \tag{2.2}$$

2.2.2.2 Decision device

In the last stage of the model a decision device provides an estimate of the sensitivity d' of the detector to the target based on the time-frequency representation given by the auditory front end. Similarly to the multiple-look concept proposed by Viemeister and Wakefield (1991), a d' value is calculated for each time-frequency look independently. The overall d' is then determined by an optimal combination of individual d' values across time, frequency and both ears. The masked threshold can be found by iteratively adjusting the SNR of the input signals to the ears until the combined d' resulting from the decision device matches the theoretical d' of the corresponding detection task.

The decision device compares the internal response to the masker alone (denoted $\mathbf{M}$) with the internal response to a mix of the masker and the target (denoted $\mathbf{S}$). Note that $\mathbf{M}$ and $\mathbf{S}$ refer to the internal representation of the stimuli, i.e. the output of the auditory front end at either the left or right ear. The sensitivity of the detector in a single time-frequency look is then given by

$$d'_{nm} = \frac{|\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}|}{\sqrt{\left(\sigma^2_{\mathbf{S}_{nm}} + \sigma^2_{\mathbf{M}_{nm}}\right)/2}}, \tag{2.3}$$

where n and m are the time index and the frequency index and μ and σ^2 denote the mean and variance respectively. The mean and variance values for each frequency look are determined from 2000 samples of the internal responses **S** and **M**. Since the masker stimulus has a much longer duration than the target stimulus, **S** is obtained by adding the target stimulus to a different, randomly chosen part of the masker for each sample.

The optimal combination of sensitivity values across time and frequency is then given by

$$d' = \sqrt{\sum_{n=1}^{N_t} \sum_{m=1}^{N_f} (d'_{nm})^2}, \tag{2.4}$$

with N_t and N_f being the number of time looks and frequency looks, respectively.

To account for decreased detector efficiency when dealing with wide-band and/or long-duration stimuli, and thereby limiting the maximal temporal and spectral integration window, Hant and Alwan (2003) introduced a weighting function w. This weighting function applies zero weighting to time-frequency looks where the difference $|\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}|$ is below a threshold θ. This extends Eq. (2.4) to

$$d' = \sqrt{\sum_{n=1}^{N_t} \sum_{m=1}^{N_f} w_{nm} \, (d'_{nm})^2}, \tag{2.5}$$

with

$$w_{nm} = \begin{cases} 1 & \text{if} \quad |\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}| \geq \theta, \\ 0 & \text{if} \quad |\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}| < \theta. \end{cases} \tag{2.6}$$

Similar to the variance of the internal noise, the threshold θ is allowed to vary with center frequency. The threshold values suggested by Hant and Alwan (2003) are given by

$$\theta(k) = 3.81 + 2.39 \cdot \left(-\frac{1}{2} + \frac{1}{2} \cdot \frac{1 - e^{\frac{k-16.15}{14}}}{1 + e^{\frac{k-16.15}{14}}} \right). \tag{2.7}$$

Equation (2.5) provides a method to determine the sensitivity of the detector to a target stimulus in the presence of a masker. However, Eq. (2.5) can only be used if the stimuli are monaural. We extended the model by a stage that also combines information across ears. In this approach, d' values are combined centrally and the overall sensitivity is given by

$$d' = \sqrt{\sum_{n=1}^{N_t} \sum_{m=1}^{N_f} \sum_{i=\{L,R\}} w_{nm,i} \left(d'_{nm,i} \right)^2}, \tag{2.8}$$

where indices L and R refer to the left and right ear, respectively, and the weighting function is also evaluated at each ear individually. A flow chart of this extended decision

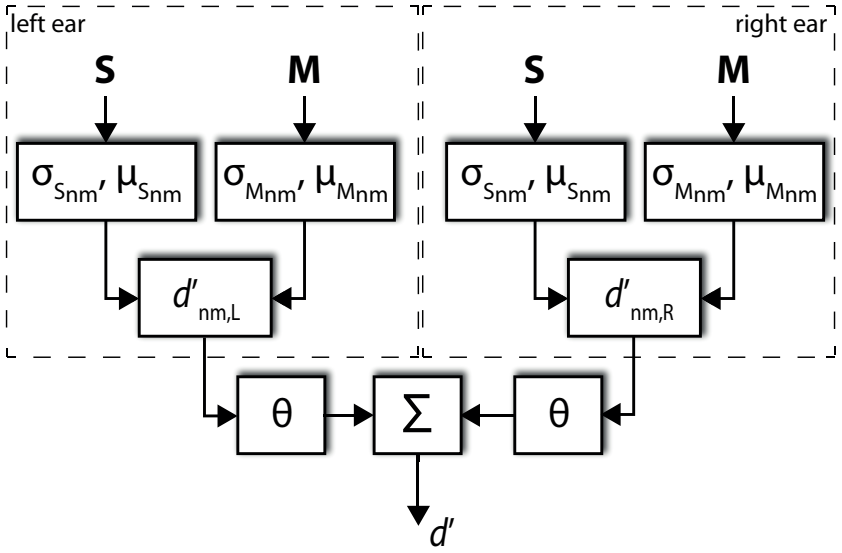


FIGURE 2.4: Flow chart of the extended decision device described by Eq. (2.8). The detectability values d'_{nm} are determined in each ear separately, then weighted and finally integrated over time, frequency and both ears.

device is depicted in Fig. 2.4.

2.2.2.3 Input signal generation and model implementation

The input stimuli for the detection model were generated such that they would resemble the signals at the eardrum of the subjects in the experiment as closely as possible. The 41 loudspeaker signals used in the experiment were filtered with head-related transfer functions (HRTFs) measured from each loudspeaker to the two ears of a Head And Torso Simulator (HATS). The filtered signals were then summed up for each ear separately to obtain appropriate input stimuli for the model. A diagram of this process is depicted in Fig. 2.5.

The sensitivity of a human observer performing a 3-AFC task is $d' = 1.26$ at threshold as measured using a 1-up 2-down adaptive track (Hacker and Ratcliff, 1979). Consequently, this value was set as the sensitivity goal for the model. The values for the two model parameters, namely the variance of the internal noise $\sigma^2_{int}(k)$ and the threshold $\theta(k)$, were left unchanged from the values suggested by Hant and Alwan (see Eqs. (2.2) and (2.7)).

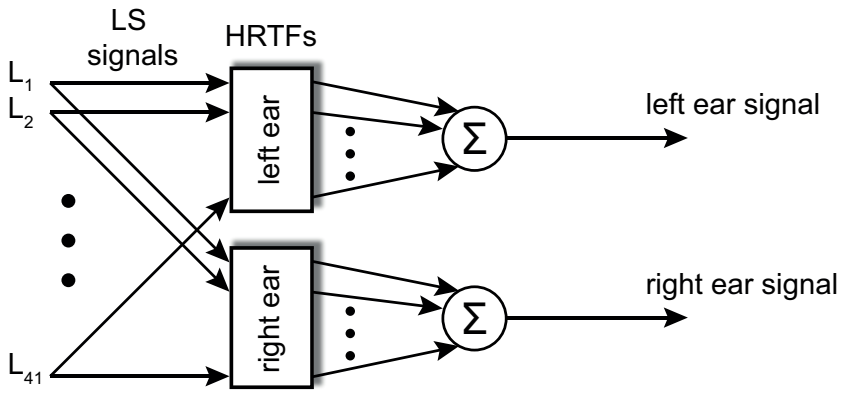


FIGURE 2.5: The input stimuli for the model are generated by first filtering the loudspeakers signals used in the behavioral experiments with HRTFs. The HRTFs were measured from each loudspeaker position to each ear. The filtered signals are then summed up for both ears individually, resulting in the input stimulus.

2.3 Results

2.3.1 Experiment

In the experiment, masked thresholds were measured adaptively for 15 different target positions (see Fig.2.2). The results are shown in Fig. 2.6. Masked thresholds are plotted against the direction of the target stimulus (with 0° being the frontal direction and positive angles referring to the right-handed side) for different target stimulus distances. Over all distances and directions, masked thresholds vary between −13.2 dB SNR (90°, 1 m) and −1.5 dB SNR (180°, 2 m), covering a range of 11.7 dB. In general the variation

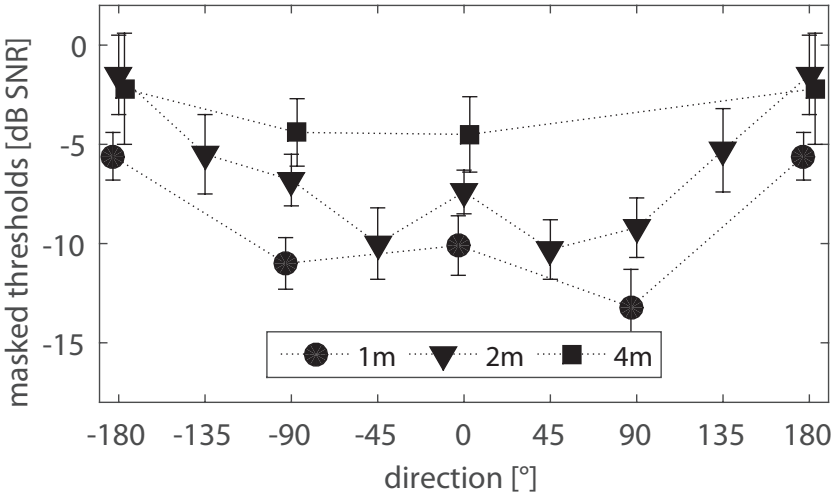


FIGURE 2.6: Masked thresholds for different target source distances as a function of target source direction, averaged over subjects. Negative angles refer to the left hand side of the listener and positive angles refer to the right hand side of the listener. Error bars denote ±1 standard deviation of the mean.

target distance	1 m	2 m	4 m
Δ_{MT} [dB]	7.6	8.8	2.3

TABLE 2.1: The range of masked thresholds Δ_{MT} for different target source distances.

of the thresholds as a function of azimuth decreased with increasing target distance (curves flattened) and shifted towards higher values (poorer performance). Table 2.1 shows the range of masked thresholds (Δ_{MT}) for different target distances, indicating a clearly reduced variation with direction for the targets in 4 m distance.

For targets at both 1 m and 2 m distance masked thresholds had a minimum at lateral positions (90° and 45° respectively) and were slightly higher for targets coming from the front. For all distances, masked thresholds showed a maximum for targets coming from the back (180°). The measurement points for targets at 1 m and 2 m distance showed high symmetry in the sagittal plane. The inter-subject standard deviation was higher for targets in 4 m distance (1.7 dB − 2.8 dB) than for targets in 1 m and 2 m distance (1.1 dB − 2.1 dB).

2.3.2 Model predictions

The model predictions are shown in Fig. 2.7 together with the corresponding experimental data replotted from Fig. 2.6. The top, middle and bottom panels show results for targets in 1 m, 2 m and 4 m distance, respectively. For most of the target locations, model predictions are within $\pm\sigma$ (indicated by the error bars) of the measured masked thresholds. At closer distances the model slightly overestimates masked thresholds for targets in front of the listener (0°) and tends to slightly underestimate masked thresholds for targets in the back of the listener as well as for lateral target locations.

Table 2.2 shows the RMS error of the model predictions with respect to the measured thresholds. The error of the model predictions decreases with increasing target distance from 1.76 dB at 1 m distance to 1.07 dB at 4 m distance. The RMS error of the model predictions over all target positions is 1.39 dB.

target distance	1 m	2 m	4 m	all pos.
e_{RMS} [dB]	1.76	1.28	1.07	1.39

TABLE 2.2: RMS error between the model predictions and the experimental data. All values are in [dB].

2.4 Discussion

Both the experimental results and the predictions of the proposed auditory model showed considerable variation of the masked thresholds over target distance as well as as target direction. The variation over direction was greatest for small target distances and the increase of the masked thresholds with increasing target distance was greatest at frontal and lateral target locations. At 1 m and 2 m target distance the masked thresholds

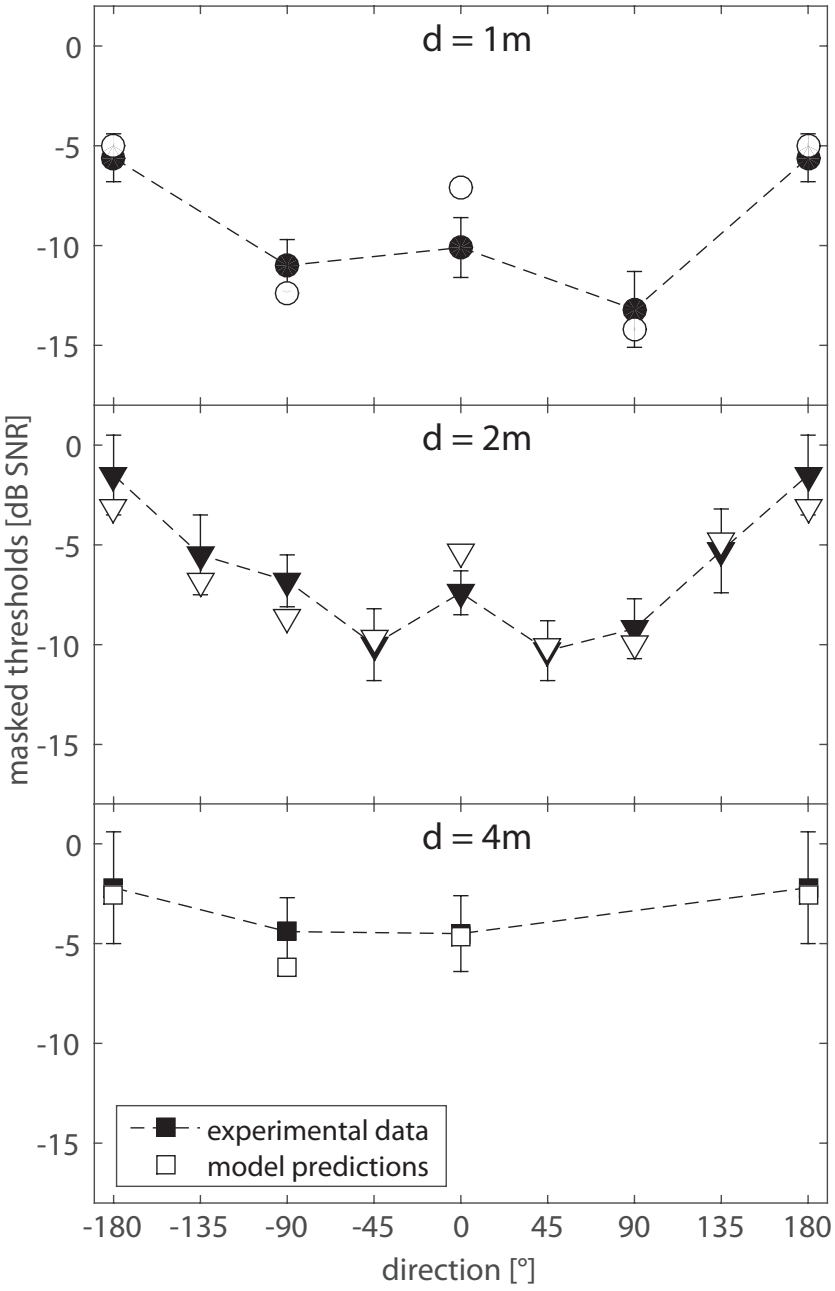


FIGURE 2.7: Model predictions (open symbols) as a function of target source direction plotted together with the corresponding experimental results from Fig. 2.6 (closed symbols). From top to bottom: target distance $d = 1, 2, 4$ m.

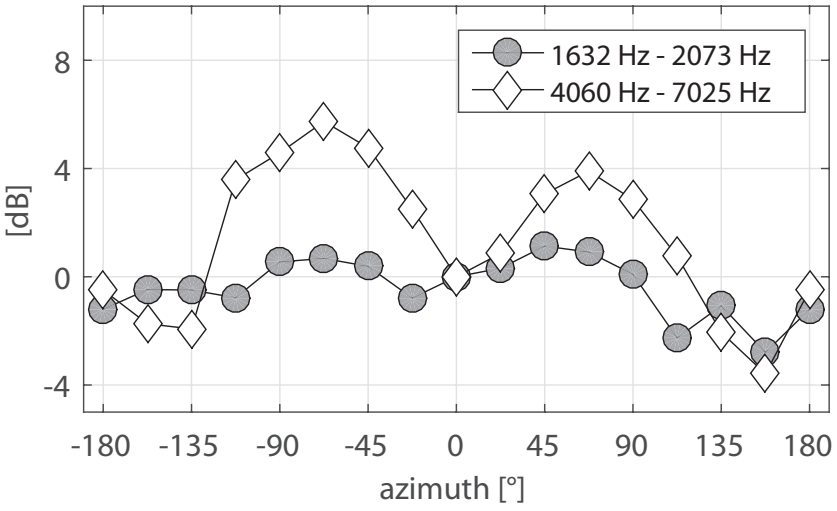


FIGURE 2.8: HRTFs integrated over both ears between 1.6 kHz and 2.1 kHz (gray circles) as well as 4 kHz and 7 kHz (white diamonds) plotted over azimuth angle. The results are normalized with respect to their 0° point.

varied by 7.6 dB SNR and 8.8 dB SNR over all directions, while at 4 m this variation only amounted to 2.3 dB SNR. Likewise, at −90° the thresholds increased by 6.7 dB SNR going from 1 m to 4 m, while at 180° they only increased by about 3.4 dB SNR. Figure 2.7 shows that this behavior of the experimental results was very well reflected in the model predictions.

In order to test if the direction dependency for targets close to the listener can simply be explained by the directivity of the head-related transfer functions (HRTFs), the HRTFs integrated (in power) over both ears and in two example frequency regions are plotted in Fig. 2.8 as functions of azimuth angle. While the HRTFs vary only by 4 dB in the

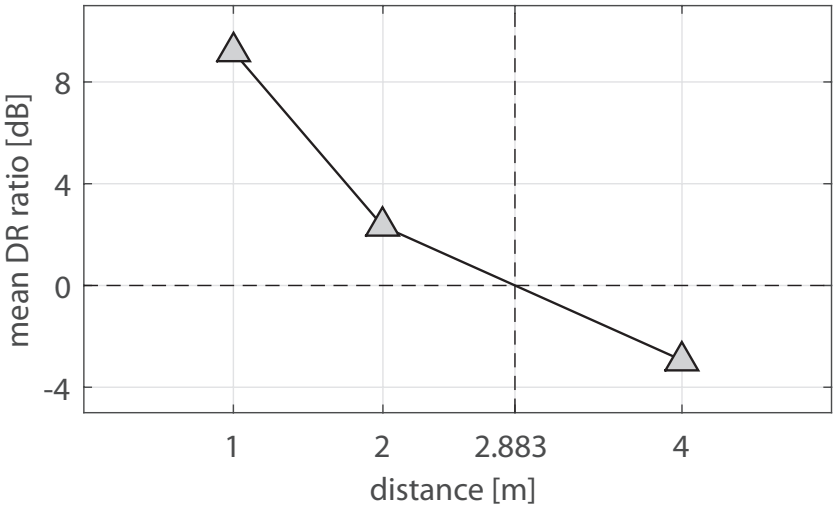


FIGURE 2.9: Direct-to-reverberant ratio averaged over all target locations at one specific distance plotted over target distance. Linear interpolation locates the critical distance at $d_c \approx 2.88$ m.

frequency region between 1.6 kHz and 2.1 kHz (corresponding to frequency channels 18 and 19 in the filterbank of the auditory front end of the model) the HRTFs between 4 kHz and 7 kHz (corresponding to frequency channels $25-30$) vary in a range of 10 dB. When comparing this behavior of the HRTFs with the masked thresholds shown in Fig. 2.7, it can be seen that the maxima of the HRTFs (between $\pm 45°$ and $\pm 112.5°$) correlate very well with the minima of the masked threshold functions for 1 m and 2 m target distance. This suggests that high frequency components play an important role in the detection of the close targets. This observation is further supported by the auditory system-internal sensitivity plots given in Fig. 2.10 as further discussed below.

With respect to Fig. 2.7, moving from 2 m to 4 m distance from the listener greatly reduces the influence of target direction on the masked thresholds and also decreases overall sensitivity. Moving further away from a sound source in a room with reflective walls usually increases the amount of reverberant energy compared to the energy of the direct sound. Figure 2.9 depicts the direct-to-reverberant (DR) ratio of sources in the simulated room as a function of the distance between the listener and the sound sources averaged over all target directions. It shows that the DR ratio drops from 9.19 dB at 1 m to -2.94 dB at 4 m. The critical distance, where direct energy equals reverberant energy, is at $d_c \approx 2.88$ m. Hence, whereas in the 1 m and 2 m target conditions the direct sound energy is mainly determining audibility, in the 4 m target conditions the reverberant energy is dominant and due to its spatially diffuse characteristic the directivity of the HRTF has less influence on the target signal. A similar effect of reverberation was observed by Plomp (1976) when measuring the intelligibility of a speech target in the front of the listener with a speech distractor at different azimuth angles. Not only were the measured intelligibility thresholds generally higher when the reverberation time was increased, but the variation of the thresholds over direction decreased significantly in more reverberant environments.

Figure 2.10 provides further analysis of the effect of target distance and direction on the masked thresholds using the described auditory model. It shows time-frequency plots of the individual sensitivity values d'_{nm} after squaring, application of the weighting and summation over both ears in Eq. (2.8). The top row contains the figures corresponding to targets located in the front of the listener, the bottom row contains the figures corresponding to target locations in the back of the listener. Figures for target locations at 1 m distance are in the left column, figures for target locations at 4 m are in the right column. For all target locations, the SNR was set to the respective predicted threshold such that the overall sensitivity was $d' = 1.26$. The corresponding SNRs are indicated in the figure panels.

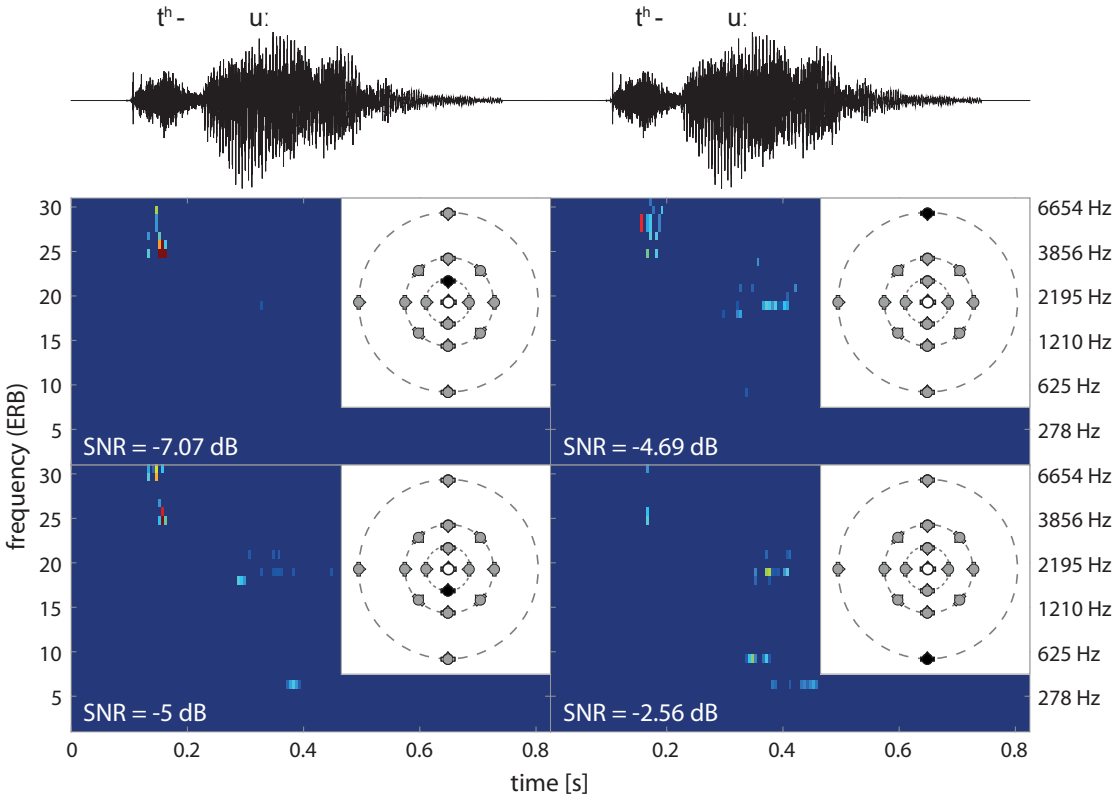


FIGURE 2.10: Time-frequency plots of the individual sensitivity values after summation over both ears. Each panel contains the plot for a different target location, each at their respective masked threshold: the top row shows values for the target in the front, the bottom row shows values for the target in the back. The left column shows values for target in 1 m distance, the right column shows values for the target in 4 m distance. Target positions are are also shown in the insets. The sensitivity values are given by the sum in Eq. (2.8) before the summation over time and frequency. Low values are shown in blue, high values are shown in red. For reference, the signal waveform of the target stimulus ("two") is plotted above.

The top left panel of Fig. 2.10 (target source at 0°, 1 m) illustrates that for a close target in front of the listener almost the entire energy contributing to its detectability is contained in the "echo-free" onset of the target stimulus, i.e. the voiceless consonant [t^h] of the word "two" [$t^h u:$]. Most of the time-frequency looks containing detectable target energy are located in frequency bands 25 to 30 ($4-7$ kHz). Hence, the frequency region where the HRTF is highly directional (see Fig. 2.8) mainly determines the detectability of the target, explaining the directional dependency of the masked threshold at short distances. For the target well beyond the critical distance (Fig. 2.10, top right panel), detectability is more strongly affected by reverberant energy in the harmonic part of the

frequency [Hz]	63	125	250	500	1000	2000	4000	8000
T_{30} [s]	0.54	0.56	0.53	0.44	0.43	0.42	0.40	0.34

TABLE 2.3: Reverberation time T_{30} of the simulated room at different frequencies

stimulus, located in frequency bands 18 and 19 (1.6−2.1 kHz). This is due to the acoustic properties of the room, which provides less reverberant energy at high frequencies than at low frequencies as indicated by the frequency dependency of the reverberation time T_{30} given in Table 2.3. As was shown in Fig. 2.8, the HRTFs vary much less over direction for lower frequencies, suggesting that not only the more diffuse characteristic of the far target signals account for the diminishing effect of target direction, but also the shift of the main auditory sensitivity to lower frequencies.

The bottom left panel of Fig. 2.10 (target source at 180°, 1 m) shows that even at short distances, the amount of detectable target energy at high frequencies is decreased when the target source is located behind the listener. Moving the source farther away from the listener further shifts detectable energy towards lower frequencies (see Fig. 2.10, bottom right panel). However, increasing the distance of sources in the back of the listener does not increase the masked threshold as much as for sources in front of the listener. This means that only a slightly increased SNR is necessary to compensate for the detrimental effect of increased reverberation when the detectable energy is already located in lower frequency regions.

Overall, the main tendencies in the experimental results are captured in the model predictions. The RMS error between the model predictions and the experimental data is on average 1.39 dB. This is especially remarkable since no changes were made to the original parameter fit suggested by Hant and Alwan (2003) and since the model does not take into account interaural cues like interaural time differences (ITDs) or interaural correlation, even though these cues are thought to be crucial in binaural hearing (as shown e.g. in experiments modeling speech intelligibility or spatial release from masking (Bronkhorst, 2000)). Culling (2007) found that, at low SNRs, changes in interaural correlation fully accounted for the listeners' ability to detect a narrowband noise target in a broadband noise masker, while at higher SNRs there seemed to be a superaddition of monarual and binaural information. He interpreted this behavior as further evidence for an equalization-cancellation (EQ) mechanism active in binaural detection. Similarly, Beutelmann and Brand (2006) incorporated an EQ mechanism when modeling speech intelligibility in a reverberant environment with spatial noise and were able to predict experimental data with a very high accuracy. Despite these contradictory findings, the purely energy-based modeling approach proposed in the current study produced very accurate predictions of the experimental data. This may, at least partly, be explained by the highly diffuse nature of the masker which distinguishes this paradigm from others that use single-loudspeaker maskers.

Therefore, despite the good agreement of the model predictions with the experimental data, no statement about the physiological plausibility of the chosen model can be made,

and future extensions to the model should in particular consider more auditory system-motivated binaural processing based on interaural cross-correlation analysis (Colburn, 1996) or equalization-cancellation theory (Durlach, 1963; Breebaart *et al.*, 2001).

2.5 Conclusions

In the chosen environment, a reverberant cafeteria with a multi-talker background noise, masked thresholds of a speech stimulus varied in a range of 11.7 dB depending on the location of the target source with respect to the listener. These results clearly show that target audibility in localization tasks or other tasks that assess spatial hearing in similar environments has to be closely controlled. However, it has to be further investigated if this holds true for different environments.

An auditory detection model proposed by Hant and Alwan (2003) was implemented and successfully extended to allow the application of binaural stimuli as experienced in loudspeaker-based VSEs. Applying this model to analyze the stimuli used in the masking experiments suggests that the masked thresholds are mainly influenced by the directivity introduced by the head shadow (as represented by HRTFs) and the direct-to-reverberant (DR) ratio of the target. These effects of target location on the masked thresholds were correctly predicted by the model. This suggests that the model is a promising candidate for controlling target audibility in future localization experiments in realistic acoustic environments. This is even further supported by the fact that the present data was successfully predicted using the original model parameters derived by Hant and Alwan (2003). However, for the testing of spatial hearing with (aided) hearing impaired listeners the model needs to be extended to include the effect of hearing loss.

Chapter 3

The effect of sensorineural hearing loss on auditory masking of speech in reverberant multi-talker environments[1]

Several studies in the past have reported that localization performance of hearing impaired listeners can be affected by decreased target audibility. Furthermore, in complex, reverberant environments with multiple distracters, target audibility does not only depend on the signal-to-noise ratio between target and masker, but also on the location of the sound sources in the scene. Weller and Buchholz (2015) showed that masked thresholds of a speech target in a reverberant multi-talker background measured in a group of normal hearing listeners varied greatly depending on the location of the target in the scene, but that it was possible to predict these masked thresholds with good overall accuracy by applying a simple, functional auditory detection model. Here, the previous experiment was repeated with a group of hearing impaired listeners. The detection model was then adjusted to obtain predictions of the masked thresholds taking into account individual differences in hearing loss by adding a threshold simulating noise. The experimental results were in good agreement with the findings of Weller and Buchholz (2015), showing that masked thresholds of hearing impaired listeners also varied over target location, but to a lesser extent than for normal hearing listeners. The model was able to predict the average sensitivity of the hearing impaired listeners rather well, but the variation of the masked thresholds across subjects and target locations was not very well reflected in the model predictions.

[1] In preparation for submission to The Journal of the Acoustical Society of America - Express Letters.

3.1 Introduction

A number of studies have measured localization in noise, but found that performance measured in normal hearing subjects only deteriorated when target signals were presented at rather low (often negative) signal-to-noise ratios (Good and Gilkey, 1996; Abouchacra *et al.*, 1998; Hawley *et al.*, 1999; Drullman and Bronkhorst, 2000). In all these studies it was highlighted that performance may have been largely affected by limited audibility of the target signal. The effect of target audibility on localization is expected to be significantly increased in hearing impaired subjects due to elevated (pure-tone) thresholds. Lorenzi et al (1999) measured both localization and detection of click trains in a broadband noise masker and found that hearing impaired listeners with low localization accuracy also had elevated masked thresholds, confirming that the poor localization performance may have been caused by the reduced audibility of the target stimulus. In a reverberant environment, control over the target stimulus audibility is even more important when measuring localization since masked thresholds will not only depend on the nature of the stimuli but also on the location of the sound sources. However, most studies on masking have applied rather artificial signals like tones, noises, or click trains (e.g. Fastl and Zwicker, 2007; Moore, 2003), but very few studies have investigated masking in more realistic environments. Recently, Weller and Buchholz (2015) measured masked thresholds of a speech target in a reverberant multi-talker environment in a group of normal-hearing (NH) listeners and found that the thresholds depended heavily on the direction and distance between the target source and the listener.

Weller and Buchholz (2015) then applied an auditory model based on the model proposed by Hant and Alwan (2003) which was able to successfully predict the measured masked thresholds with an overall root mean square (RMS) error of 1.4 dB. An analysis of the model processes suggested that, at threshold, most of the detectable energy of the target stimulus was concentrated at high frequencies above 4 kHz. At such high frequencies head shadow effects are substantial and could explain a large part of the variation of the measured masked threshold with target direction (at least for nearby sources where the influence of reverberation was rather small). Since sensorineural hearing loss is usually strongest at high frequencies, this raises the question how masked thresholds would change for HI listeners in particular as a function of target location. Therefore, in the current study, we replicated the setup used by Weller and Buchholz (2015) to measure the masked thresholds of a speech target in a reverberant environment in a group of HI listeners with a variety of audiometric configurations.

To better understand the behavior of the measured masked thresholds as well as to, hopefully, better control the audibility of target signals when measuring localization in HI subjects in complex environments, the same masking model that was successfully applied in Weller and Buchholz (2015) to describe auditory detection in NH subjects was also applied here. To account for the reduced sensitivity due to sensorineural hearing loss we adjusted this model by introducing a threshold simulating noise that was added to the input signals of the model. The applied model follows a rather simple, functional approach and does not include any of the important non-linear auditory mechanisms as for instance described by the computational auditory signal processing and perception (CASP) model proposed by Jepsen and Dau (2011). Nevertheless, the applied model still provides an interesting first approximation for describing and analyzing auditory detection in HI subjects in complex acoustic environments.

3.2 Methods

3.2.1 Experimental methods

3.2.1.1 Subjects

Nine HI subjects participated in the study. Subjects were aged between 69 and 86 years (mean 77.4 years) and had symmetrical sensorineural hearing losses (differences of less than ±10 dB between both ears at audiometric frequencies up to 4 kHz), with the exception of one subject who had thresholds of 40 and 60 dB HL at the left and right ear respectively at a frequency of 0.25 kHz. All subjects were recruited from a database available at NAL and had no reported cognitive disability. Ethics approval was given by the Australian Hearing Human Research Ethics Committee and all subjects had given written consent prior to participating in the study. After participating, subjects received a gratuity for their participation.

3.2.1.2 Stimuli and setup

The study was conducted in an anechoic chamber located at the National Acoustic Laboratories (NAL). This anechoic chamber featured an array of 41 Tannoy V loudspeakers. The loudspeakers were arranged in horizontal rings of different elevation covering the surface of a sphere with a radius of 1.8 m. The stimuli were played back using a PC with an RME MADI soundcard. The output of the soundcard was connected to two RME M-32 D/A converters. The analog signals were then amplified by 11 4-channel Yamaha XM4180 amplifiers.

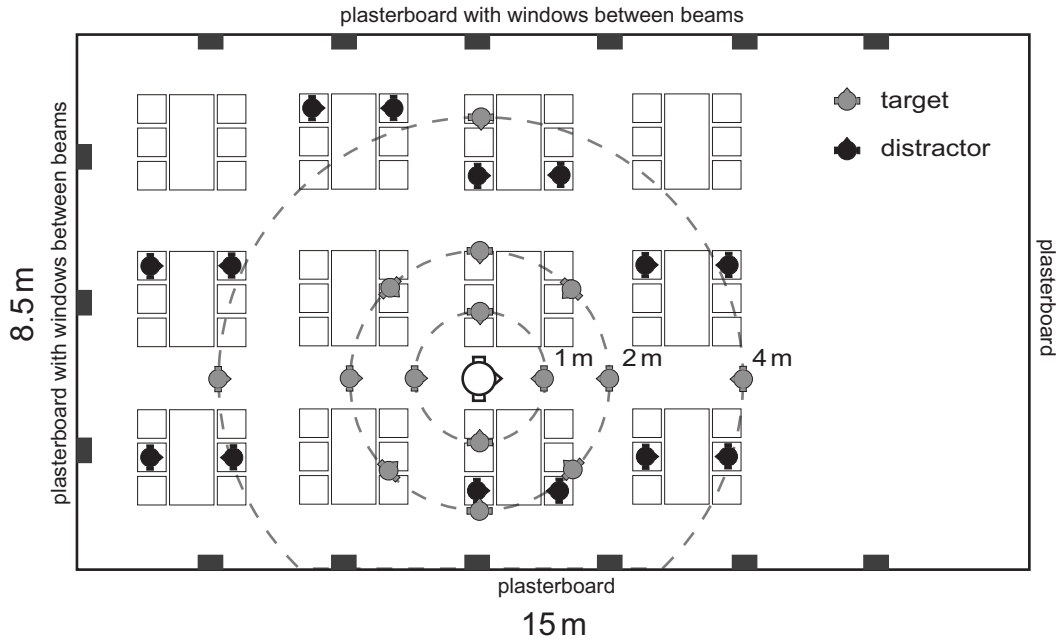


FIGURE 3.1: Top-down perspective of the cafeteria scene that was simulated. The transparent head represents the listener, gray heads indicate the target positions in 1, 2 and 4 m distance and black heads indicate the positions of sources contributing to the masker. Tables and chairs were modeled by wooden plates, the floor by 6 mm carpet with closed-cell foam on concrete, and the ceiling was suspended by 200 mm with 25 mm of wood-wool in the cavity.

The target was a single utterance of the word "two" (duration $\approx$ 350 ms) spoken by a female. The masker was a mixture of seven concurrent dialogues consisting of scripted conversations that were taken from the "listening component" of past examinations of the International English Language Testing System (IELTS) and were read by eight females and six males.

A big shoebox room (15 m$\times$8.5 m$\times$2.8 m) with reflective walls, floor, ceiling and furniture (tables and chairs) was modeled in *Odeon*, a room acoustic simulation software (Rindel, 2000). The resulting reverberation time of the modeled room was $T_{30} \approx 0.5$ s. The listener was positioned slightly off center in the simulated room to avoid undesirable sound coloration effects. Fifteen target positions were selected on circles with a radius of 1, 2 and 4 m respectively around the listener as follows: four target positions with 90° angular spacing at 1 m from the listener, eight target positions with 45° angular spacing at 2 m from the listener and three target positions at $-90°$, 0° and 180° around the listener at a distance of 4 m. For the masker stimulus 14 positions at different tables across the room were chosen, forming seven pairs of talkers. A drawing of the top-down perspective of the cafeteria scene is shown in Fig. 3.1, with talkers contributing to the masker indicated by black heads and target locations indicated by gray heads. Using *Odeon*, room impulse responses (RIRs) were generated between all target and masker talker positions and the listener position.

These RIRs were then auralized using the LoRA toolbox for MATLAB (Favrot and Buchholz, 2010). This toolbox divides a RIR into three different parts containing the direct sound, the specular early reflections (here described by image sources up to order 3) and the diffuse portion of the sound, i.e., the scattered energy of the early reflections as well as late reverberation. The direct sound and specular early reflections were each realized by the playback loudspeaker that was closest to their direction. The reverberant part of the RIR was realized by multiplying directional (first order) energy envelopes with noise that is uncorrelated across loudspeakers. Details can be found in Favrot and Buchholz (2010). Finally, 41 filters were obtained for each sound source, mapping the respective RIR to each of the 41 loudspeakers in the array. After applying the filters to the anechoic speech signals, the masker was created by adding up all seven auralized dialogues, resulting in a multi-talker scene with a duration of about three minutes. The anechoic speech signals contributing to the masker had been normalized to the same root mean square (RMS) value prior to the processing, and therefore their RMS level relative to each other at the listener's position depended on their distance from as well as their direction and orientation towards the listener.

For calibration purposes, the masker signal was played back by the loudspeaker array. Using a sound level meter, the average sound level of the masker at the listening position in the center of the array was adjusted to a root mean square (RMS) value of 65 dB SPL. According to predictions by *Odeon* this was the expected total sound pressure level of the described scene when each talker contributing to the masker applied a normal voice level. The time-averaged signal power of the calibrated masker was then calculated in MATLAB and the target level adjusted accordingly for each target stimulus to obtain the desired signal-to-noise ratios (SNRs).

3.2.1.3 Procedure

In the beginning of the experiment subjects were asked to sit down on the chair in the center of the loudspeaker array. The height of the chair was adjusted such that the subject's head was in the center of the loudspeaker array. Subjects were asked to face the front loudspeaker at all times during the experiment but were not physically restrained. All subjects were tested without their hearing aids.

Masked thresholds were measured for the 15 target locations shown in Fig. 3.1 using an adaptive 1-up 2-down staircase procedure, thus measuring the 70.7 %-correct point on the psychometric function (Levitt, 1971). Each track was started at a SNR of 5 dB and the initial step size was 8 dB. The step size was decreased to 4 dB after the first two reversals in the adaptive track and to 2 dB after the next two reversals, and after twelve

reversals the track was terminated. The masked threshold was then determined as the mean value of the SNR at the last eight reversal points. Each trial in the adaptive track was designed as a three-alternative forced-choice (3-AFC) task.

The progress of the experiment was controlled by the subjects using a graphical user interface (GUI) presented on an iPad. The GUI included a start button and one button for each interval in the 3-AFC task. After starting an adaptive track, the masker stimulus was played in a loop. In each trial, three intervals were then played with only one of the three intervals containing the target in addition to the continuous masker. The active interval was indicated to the subjects by the corresponding button on the GUI being colored. After all three intervals had been played the subjects could choose the interval in which they had detected the target stimulus by pressing the corresponding button.

3.2.2 Auditory detection model

The applied auditory detection model was originally proposed by Hant and Alwan (2003) and extended by Weller and Buchholz (2015) and consists of two parts. The first part is an auditory front-end that simulates the peripheral processing of the auditory system and generates an internal representation of the stimuli. The second part is an optimal detector that combines information across time, frequency and both ears and produces an estimate of the sensitivity d' for a given target and masker signal. The target level (SNR) is then adjusted iteratively until the predicted sensitivity matches a theoretical value corresponding to the psychophysical task at hand. For the applied 3-AFC task in a 1-up 2-down procedure the sensitivity is $d' = 1.26$ at threshold(Hacker and Ratcliff, 1979).

3.2.2.1 Auditory front-end

First the stimuli arriving at the listener's left and right ear are separately processed by an auditory filterbank. The auditory filterbank contains 31 4-th order gammatone filters (Patterson *et al.*, 1982) with center frequencies between 105 Hz and 7411 Hz and separated by 1.019 equivalent rectangular bandwidths (ERBs). The ERB depends on the center frequency c_f (in kHz) and is given by

$$\mathrm{ERB} = 24.7 \cdot (4.37 \cdot c_f + 1). \tag{3.1}$$

The narrowband signals are then squared and processed by a temporal integrator using a flat window with raised-cosine skirts. The flat part of the window is 4 ms long and

the skirts have a duration of 1 ms each. The window is applied every 5 ms resulting in an overlap of 1 ms between temporally adjacent windows. After temporal integration, the signal is logarithmically compressed before internal noise is added. The internal noise is modeled as Gaussian noise with zero mean and frequency-dependent variance $\sigma^2_{int}(k)$, where k is the ERB number (Moore and Glasberg, 1983). Hant and Alwan (2003) suggested values of the variance given by the sigmoidal fit

$$\sigma^2_{int}(k) = 16.62 + 7.88 \cdot \left(-\frac{1}{2} + \frac{1}{2} \cdot \frac{1 - e^{k-21.8}}{1 + e^{k-21.8}} \right). \tag{3.2}$$

3.2.2.2 Decision device

The decision device compares the internal response to the masker alone (denoted **M**) with the internal response to the sum of the masker and the target (denoted **S**). Note that **M** and **S** refer to the internal representation of the stimuli, i.e. the output of the auditory front end at either the left or right ear. The sensitivity of the detector in a single time-frequency look is then given by

$$d'_{nm} = \frac{|\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}|}{\sqrt{\left(\sigma^2_{\mathbf{S}_{nm}} + \sigma^2_{\mathbf{M}_{nm}}\right)/2}}, \tag{3.3}$$

where n and m are the time index and the frequency index and μ and σ^2 denote the mean and variance respectively. The mean and variance values for each frequency look are determined from 2000 samples of the internal responses **S** and **M**. The individual sensitivity values are then combined across time, frequency and both ears using an optimal detector resulting in the overall sensitivity given by

$$d' = \sqrt{\sum_{n=1}^{N_t} \sum_{m=1}^{N_f} \sum_{i=\{L,R\}} w_{nm,i} \left(d'_{nm,i} \right)^2}, \tag{3.4}$$

where N_t is the number of looks in time, N_f is the number of frequency looks and indices L and R refer to the left and right ear, respectively. $w_{nm,i}$ is a binary weighting function that applies zero weighting to time-frequency looks where the difference $|\mu_{\mathbf{S}_{nm}} - \mu_{\mathbf{M}_{nm}}|$ is below a threshold θ (for values of θ refer to Hant and Alwan, 2003).

3.2.2.3 Input signals

The input signals were generated by performing a binaural synthesis (Hammershøi and Møller, 2005) on the loudspeaker signals used in the experiment. A set of 41 head-related transfer functions (HRTFs) between the loudspeakers in the array and the ears of a

Brüel&Kjær head and torso simulator (HATS) was measured. Then, the loudspeaker signals were filtered with the measured HRTFs and summed up for each ear separately. This binaural synthesis was carried out for the masker signal and each of the 15 target signals, and the resulting input signals were calibrated such that the average level of the masker signal was 65 dB across both ears.

To account for the individual hearing loss, a threshold simulating noise signal was added to the input signals at each ear. This threshold simulating noise was adjusted for each of the HI listeners separately by filtering white noise with a filter designed using the individual pure-tone thresholds. For each listener, the monaural version of the auditory detection model was applied to 300 ms-long pure tone targets (separately at audiometric frequencies of 0.25, 0.5, 1, 2, 3, 4 and 6 kHz) and a noise masker. The spectrum of the noise masker was then adjusted iteratively until the thresholds predicted by the model matched the listener's pure tone thresholds. In the first iteration a white noise masker was used. Next, a 256-tap minimum phase finite impulse response (FIR) filter was designed based on the thresholds predicted in the first iteration and applied to the white noise masker. In each subsequent iteration (typically up to a maximum of two iterations), the frequency response of the FIR filter was adjusted until the error between the model predictions and the audiometric data was within a margin of ±2 dB.

3.3 Results

Figure 3.2 (left column) shows the subjects' mean and standard deviation of the masked thresholds together with the individual thresholds (thin gray lines) at all 15 target positions as a function of target direction. Here, 0° refers to the frontal direction, negative angles refer to the left-hand side and positive angles refer to the right-hand side. The three different panels show the masked thresholds for different target distances d: $d = 1$ m (top), $d = 2$ m (center) and $d = 4$ m (bottom). At all target distances, masked thresholds were generally lowest when the target was in front of the listener and highest when the target was behind the listener. Across direction, the mean masked thresholds covered a range of about 4 dB at distances of 1 m and 2 m, but only about 2 dB at the farthest distance of 4 m.

Overall, moving the target farther away from the listener elevated masked thresholds. However, due to the reduced variation over target direction, the mean masked thresholds of targets at 0° increased more strongly (an increase of 4.3 dB going from 1 m to 4 m) than masked thresholds of targets in the back (an increase of 2 dB going from 1 m to 4 m).

This general behavior of the masked thresholds over direction and distance was very similar to that of masked thresholds measured in nine NH listeners (Weller and Buchholz, 2015). There are, however, two important differences between the data of the two groups. First, masked thresholds of the HI listeners were overall higher than masked thresholds of NH listeners. This difference between the two groups varied slightly with target direction and distance, but on average masked thresholds in the HI group were 5.8 dB higher than those in the NH group. Secondly, in the NH group, masked thresholds of targets at 1 m and 2 m distance were lowest at lateral directions rather than at the front. This behavior was not shown by the masked thresholds of the HI group.

The right column of Fig. 3.2 compares the model predictions with the masked thresholds measured in the experiment as a function of target direction. It can be seen that the masked thresholds predicted by the model were in a similar SNR range as the measured

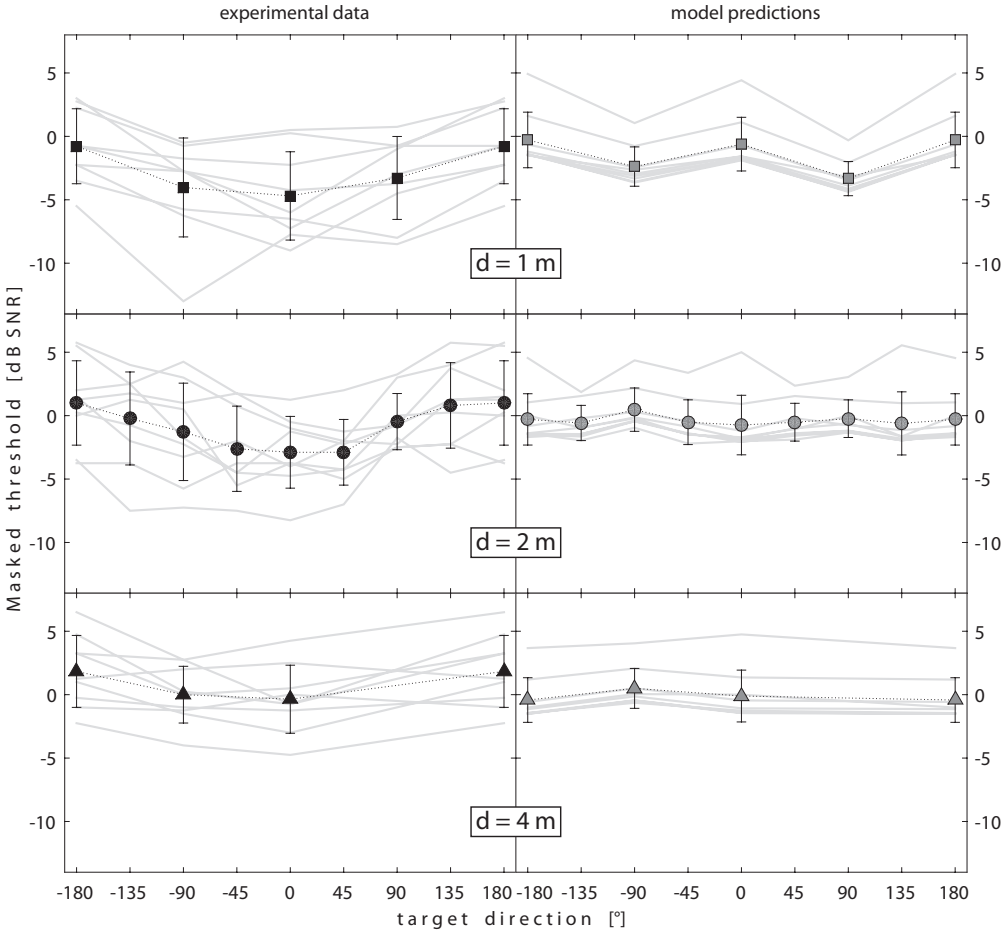


FIGURE 3.2: Masked thresholds of the target word "two" in a reverberant multi-talker background as a function of target direction, measured in a group of HI listeners (left column) and predicted by an auditory detection model (right column). Each row shows the results for a different target distance: 1 m (top row), 2 m (center row) and 4 m (bottom row). The solid black lines show the mean across subjects and the error bars represent ± one standard deviation from the mean. The gray lines correspond to individual data of the nine subjects.

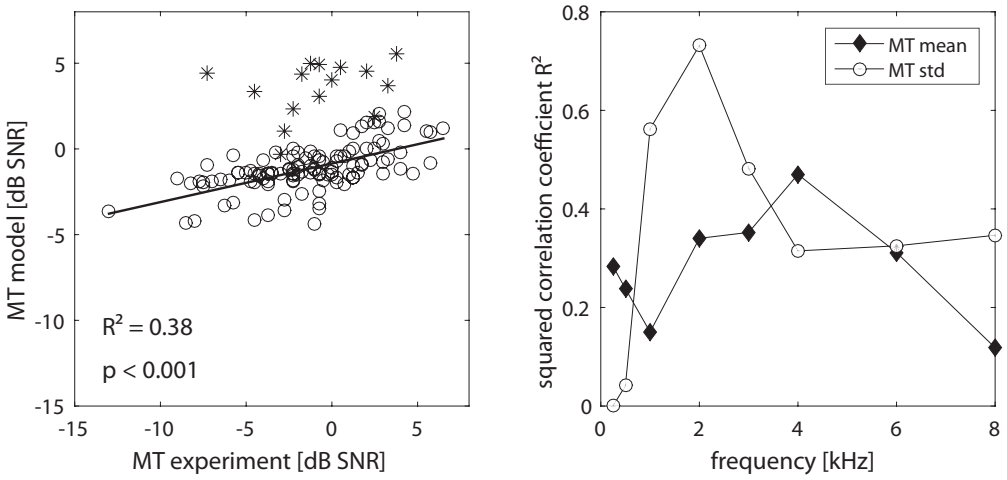


FIGURE 3.3: Analysis of the experimental results and the model predictions. The left panel shows the individual model predictions as a function of the measured masked thresholds. The right panel shows the correlation of the experimental results with hearing loss at different audiometric frequencies. The overall sensitivity (represented by the mean threshold over all target locations) was most strongly correlated with hearing loss above 2 kHz, whereas the variation of the masked thresholds over target location (represented by the standard deviation over all target locations) was mainly correlated with hearing loss between 1 kHz and 3 kHz.

masked thresholds but that they had a substantially reduced variation across subjects as well as target locations. The mean inter-subject standard deviation of the measured masked thresholds was 3.1 dB compared to 1.8 dB for the model predictions. The mean masked thresholds predicted by the model increased by 1.7 dB going from 1 m to 4 m compared to an average increase by 3.6 dB of the measured masked thresholds.

To test if the model predictions followed similar trends as the measured masked thresholds, they are plotted against each other in a scatter plot in Fig. 3.3 (left panel). It can be seen that the model predictions correlated very well with the experimental data for all subjects except for subject 4 as indicated by the asterisk. For subject 4 the model predicted highly elevated thresholds. This may be explained by the subject's hearing loss, which was basically flat at around 50 dB HL and was by far the highest of all subjects at low frequency. Hence, it may be speculated that the model overly emphasizes on the audibility at low frequencies, which is further discussed below. Applying a linear regression analysis to the data excluding subject 4 revealed a modest but significant correlation between modeled and measured masked thresholds ($R^2 = 0.38$, $p < 0.001$). However, the slope of the approximated linear function was very shallow ($a = 0.22$) highlighting the aforementioned observation that the modeled masked thresholds showed a highly reduced variation across subjects and target location.

In order to better understand what mainly drives the large variation of the masked

thresholds across subjects, a correlation analysis was carried out between the (pure-tone) hearing thresholds at different audiometric frequencies and two different measures of the individual experimental data. The mean value over all masked thresholds was used as an estimate of a subject's overall sensitivity and the standard deviation as an estimate for the overall variation of the data as a function of source location. The resulting correlation coefficients are shown in the right panel of Fig. 3.3. It can be seen that the standard deviation of the masked thresholds was highly correlated with the hearing loss at around 2 kHz ($R^2 = 0.78$). The mean value of the masked threshold was mainly correlated with the hearing loss at 4 kHz ($R^2 = 0.47$), but to a lesser degree and with a much less frequency-dependent behavior than the standard deviation.

3.4 Discussion and conclusions

In this study, masked thresholds for the detection of a speech target in a reverberant multi-talker background were measured for a cohort of HI listeners. The results showed that when compared to the masked thresholds of NH listeners measured under the same conditions, masked thresholds of the HI listeners were overall higher and varied less over both target direction and target distance. Considering that most of the participating listeners had a substantial hearing loss at high frequencies, this is in line with the observations made by Weller and Buchholz (2015). They found that most of the detectable target energy at threshold was located at high frequencies in the voiceless onset of the target word "two". This is further supported by the fact that in the current study, the listener with the lowest high-frequency hearing loss (44 dB between 2 kHz and 6 kHz) also had the lowest thresholds by a considerable margin. At a target distance of 4 m, where energy at high frequencies is less dominant due to the low direct-to-reverberant (DR) ratio, the masked thresholds of this listener were within ±0.35 dB of the average masked thresholds in the NH group measured by Weller and Buchholz (2015). In addition to reduced audibility, other consequences of cochlear damage (see Moore (2007) for a review) may have contributed to the raised thresholds in the HI group. Poorer temporal resolution reduces the ability to detect gaps in fluctuating stimuli (e.g. Fitzgibbons and Wightman, 1982; Bronkhorst and Plomp, 1992), while reduced spectral resolution has been shown to reduce speech intelligibility in noise (e.g. Baer and Moore, 1993, 1994), although this connection is less clear (George *et al.*, 2006).

Individual differences between listeners in the HI group were generally much larger than between listeners in the NH group. This is not surprising considering the relatively large range of different audiometric configurations in the HI group. The four-frequency average hearing loss (4-FAHL) ranged between 36 dB HL and 66 dB HL and the hearing

loss at high frequencies between 3 kHz and 6 kHz (the frequency region that was identfied as particularly important for the given task by Weller and Buchholz (2015)) even varied in a range of 40 dB. Additionally, other factors (e.g. cognitive capacities), where large differences between listeners are more likely to occur in older populations, contribute to the spread of the data in the HI group.

The model that was applied to predict the outcome of the experiment was able to successfully predict the overall sensitivity of the HI listeners, but considerably under-estimated the variation of the masked thresholds across subjects as well as target location. To better understand the reasons for this systematic error, Fig. 3.4 shows the (long-term) spectrum at the output of the model's auditory front-end for the target signal averaged over all locations (solid line), the masker (dash-dotted line), and the threshold simulating noise (dashed line). The variation of the target spectrum with location is indicated by the gray-shaded area which refers to ±1 standard deviation around the mean value over all (long-term) target spectra. For the threshold simulating noise a hearing loss was considered that was equal to the mean audiogram over all subjects. This figure demonstrates that the overall masking signal (consisting of the masker stimulus and the threshold simulating noise) is dominated by the masker below about 1 kHz and by the threshold simulating noise above about 1 kHz. Comparing the spectra of the masker stimulus and the target stimulus, it can be deduced that the SNR is highest at frequencies around $150-500$ Hz and above about 1.5 kHz. It was shown by Weller and Buchholz (2015) that especially the high frequency region was responsible for the large variation of masked thresholds over direction (mainly due to head-shadow effects) and distance (due to reverberation). However, at least for the average HI subject, this frequency region was not accessible due to the masking caused by the threshold simulating noise.

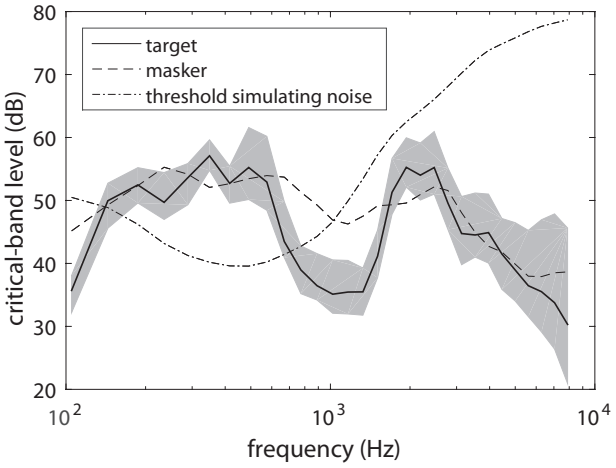


FIGURE 3.4: Critical band spectra of the different signals at the output of the auditory front-end of the detection model averaged over time. The gray area corresponds to ±1 standard deviation around the mean value of the target stimuli across all 15 target locations.

This means that the model as described here solely takes into account low frequencies (below 1 kHz) when predicting masked thresholds, where head shadow effects are weak and hence no significant variation of the model predictions over target direction can be observed.

This low frequency emphasis also results in model predictions that mainly reflect the differences in low frequency hearing loss between HI subjects. Since the low frequency hearing varied very little between subjects in the current study, the model predictions are grouped together as well. The only subject for which the model predicted substantially different thresholds than for the rest of the group (subject 4, see Fig. 3.3) was also the only subject that had a considerable low frequency hearing loss (50 dB at 250 Hz and 500 Hz). Considering the analysis of Fig. 3.3 (right panel), which suggests that mainly the frequency region between $1-4$ kHz drives the variation of measured masked thresholds over target location, it is not surprising that the model under-predicts this variation.

The low-frequency dominance in the auditory detection model and the resulting limited ability of the model to adequately reproduce the variation of the masked thresholds across both target location and subjects may be explained by the fact that individual hearing loss was solely simulated by the addition of a threshold simulating noise. Such an approach disregards other aspects of (sensorineural) hearing loss than just reduced sensitivity, such as reduced temporal and spectral resolution and reduced compression in the cochlea (or loudness recruitment). Other approaches that use a more detailed model of the cochlear processing have been more successful when accounting for individual hearing loss (Jepsen and Dau, 2011; Moore and Glasberg, 2004), albeit with simpler stimuli than the complex environment used in the present study. Hence, in order to more accurately predict the effect of an individual hearing loss on masked thresholds of speech in complex environments, future research may need to consider more elaborate auditory models that take into account the effects of hearing loss on auditory perception in more detail than the approach presented in this paper.

Chapter 4

The effect of spatial uncertainty on speech detection in a reverberant multi-talker environment[1]

When auditory localization is measured in complex environments it is important to maintain good control of the audibility of the target stimulus. This can be achieved by measuring masked thresholds for the relevant stimuli and source locations. However, measuring localization naturally introduces uncertainty about the location of the target stimulus. The goal of this study was to analyze if the effect of this spatial uncertainty on masked thresholds is an important factor to consider in the design of localization experiments. We measured psychometric functions for the detection of a short speech target in a simulated reverberant multi-talker background for eight target directions and in two conditions: a fixed-direction condition where the target direction was fixed over a block of trials and a random-direction condition where the target direction was randomly varied from trial to trial. Thresholds were 0.8 dB higher in the random-direction condition than in the fixed-direction condition for both normal hearing and hearing impaired listeners while the slopes of the psychometric functions remained unchanged in both groups. The effect of spatial uncertainty did not significantly differ between the different target locations. The results suggest that target location uncertainty plays a minor role in signal detection in reverberant multi-talker environments.

[1]Manuscript submitted for publication to The Journal of the Acoustical Society of America.

4.1 Introduction

Measuring auditory localization requires appropriate control of the audibility of the target stimulus. If the level of the target stimulus (in quiet) or the target-to-masker ratio (in noise) is set too low the data will be corrupted by trials where the target was inaudible (Lorenzi *et al.*, 1999b,a). This issue is further complicated if the acoustic environment used for the localization task is very complex, e.g. if it contains multiple, spatially distributed speech maskers in a reverberant space (Kopčo *et al.*, 2010). Controlling for audibility is especially important when trying to study sound localization in listeners with hearing loss (Best *et al.*, 2011).

One way to ensure the audibility of the target stimulus is to measure masked thresholds under the same conditions prior to the localization experiment. However, since during a typical localization experiment the target is presented from different directions, several masked thresholds (optimally one for every direction) have to be measured. Depending on the localization test design, this process may even need to be repeated for different target signals. Since this is a very time consuming task, an alternative method is to instead apply an auditory model and predict the thresholds. In this regard, Weller and Buchholz (2015) measured masked thresholds for a speech target at different locations in a reverberant, multi-talker background. They also applied a functional auditory model based on the signal energy at the ears that was able to predict these thresholds with reasonable accuracy.

However, the procedure in a localization task usually differs substantially from that in a detection task in that the target stimulus in each trial is presented from a random location that is unknown to the test subject. In a localization task, by definition, the subject does not have *a priori* knowledge about the target location. Neither of the aforementioned methods to control target audibility in a localization task (measuring and modeling masked thresholds) take into account this added effect of uncertainty about the target location (i.e. the spatial uncertainty). Masked thresholds measured for a fixed (and known) location may overestimate the audibility of a target in the uncertain context of a localization experiment.

It has been shown in the past that stimulus uncertainty can be detrimental to the listener's performance in different tasks. However, the magnitude of this effect appears to depend on the stimulus and task, and in particular on the presence of energetic vs informational masking. While introducing uncertainty about the frequency of a tonal target only decreased detectability of the target in white noise by about 3 dB (Green, 1961), uncertainty about the spectral content of a multi-tone masker was shown to increase thresholds of the tonal target by up to 40 dB even if there was no masker

energy in the frequency region around the target (Neff and Green, 1987; Wright and Saberi, 1999; Durlach *et al.*, 2005; Leibold and Werner, 2006).

Moreover, temporal uncertainty, i.e. uncertainty about the timing of the onset of the target signal, negatively affects auditory detection thresholds. The effect of temporal uncertainty on pure tone detection was found to be small in the presence of energetic masking (Egan *et al.*, 1961; Green and Weber, 1980) but significantly larger in the presence of informational masking as introduced by varying the spectral properties of the masker (Bonino and Leibold, 2008). This increase in the effect of temporal uncertainty on the listener's performance due to the presence of informational masking is further supported by the results of Best *et al.* (2007), who measured thresholds for the discrimination of birdsongs and the discrimination of spoken digits. They found that when the target and masker were highly similar and unfamiliar (birdsongs) a visual cue indicating the time interval containing the target was more helpful than when target and masker were less similar and more familiar (spoken digits).

There is also evidence that listeners are able to benefit from a focus of attention on a particular location. Mondor and Zatorre (1995) found that when the location of the stimulus changed from trial to trial, increasing the time between an auditory location cue and the presentation of the stimulus (and hence increasing the time available to focus attention) resulted in better performance in a tone discrimination task. They also found that performance declined with increasing distance between the cued location and the actual target location, suggesting a spatial gradient of attention. These findings are supported by outcomes of other studies (Teder-Sälejärvi and Hillyard, 1998; Allen *et al.*, 2009). Fan *et al.* (2008) also found that there was an effect of spatial uncertainty on the masked detection of narrowband noise signals, but that it was small relative to the effect of spectral uncertainty and overall level uncertainty.

In more complex scenarios, spatial attention has also been shown to improve performance. Kidd *et al.* (2005) found a significant effect of *a priori* knowledge about the target location on the listener's ability to identify the target speech in a mix with two, spatially separated, competing speech maskers. Brungart and Simpson (2007) observed an improvement in speech identification when the target location remained fixed between trials compared to when the target location was changed frequently. This effect was consistent over different spatial configurations with $2-4$ concurrent talkers. In a study measuring the ability to identify and recall a sequence of target digits presented with distracter digits at different locations, Best *et al.* (2008b) found that performance was better when the target location was fixed rather than changing between digits.

In contrast, Jones and Litovsky (2008) measured the intelligibility of target spondees in a mixture with one to three masker sentences. They found that varying the predictability

of the spatial configuration of the masker had no significant effect on the results and concluded that there is no added masking due to uncertainty about the masker locations in a complex environment.

The goal of this study was to investigate the effect of spatial uncertainty on the detection of a target word in a reverberant multi-talker background. This scenario allows direct comparison to a previous study (Weller and Buchholz, 2015) that measured masked thresholds in the same environment for a large number of target locations. Additionally, such a scenario is considered to be representative for many complex conditions encountered in real life. Motivated by previous results showing a significant effect of spatial uncertainty it was our aim to find out if spatial attention has to be accounted for when controlling target audibility in auditory localization tasks. To this end we measured psychometric functions for the detection of a speech target in a reverberant multi-talker background under two conditions: one condition where the target location was fixed during a block of trials and one condition where the target location was randomly changed from trial to trial. Since previous results have indicated that hearing impaired listeners may receive a reduced benefit from focusing attention (Best *et al.*, 2008a), we analyzed the effect of this added spatial uncertainty on the detection performance of both normal hearing (NH) and hearing impaired (HI) listeners.

4.2 Methods

4.2.1 Subjects

Nine NH and eight HI subjects participated in the study. The NH subjects were aged between 15 and 39 years (mean 29.3 years) and had hearing thresholds of less than 20 dB HL up to 8 kHz. The HI subjects were aged between 69 and 86 years (mean 78.4 years) and had symmetrical sensorineural hearing losses (differences of less than ±10 dB between both ears at audiometric frequencies up to 4 kHz), with the exception of one subject who had thresholds of 40 and 60 dB HL at the left and right ear respectively at a frequency of 0.25 kHz. The individual audiograms and the mean are shown in Fig. 4.1. NH subjects were either colleagues at the National Acoustic Laboratories (NAL) or students at Macquarie University. HI subjects were recruited from a database available at NAL and had no reported cognitive disability. Ethics approval was given by the Australian Hearing Human Research Ethics Committee and all subjects had given written consent prior to participating in the study. All subjects that were not employed at NAL received a gratitude for their participation.

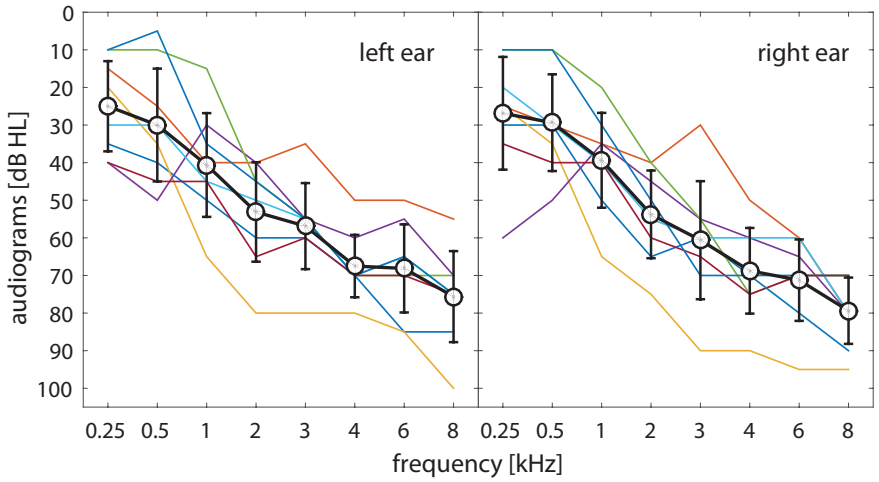


FIGURE 4.1: Individual audiograms and their mean values of the eight HI subjects that participated in the study. The error bars represent ±1 standard deviation.

4.2.2 Stimuli and setup

The speech signals were used for both the target and the masker. The target was a single utterance of the word "two" (duration ≈ 350 ms) spoken by a female and the masker was a mixture of seven concurrent dialogues. These dialogues consisted of scripted conversations that were taken from the "listening component" of past examinations of the International English Language Testing System (IELTS) and were read by eight females and six males. Anechoic recordings of these stimuli were spatialized using acoustic room simulation software, auralized using a MATLAB toolbox and then played back with an array of loudspeakers located in an anechoic chamber at NAL.

A shoebox room of dimensions 15 m × 8.5 m × 2.8 m was simulated in *Odeon* (Rindel, 2000) with reflective walls, floor, ceiling and furniture. The resulting reverberation time of the simulated room was $RT_{30} \approx 0.5$ s. This was considered a good representation of a realistic cafeteria-like environment. The seven pairs of talkers contributing to the masker were distributed across the room and the target source was simulated from eight different positions on a circle at 2 m distance from the listener's position (see Fig. 4.2), facing the listener.

The room impulse responses (RIRs) generated with *Odeon* were auralized using the loudspeaker-based room auralization (*LoRA*) toolbox for MATLAB (Favrot and Buchholz, 2010). The *LoRA* toolbox divides the RIRs simulated by the *Odeon* software into three components: the direct sound, the specular early reflections (here described by image sources up to order 3) and the late reverberant component. The direct sound and early reflections were mapped to the loudspeaker in the array whose position with respect to the listener was the closest to the sound component's direction of arrival at the listener's

position in the simulated room. The reverberant part of the RIRs was then realized by multiplying directional (first order) energy envelopes with noise that was uncorrelated across loudspeakers. The loudspeaker array used in the present study consisted of 41 loudspeakers covering a sphere with a radius of 1.85 m. The stimuli were then created by convolving the anechoic source signals with the transfer functions generated by the *LoRA* toolbox. The final masker was obtained by mixing the 14 distracter signals after the auralization. The anechoic speech signals contributing to the masker had been normalized to the same root mean square (RMS) value prior to the processing, and therefore their RMS level at the listener's position depended on their distance from and direction to the listener.

The stimuli were played back using a PC with an RME MADI soundcard. The output of the soundcard was connected to two RME M-32 D/A converters. The analog signals were then amplified by 11 4-channel Yamaha XM4180 amplifiers. For calibration purposes the masker signal was played back by the loudspeaker array. Using a sound level meter, the average sound level of the masker at the listening position in the center of the array was adjusted to a value of 65 dB SPL. According to predictions by *Odeon* this was the expected total sound pressure level of the described scene when each talker applied a normal conversational voice level. After summing up the 41 channels of the masker signal, the RMS level of the calibrated masker was then calculated in MATLAB

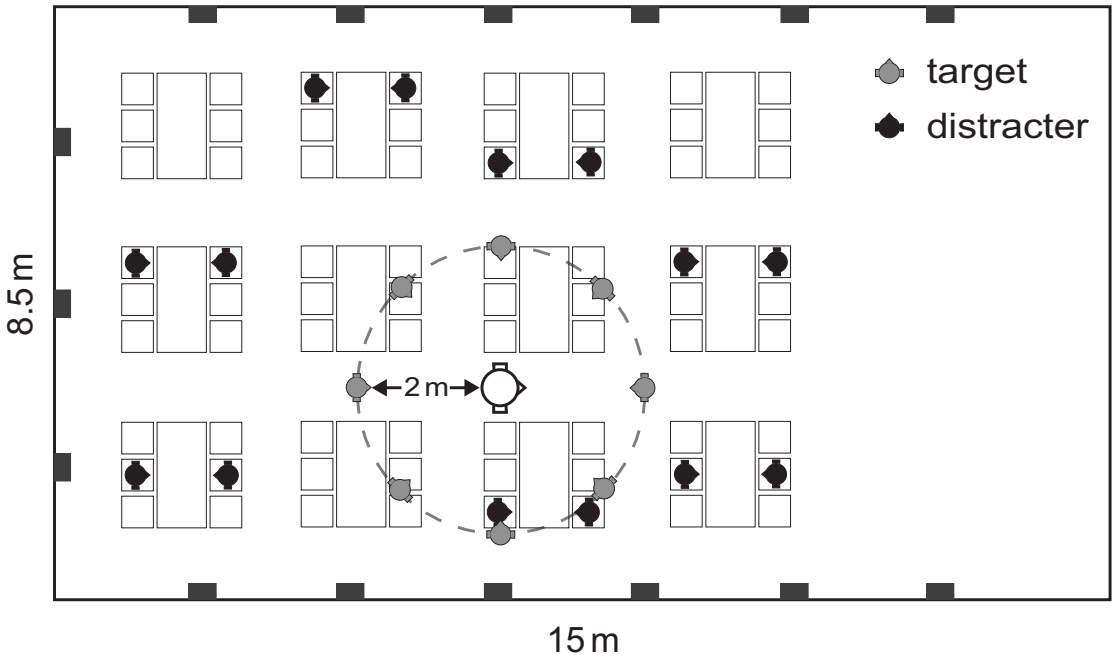


FIGURE 4.2: Top-down drawing of the simulated cafeteria scene. The transparent head represents the listener, gray heads indicate the target positions in 2 m distance from the listener and black heads indicate the positions of distracting sources.

and the target level adjusted accordingly for each target stimulus to obtain the desired signal-to-noise ratios (SNRs).

4.2.3 Procedure

In preparation for the measurement of the psychometric functions, masked thresholds were measured adaptively for each subject at all eight target locations (see Fig. 4.2, gray heads). These individual masked thresholds were used as reference points for the measurement of the psychometric functions. The psychometric functions were subsequently measured in two conditions: In the fixed-direction condition the target direction was fixed until all points on the psychometric function for this target direction were measured. In the random-direction condition the target direction was changed randomly from trial to trial. The experiment was divided into two appointments of approximately two hours, each with at least one break after the first hour. The adaptive measurement of the masked thresholds was performed during the first appointment and the measurement of the psychometric functions was performed during the second appointment. In the second appointment the fixed-direction condition was always measured first. The two appointments were spaced apart by as close to two weeks as possible (schedules allowing) for all subjects.

Throughout all stages of the experiment, subjects were sitting in the loudspeaker array such that their head was in the epicenter of the array. Subjects were asked to face the front loudspeaker at all times during the experiment but were not physically restrained. HI subjects were measured without their hearing aids.

The adaptive measurement of the masked thresholds was carried out using a 1-up 2-down adaptive track, thus measuring the 70.7 %-correct point on the psychometric function (Levitt, 1971). Each trial of the adaptive track was designed as a three-alternative forced choice (3-AFC) task. Subjects completed the 3-AFC task using a graphical user interface presented to them on an iPad. After starting the experiment, the 3 minutes long masker signal was looped and played continuously without interruption until the end of the experiment. Each test trial consisted of three intervals, only one of which contained the target stimulus in addition to the masker. The intervals were indicated to the subjects using three buttons in the GUI that were colored blue when the interval corresponding to the button was active and gray otherwise.

At each of the eight target locations the adaptive track was started at an SNR of 5 dB and the initial step size was 8 dB. The step size was decreased to 4 dB after the first two reversals in the adaptive track and to 2 dB after the next two reversals, and after twelve reversals the track was terminated. The masked threshold was then determined as the

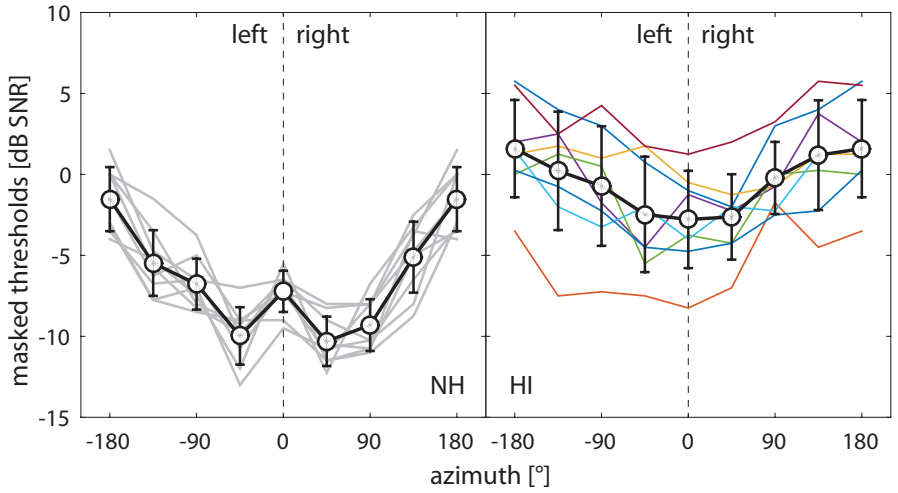


FIGURE 4.3: Individual masked thresholds and their inter-subject mean and standard deviation measured adaptively at all eight target locations. Left panel: normal hearing (NH) subjects. Right panel: hearing impaired (HI) subjects. The colors in the right panel correspond to those in Fig. 4.1.

mean SNR across the last eight reversals. The individual masked thresholds and the inter-subject mean and standard deviation are shown in Fig. 4.3. The masked thresholds varied with direction in both groups, but more so in the NH group. Generally, masked thresholds in the HI group were higher than in the NH group and they showed more variation between subjects. A detailed discussion and an auditory model-based analysis of these results can be found in Weller and Buchholz (2015).

The adaptively measured thresholds were then used as reference values for the measurement of the psychometric functions in the fixed-direction and random-direction conditions. In both conditions, five points on the psychometric function were measured at each target location: In the fixed-direction condition detection performance was measured at -6, -3, 0, 3 and 6 dB with respect to the reference SNR value, and in the random-direction condition performance was measured at -3, 0, 3, 6 and 9 dB with respect to the reference SNR value. The SNR values in the random-direction condition were chosen 3 dB higher than in the fixed-direction condition because results from pilot measurements suggested a slight upwards shift of the psychometric function introduced by the randomized presentation of the target locations.

The setup for the measurement of the psychometric functions was the same as for the adaptive measurement of the masked thresholds. In the fixed-direction condition, all target directions were measured one after the other from 0° to 315°, with the current direction indicated to the subject using a large text field on the GUI saying “front”, “front right”, etc. For each target direction, the SNR values were presented randomly with a total of 10 trials at each SNR for the NH subjects and 8 trials at each SNR for the HI subjects. In the random-direction condition the same number of trials was measured,

but in addition to the SNR values the target directions were changed randomly from trial to trial and no indication of the current direction was given to the subjects.

4.3 Results

To analyze the results, a logistic sigmoidal psychometric function was fitted to each data set individually using the *mpsignifit* toolbox for MATLAB (Fründ *et al.*, 2011), resulting in two psychometric functions for each subject at each target direction (one for the fixed-direction condition and one for the random-direction condition). The *mpsignifit* toolbox uses a Bayesian inference approach to fit psychometric functions. No constraints were put on the midpoint and width of the psychometric functions prior to the fitting. The SNR at the 0.707 proportion correct point of each of these psychometric functions provided an estimate of the masked threshold. Additionally, the slopes of the psychometric functions at their turning point were measured and compared.

The difference between the thresholds derived from the psychometric functions in the fixed condition and the adaptively measured thresholds was evaluated to assess the reliability of the measurements for each subject. Figure 4.4 (left panel) shows a scatter plot of the masked thresholds derived from the psychometric functions plotted against the adaptively measured masked thresholds. Most of the data points are grouped reasonably close to the diagonal with the exception of several data points belonging to subject 5 in the group of HI listeners. Figure 4.4 (right panel) displays the root mean square (RMS) error between the masked thresholds in the fixed condition and the adaptively

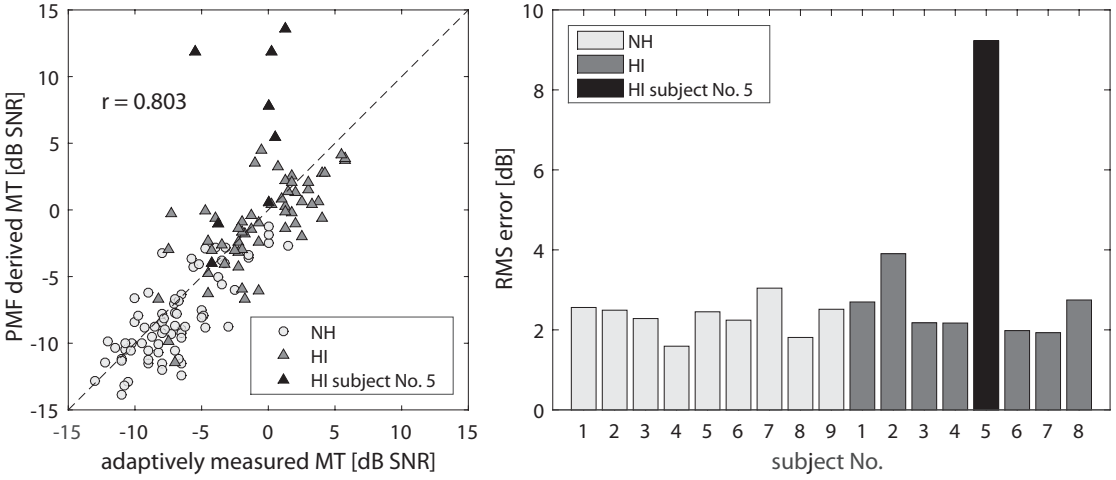


FIGURE 4.4: Left panel: scatter plot of the masked thresholds derived from the fitted psychometric functions in the fixed-direction condition plotted against the adaptively measured masked thresholds. Right panel: the RMS error between the thresholds in the fixed-direction condition and the adaptively measured thresholds plotted for all subjects.

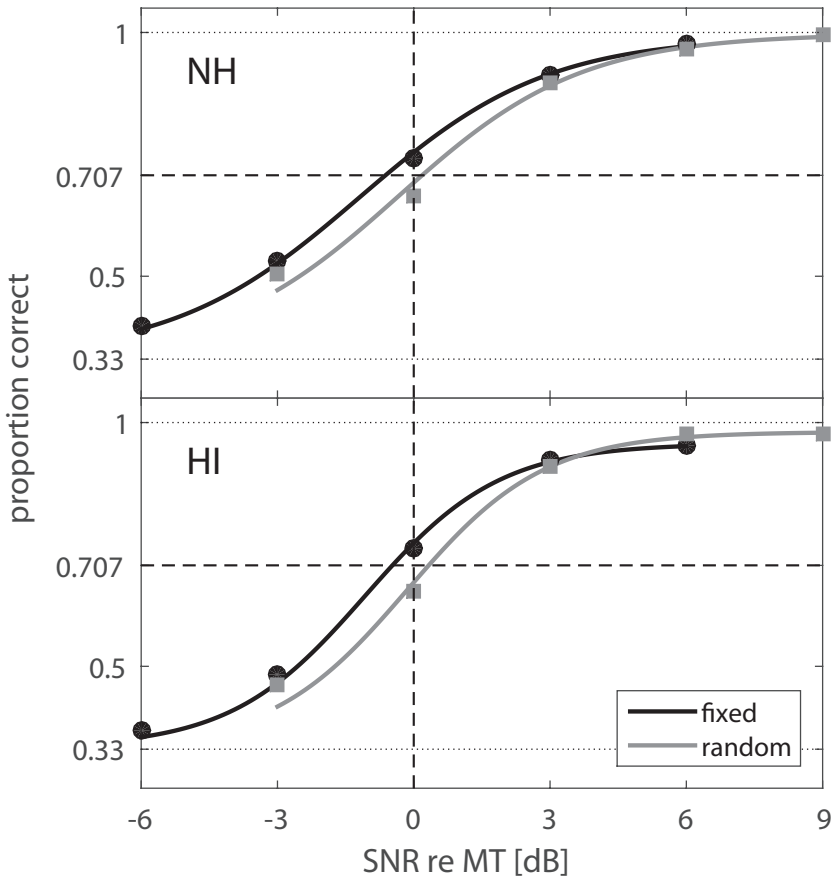


FIGURE 4.5: Psychometric functions for the two conditions fitted to the data averaged across all subjects in a group and all target directions, NH group (top panel) and HI group (bottom panel). The SNR values are given with respect to the reference values, i.e. the adaptively measured thresholds (dashed line).

measured masked thresholds for each subject averaged over the eight different target locations. It shows rather small and consistent RMS errors for all subjects except for HI subject 5 who shows a drastically increased RMS error. Two methods to detect statistical outliers were applied to the RMS error data, both of which identified the data of HI subject 5 as a statistical outlier: Grubb's test (Grubbs, 1950): $G > 2.38$, $\alpha = 0.001$, and Dixon's Q test (Dean and Dixon, 1951): $Q > Q_{lim} = 0.634$, $\alpha = 0.01$ (for a table of Q_{lim}-values see Rorabacher, 1991). Their results, paired with the graphical evidence in Fig. 4.4, motivated the exclusion of HI subject 5 from further analysis. The correlation coefficient between the thresholds measured adaptively and the thresholds derived from the psychometric functions (see Fig. 4.4, left panel) was $r = 0.776$ for the NH group and $r = 0.725$ for the HI group (excluding subject 5). The high overall correlation and the low RMS error (Fig. 4.4, right panel) of approximately 2 dB on average suggests consistency of thresholds across measurement methods.

In Fig. 4.5 the psychometric functions fitted to the data averaged over all subjects and target directions are shown for both groups. For both NH and HI subjects, the threshold

extracted from these fits was 0.8 dB higher in the random condition than in the fixed condition. No difference in the slopes of the average psychometric functions was found for NH subjects (0.08/dB in both conditions) nor for HI subjects (0.11/dB in both conditions), but slopes were slightly steeper for HI subjects than for NH subjects.

To determine if the difference in thresholds between the fixed-direction condition and the random-direction condition (i.e. the effect of spatial uncertainty) was significant, a linear mixed-effects model (MIXED) analysis was performed on the individual thresholds derived from the psychometric functions with the factors being group, condition and direction. A MIXED analysis was chosen over a two-way repeated measures analysis of variance (ANOVA) due to the unbalanced sample size between the two groups. Additionally, no threshold could be extracted from the data of HI subject 2 in the random-direction condition at 45° azimuth, rendering the data set within the HI group unbalanced as well.

The MIXED analysis revealed no significant effect of group, a marginally significant effect of condition ($F(1, 223) = 4.321$, $p = 0.039$), a significant effect of direction ($F(7, 223) = 3.908$, $p < 0.001$) and a significant interaction between group and direction ($F(7, 223) = 3.336$, $p = 0.002$). The interactions between group and condition as well as condition and direction and the three-way interaction between all of the factors were not significant. Despite the lack of significant interaction between condition and direction, a closer look was taken at the role of the target direction. In both groups, paired comparisons with Bonferroni correction between the thresholds in the two conditions were carried out for all eight directions. A significant effect of condition was only found for the thresholds measured at a target direction of 90° (NH: $t = 2.399$, $p = 0.043$, HI: $t = 3.583$, $p = 0.012$). The means and standard deviations of the paired comparisons are shown in Fig. 4.6.

4.4 Discussion

The analysis of the data revealed a statistically significant effect of spatial uncertainty (i.e. the listening condition) on masked detection thresholds. However, this effect was very small (0.8 dB on average for both NH and HI subjects). In the context of controlling audibility in localization experiments the effect of spatial uncertainty, since it was largely independent of target direction, could be modeled by a simple SNR offset. Compared to the overall difference between thresholds at different directions (≈ 9 dB at a distance of 2 m, Weller and Buchholz, 2015), however, the effect of spatial uncertainty was so small that it may even be neglected.

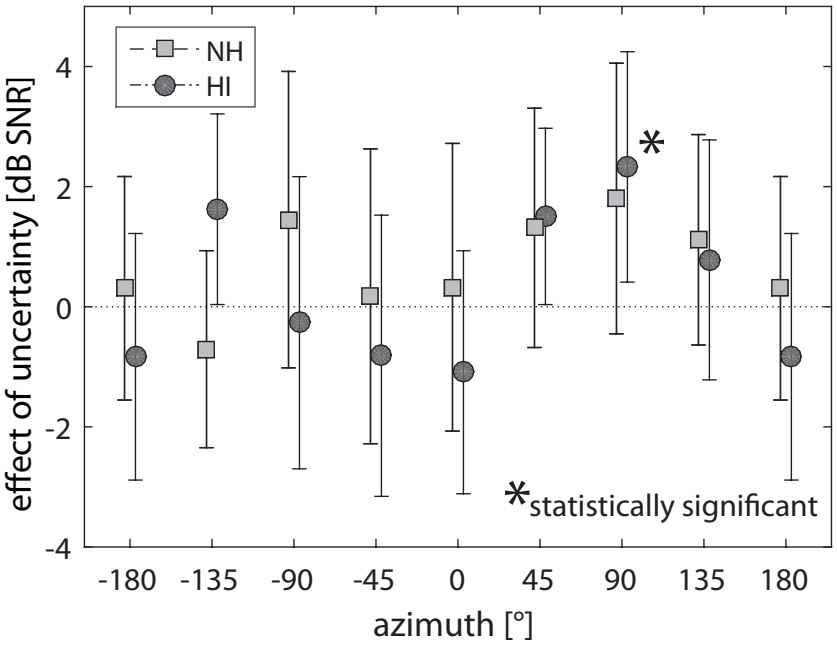


FIGURE 4.6: Inter-subject means and standard deviations of the difference between masked thresholds in the fixed-direction condition and the random-direction condition (i.e., the spatial uncertainty) plotted over target direction. The effect of spatial uncertainty was only found to be statistically significant at a target direction of 90°.

This very modest effect of spatial uncertainty that was found here for speech detection is somewhat surprising given that a number of studies have reported rather large uncertainty effects. Kidd *et al.* (2005), for instance, analyzed the effect of uncertainty about the target stimulus location in a speech identification task. In each trial they presented three sentences taken from the Coordinated Response Measure (CRM) corpus from three different directions (−60°, 0° and +60°) and subjects had to identify and repeat the color and number of the sentence containing the callsign. In their "callsign after" condition (i.e. the callsign was not known before the trial) they found that speech identification performance dropped from 0.92 proportion correct when the direction of the target sentence varied to approximately 0.35 proportion correct when the target direction was constant. Even when the callsign was known before the trial ("callsign before" condition), performance dropped from 0.91 to 0.67 proportion correct for varying target location.

Similarly, Brungart and Simpson (2007) measured speech identification performance of NH subjects in conditions of varying degree of spatial uncertainty using the CRM corpus. Their spatial setup differed slightly from the one used by Kidd *et al.* (2005) and the spatialized stimuli were presented over headphones. They found that in blocks of trials where there were no transitions of the talker locations between trials, approximately 89 % of the target sentences were identified correctly when there were three talkers active (one target and two distracters) and 75 % were identified correctly when four talkers were

active (one target and three distracters). When the probability of inter-trial transitions of the talker locations was set to 1, performance decreased to 71 % correct and 58 % correct respectively.

Best *et al.* (2008b) measured the effect of spatial uncertainty on the ability to recall a sequence of four spoken digits. Five sequences were presented simultaneously from five loudspeakers distributed horizontally in front of the listeners, one of them being the target sequence. Spatial uncertainty was introduced by presenting the four digits of the target sequence from different, randomly chosen loudspeakers as opposed to presenting all digits in the target sequence from a single loudspeaker. For every digit, the loudspeaker presenting the digit of the target sequence was indicated using a visual cue. The results show a "cost" of switching spatial attention, i.e. the difference between the presence and absence of spatial uncertainty, between 7 percent points and 17 percent points depending on the time delay between digits (the higher the delay, the lower the cost).

In the current study the masked thresholds in both groups were on average 0.8 dB higher in the random-direction condition than in the fixed-direction condition. Assuming an average slope of the underlying psychometric function of about 0.08/dB for NH subjects and 0.11/dB for HI subjects (Fig. 4.5), this translates into a decrease in performance of about 0.064 and 0.088 proportion correct, respectively. This effect is much smaller than the one found by Kidd *et al.* (2005) and Brungart and Simpson (2007), but it is similar in size to the effect found by Best *et al.* (2008b). However, in the latter experiment a visual cue for the target direction was provided to the listeners synchronously with the presentation of the stimuli, which may have reduced the observed effect size.

There are several differences between the studies mentioned above and the current study that may explain the different outcomes. Firstly, all previous studies used different kinds of speech identification tasks while the current study employed a speech detection task. Additionally, the different studies involved very different levels of informational masking. It has been shown that the presence of informational masking enhances the effect of temporal uncertainty (Best *et al.*, 2007) as well as spatial uncertainty (Fan *et al.*, 2008). Whereas the previous three studies all involved a substantial amount of informational masking, this was not the case for the present study. Westermann and Buchholz (2015) showed for a very similar reverberant environment with seven distracter-pairs as used here that informational masking plays a very minor (or even negligible) role. Hence, the low level of informational masking may be responsible for the small effect of spatial uncertainty that was observed in the present study. Jones and Litovsky (2008) did not find a significant effect of spatial uncertainty when measuring speech intelligibility in a cocktail party scenario, even though their study was explicitly designed to produce

a high level of informational masking. However, this study differs substantially from the ones mentioned above in that spatial uncertainty was introduced by varying the locations of the distracters while the location of the target was known to the listener.

Even though a significant direction-specific effect of spatial uncertainty was only found in the current study for the thresholds measured with a target direction of 90°, the mean effect of spatial uncertainty showed a certain tendency over direction: With the exception of the NH group at 135° the mean effect of spatial uncertainty was higher for lateral target directions (between ±45° and ±135°) than for target directions in the front (0°) or back (±180°). The mean effect of spatial uncertainty was also stronger for target locations on the right-hand side of the listeners than on their left-hand side. Previous studies in more complex environments reported that the listener's ability to attend to speech was better in the right hemisphere than in the left hemisphere (Bolia *et al.*, 2001; Ericson *et al.*, 2004), and Brungart and Simpson (2007) suggested that there may be a general bias to selectively focus on speech on the right-hand side. The adaptively measured masked thresholds (see Fig. 4.3) were indeed slightly lower on the right side than on the left side, however, this may also have been caused by the non-symmetric nature of the simulated scenario. If and how such a general bias would influence the effect of spatial uncertainty is unclear and cannot be predicted from the present data.

Given that each subject was tested at a different absolute SNR, according to their adaptively measured masked thresholds, we examined whether this had any influence on the measured effect of spatial uncertainty. Figure 4.7 shows a scatter plot of the effect of spatial uncertainty as a function of absolute SNR instead of SNR relative to the reference points. A simple linear regression analysis of the relation between the effect of spatial uncertainty and SNR for both subject groups combined showed a weak, but significant correlation ($R^2 = 0.103$, $p < 0.001$). Performing the same analysis on both groups separately showed a significant correlation in the NH group ($R^2 = 0.222$, $p < 0.001$) and only a marginally significant correlation in the HI group ($R^2 = 0.08$, $p < 0.036$). A negative correlation between the SNR and the effect of spatial uncertainty may, at least partly, explain the observed variation of the effect of spatial uncertainty over target direction (see Fig. 4.6). The SNR was typically lower at lateral target locations (±90°) than at the front (see Fig. 4.3) and thus, should exhibit a larger effect of spatial uncertainty.

It is also worth noting that the statistical analysis of the masked threshold data revealed a significant overall effect of the target direction on the masked thresholds. This is rather surprising since the applied normalization of the masked thresholds to adaptively measured reference SNRs (measured for each subject and direction individually) should

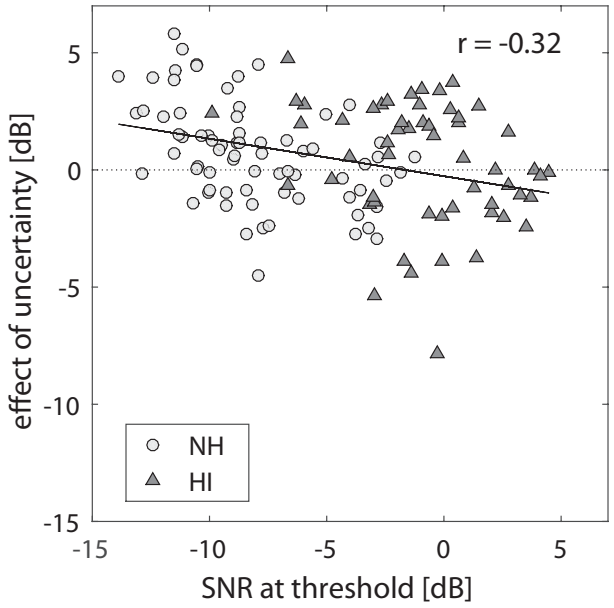


FIGURE 4.7: Scatter plot of the effect of spatial uncertainty as a function of SNR between the target and the masker in the fixed-direction condition. The correlation coefficient for the combined data of both groups was $r = -0.32$.

have removed any dependency on the target direction. Thus, it can not be ruled out that there is an effect of the measurement method that affects target directions differently.

It should be noted that the presentation order of the two listening conditions was not counterbalanced in the experiment, i.e. for all subjects, the fixed-direction condition was presented before the random-direction condition. This may have resulted in an underestimation of the effect of spatial uncertainty due to a training effect (subjects learning the procedure and improving their performance over the course of the experiment). To determine if subjects improved over the course of the experiment we compared the results from the first half of the random-direction condition to the results from the second part. A 2-way repeated measures ANOVA of the data averaged across directions with time (first half/second half) and SNR (the five measurement points on the psychometric function) being the factors showed no significant effect of time and no significant interaction between time and SNR. Therefore, it may be assumed that even if there was a training effect it would have been very small.

4.5 Conclusions

We measured psychometric functions for the detection of a short speech target in a reverberant multi-talker background for NH and HI listeners. These psychometric functions were measured in two conditions, one where the target direction was fixed and one where the target direction was varied randomly from trial to trial. The difference

in thresholds between these two conditions provided a measure of the effect of spatial uncertainty on masked thresholds.

The slopes of the psychometric functions did not differ between these two conditions, there was, however, a significant difference in masked thresholds, demonstrating a small but significant effect of spatial uncertainty of about 0.8 dB. While the masked thresholds themselves varied strongly over target direction, the target direction did not significantly change the effect of spatial uncertainty on the masked thresholds. Hearing status had a small effect on the slope of the psychometric functions but it did also not change the effect of spatial uncertainty on the masked thresholds. The observed uncertainty effect was significantly smaller than previously reported in the relevant literature, which may be explained by the relatively low level of informational masking in our simulated cafeteria.

In the context of measuring masked thresholds to control the audibility of the target stimulus in a localization task, the effect of spatial uncertainty that is introduced by the localization task is negligible compared to the large variation of the masked thresholds over target location. It is possible, however, that larger spatial uncertainty effects would be observed in different listening environments that produce a higher level of informational masking.

Chapter 5

Multi-talker localization in reverberant environments of varying complexity[1]

Deficits in spatial hearing can have a negative impact on a listener's ability to follow conversations in noisy backgrounds and may exacerbate the experience of hearing loss as a handicap. However, there is still a lack of methods that reliably capture the spatial hearing performance of listeners and accurately reflect their experience in daily life, in particular, in challenging acoustic environments. In this study, a novel approach was therefore proposed that measured auditory spatial awareness in a multi-talker scenario. Ten normal hearing and 16 hearing impaired subjects were presented with sounds scenes of 45 seconds duration that contained between one and six active talkers in a simulated reverberant environment. The task was to count and locate the talkers in direction and distance within the given time frame. Whereas the normal hearing subjects were able to reliably analyze scenes with up to four simultaneous talkers, most hearing impaired subjects already struggled with identifying four talkers at a time. Moreover, localization performance decreased in both groups with increasing number of talkers and was significantly reduced in the hearing impaired group. Overall performance revealed a significant correlation with hearing loss. These results showed that the proposed approach was useful to measure localization in direction, but that the design of the experiment was not well suited for measuring localization in distance. Future studies should refine the presented approach, for example by providing unique talker identities or by improving measurement as well as data analysis methods that better evaluate distance perception.

[1]Manuscript submitted for publication to The Journal of the American Academy of Audiology.

5.1 Introduction

It is widely acknowledged that hearing impaired (HI) listeners experience difficulties with spatial hearing. This is often reported anecdotally or in clinical reports, and was also shown in a questionnaire-based study conducted by Gatehouse and Noble (2004). The latter also concluded that spatial hearing deficits have a substantial impact on how hearing disability is experienced as a handicap. Even though the restoration of speech intelligibility is usually regarded as the main task of a hearing aid, these findings have led to many new hearing aids also aiming to preserve or even enhance spatial hearing.

However, surprisingly little is known about the causes of spatial hearing deficits in listeners with sensorineural hearing loss. Hawkins and Wightman (1980) found that discrimination thresholds for interaural time differences were reduced in HI listeners, but results varied greatly between listeners and there was no clear relationship between discrimination thresholds and hearing loss configuration. Akeroyd and Guy (2011) reported that localization dominance, a phenomenon related to the precedence effect, was significantly reduced in HI listeners, but they observed high individual variation of the results as well. This reduction of localization dominance may be linked to self-reported difficulties in everyday spatial perception of sound.

Localization, perhaps the most accessible aspect of spatial hearing, has been widely studied in the past. Localization performance in quiet was reported to be deteriorated only in listeners with unilateral hearing loss, substantial low-frequency hearing loss (Häusler *et al.*, 1983; Rosenhall, 1985; Noble *et al.*, 1994), or with conductive hearing loss (e.g. Durlach *et al.*, 1981; Dillon, 2012). In contrast, there are only a few recent studies that have measured localization in noise. Lorenzi *et al.* (1999b,a) measured the localization of broadband click trains in white noise, and, more recently, Kopčo *et al.* (2010) and Best *et al.* (2011) measured localization in more realistic environments with speech maskers. All these studies indicated that localization performance in noise may be reduced in hearing impaired subjects, but at the same time, they also highlighted that the reduced localization performance may be, at least partly, due to limited target audibility.

Although distance perception plays a significant role in the spatial awareness of acoustic scenes, very few studies have investigated distance perception in HI listeners. Based on their speech, spatial and qualities of hearing scale (SSQ) questionnaire Noble and Gatehouse (2004) pointed out that HI listeners reported deficits in their capability to perceive distance or motion of sound sources. Considering relative distance perception (i.e. just noticeable differences (JNDs) in distance) Akeroyd *et al.* (2007) showed that when level cues dominated distance perception, (elderly) HI listeners performed equally well as

normal hearing (NH) listeners. However, when reverberation cues became dominant HI listeners performed significantly poorer than NH listeners.

Single-source localization tasks are, however, not a very accurate representation of situations that listeners experience in their everyday life. These situations are often very dynamic, containing multiple sources at different locations (distances and directions) that may have to be attended to and moving sources or objects that constantly change the auditory scene. Consequently it is difficult to believe that performance measured in a single-source localization task fully captures a listener's spatial hearing ability. In this context, Brungart *et al.* (2014) conducted a study where they measured source identification and localization of short environmental sounds in a complex acoustic scene using three different tasks: A single-source localization task, an added source task where listeners had to identify and localize a sound added to an existing scene, and a removed source task where listeners had to identify and localize sound removed from an existing scene. Their general results indicated that performance decreased with increasing task complexity and increasing number of competing sound sources (in the added source task and removed source task) in both NH and HI listeners. They were able to show that the added source task was as sensitive to differences in performance between different listening configurations (unaided, monaurally aided, bilaterally aided) in a single listener as the single-source localization task, but that the results of the single-source localization task and the added source task across listeners were only moderately correlated. They suggested that this may indicate that these two tasks measured different aspects of spatial hearing. These results emphasize the necessity for tests that assess spatial hearing in more complex environments that are reliable and ecologically valid.

In contrast to Brungart *et al.* (2014), who considered localization (and identification) performance of a single object that changed (was added or removed) within a given scene, we were interested in measuring listeners' performance in analyzing a continuous spatial scene. In a first implementation of such a test we exposed listeners to auditory scenes containing multiple talkers in a simulated reverberant environment. Their task was to count and locate the different talkers in the scene in a limited amount of time. To our knowledge it is the first test of its kind that obtains localization estimates for all sources, rather than just one, in a complex scene. We considered such a task to represent a scenario that is relevant in daily life, where accurate localization may be important for listeners following a conversation between multiple talkers to spatially focus their attention on the active talker and suppress other talkers. In this context the presented spatial "awareness" task may be considered an indirect measure for the ability of a listener to participate in dynamic multi-talker conversations. In this study we describe the design of the test, discuss some of the issues that arose during data analysis, and present preliminary results in listeners with normal and impaired hearing.

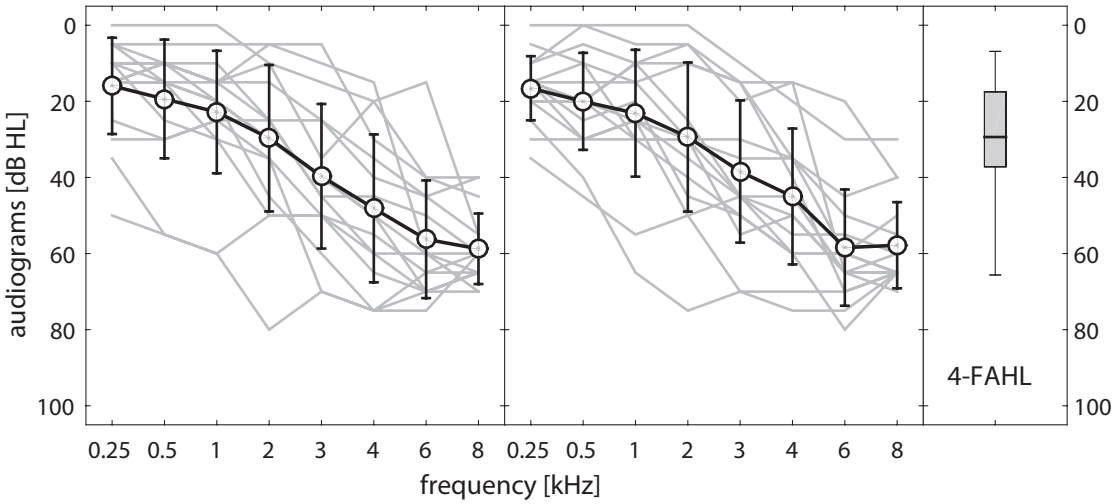


FIGURE 5.1: Individual audiograms and their means of the HI subjects that participated in the study. The error bars represent ±1 standard deviation. Left ear audiograms (left panel), right ear audiograms (center panel) and the four frequency average hearing loss (4-FAHL, right panel).

5.2 Methods

5.2.1 Subjects

Ten normal hearing (NH) and 16 hearing impaired (HI) subjects participated in the study. The NH subjects were aged between 20 and 31 years (mean 25.7 years) and had pure tone thresholds of less than 20 dB HL at audiometric frequencies up to 8 kHz. The HI subjects were aged between 52 and 85 years (mean 69 years) and had symmetrical sensorineural hearing losses (i.e differences between ears < 15 dB at all audiometric frequencies up to 4000 Hz) with the four-frequency (500, 1000, 2000, and 4000 Hz) average hearing loss (4-FAHL) ranging between 7 and 60 dB HL (mean 29.7 dB HL). The individual audiograms and their mean and standard deviation are shown in Fig. 5.1 together with a box plot of the 4-FAHL across subjects. NH subjects were either colleagues at the National Acoustic Laboratories (NAL) or students at Macquarie University. HI subjects were recruited from a database available at NAL and had no reported cognitive disability. Ethics approval was given by the Australian Hearing Human Research Ethics Committee and all subjects had given written consent prior to participating in the study. All subjects that were not employed at NAL received a gratuity for their participation.

5.2.2 Setup

The experiment was conducted in a virtual sound environment (VSE) located in the Australian Hearing Hub. This VSE consisted of a loudspeaker array with 41 Tannoy

V loudspeakers installed in an anechoic chamber. The loudspeakers were arranged in horizontal rings at different elevation angles covering the surface of a sphere with a radius of 1.85 m. There were 16 loudspeakers at 0° elevation, 8 loudspeakers each at ±30°, 4 loudspeakers each at ±60° and 1 loudspeaker in the zenith. The test subjects were seated on a height-adjustable chair such that their heads were always inside the epicenter of the loudspeaker array. The loudspeakers were driven by 11 4-channel Yamaha XM4180 amplifiers, which were connected to a PC outside the anechoic chamber via two RME M-32 analog-to-digital converters and a RME MADI sound card. All sound signals were generated by the PC running MATLAB.

5.2.3 Stimuli

For the experiment 96 different acoustic scenes were generated. Each scene contained a mixture of between one and six concurrent talkers in a reverberant environment and was auralized using the room acoustic simulation software *Odeon* (Rindel, 2000) and the Loudspeaker-based Room Auralization (LoRA; Favrot and Buchholz, 2010) toolbox in MATLAB.

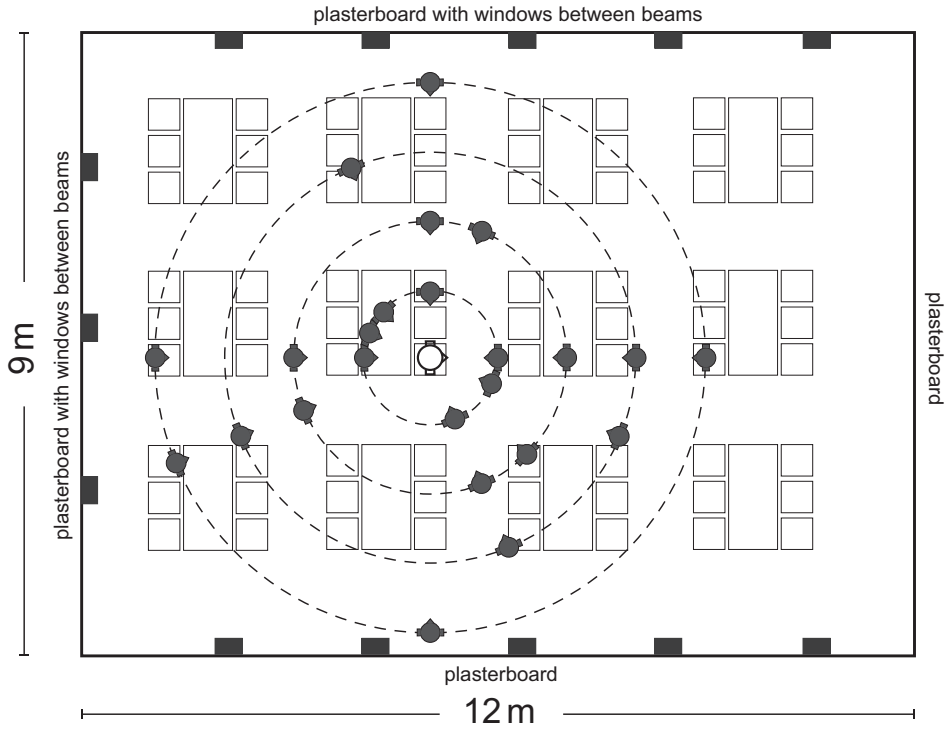


FIGURE 5.2: Top-down drawing of the simulated room. The listener location is indicated by the transparent head and the 24 source locations used in the experiment are indicated by the gray heads. Tables and chairs were modeled by wooden plates, the floor by 6 mm carpet with closed-cell foam on concrete, and the ceiling was suspended by 200 mm with 25 mm of wood-wool in the cavity.

A shoebox room of dimensions $9\,\text{m} \times 12\,\text{m} \times 2.8\,\text{m}$ (width $\times$ depth $\times$ height) was modeled in *Odeon* with reflective walls, floor and ceiling, and some furniture. The reverberation time of the simulated room was $\text{RT}_{30} \approx 0.45\,\text{s}$. The listener was positioned slightly off center in the simulated room to avoid undesirable sound coloration effects. To choose an appropriate set of source locations inside this room, first a grid of potential source locations was generated with distances of 1, 2, 3, and 4 m and 16 directions with a spacing of 22.5°. These directions coincided with the physical locations of the 16 loudspeakers inside the horizontal ring of the playback array. To allow for a balanced distribution of source locations across all scenes, a subset of 24 locations was then chosen out of the total of 64 possible locations. Nine of these 24 source locations were chosen manually to be at 0°, −90° and −180° in 1, 2 and 4 m distance to allow for a systematic analysis of the effects of distance and direction, the remaining 15 source locations were then chosen randomly. A schematic of the room and all relevant locations are shown in Fig. 5.2.

To create the scenes, anechoic recordings of 14 scripted monologues that were read by nine female and five male native Australian speakers were cut into snippets of 45 seconds, resulting in a total of 143 snippets. For each of the 96 scenes either 1, 2, 4 or 6 of these snippets were then chosen such that the same talker would not appear more than once in the same scene and all talkers were used roughly equally often across all scenes. Each snippet was then assigned a source location randomly from the subset of 24 source locations. Overall 96 scenes were created, 24 each with 1, 2, 4 and 6 talkers. In each of the 24 scenes corresponding to a particular number of talkers all source locations were used equally often, i.e. each source location was used once in the 24 scenes containing one talker, twice in the 24 scenes containing two talkers, four times in the 24 scenes containing four talkers and six times in the 24 scenes containing 6 talkers. Across all 96 scenes each source location was used 13 times.

The scenes were compiled by auralizing each 45-second speech snippet separately corresponding to the assigned source location and then mixing the auralized signals of all talkers present in the scene. The auralization of the signals was done in a two-step process. First, *Odeon* was used to obtain a representation of the room impulse response (RIR) between the location of each individual source and the listener location. Then, the *LoRA*-toolbox was used to create a set of filters that map the individual RIRs to the 41 loudspeakers in the playback array. Next, the corresponding anechoic speech snippets were convolved with this set of filters to generate the 41-channel loudspeaker signals. Finally, the loudspeaker signals of all talkers present in the scene were summed up resulting in one 41-channel signal representing a scene of a number of concurrent talkers at different directions and distances around the listener in a reverberant room.

In the described auralization process, the sound pressure level of all the realized sources (and scenes) at the listener's location is provided by the *Odeon* software and preserved by the LoRA processing. However, the absolute playback level inside the loudspeaker array needed to be calibrated. This was achieved by measuring the time-averaged sound pressure level of two different scenes (one with one talker and one with six talkers) with a sound level meter placed at the listener's position in the loudspeaker array. The measured values were then compared to the sound pressure level predicted by *Odeon* (assuming all talkers applied a normal conversational voice level) and the overall loudspeaker gain adjusted accordingly. Finally, the average of the two derived gains was applied to all the scenes. The resulting sound level of the presented scenes varied between 57 dB SPL (one talker) and 68 dB SPL (six talkers).

5.2.4 Procedure

The experiment consisted of a training phase and a measurement phase. In both phases, subjects had to analyze a number of different scenes regarding the number of present talkers, their position in the virtual room with respect to the listener's position (in distance and azimuth) and their sex.

Prior to the start of the experiment the subjects were asked to remove their hearing aids if they were wearing any. They were then asked to sit down on the chair in the center of the loudspeaker array, and the height of the chair was adjusted such that their head was in the epicenter of the loudspeaker array. The subjects had several minutes to go through the instruction sheet before the experimenter explained the procedure to them. They were also encouraged to turn their head during the experiment but at the same time they were asked to refrain from moving the upper body. The latter was necessary to avoid a less than perfect sound reproduction as it is experienced off center in the loudspeaker array (Favrot and Buchholz, 2010).

The progress of the experiment was controlled by the subjects themselves. At the start, the subjects were given an *iPad* with a graphical user interface (GUI) that was manipulated using the touchscreen of the device. The *iPad* was connected via WI-FI to the control computer outside the anechoic chamber using the *Splashtop Streamer* software. The GUI contained a top-down representation of the simulated room, a 45-second timer, a scene counter to display the progress in the experiment and a start button in the top left corner. In the top-down representation of the simulated room, a button was placed at each of the 64 potential source locations. A picture of the GUI is shown in Fig. 5.3.

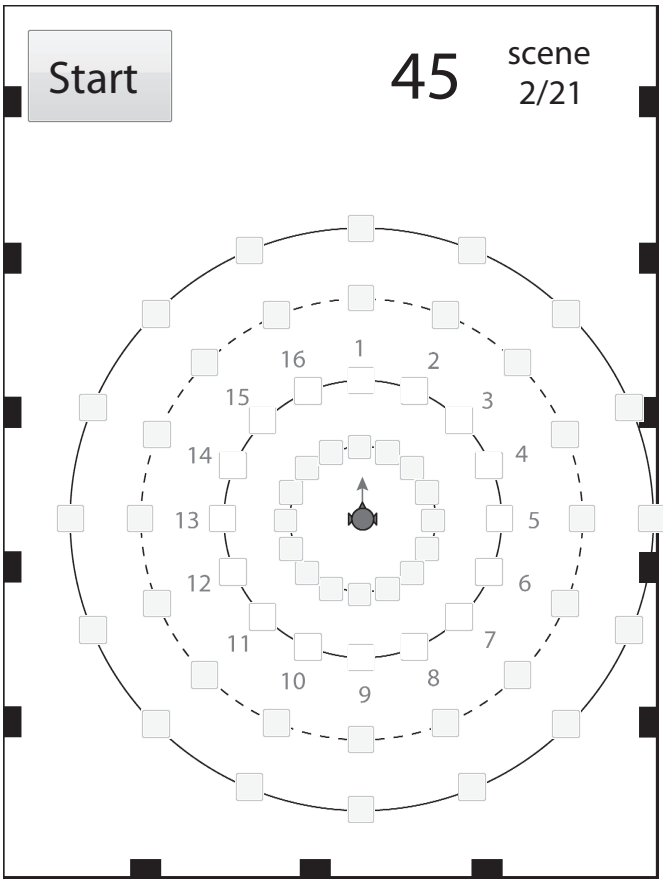


FIGURE 5.3: The graphical user interface (GUI) that the subjects used to give their responses and control the progress of the experiment. The GUI contains a top-down representation of the simulated room, a button in the top left corner, a timer and a progress display. Each of the potential source locations in the simulated room is represented by a push button. The numbers correspond to the loudspeakers in the central ring of the loudspeaker array and were displayed to aid with orientation.

The experiment was started by pressing the start button. After the scene had loaded the subjects had 45 seconds to count the number of talkers active in the scene, locate them and identify their sex. To respond, the subjects used the buttons corresponding to the location where they perceived a talker. Pressing the button once turned the button blue (indicating a male talker), pressing the button twice turned the button red (indicating a female talker) and pressing the button a third time reset the button to its original state. Subjects were given a capacitive *Stylus* touch pen to operate the buttons. After 45 seconds the scene stopped playing and the subjects had an additional five seconds to finalize their response. When the five seconds had passed the next scene was automatically loaded and the buttons reset. If the subjects had completed their response before the end of the 45 seconds they were able to skip to the next scene by pressing the start button. The time remaining for the current scene was indicated using the timer at the top of the GUI.

Before starting the measurement phase all subjects had to complete a training sequence. The training sequence consisted of a total of 21 scenes, twelve scenes containing one talker, four scenes containing two talkers and one scene each containing three, four, five, six and seven talkers, respectively. The scenes containing one talker were designed to help the listener developing a sensitivity for different target distances and target directions, while the scenes containing multiple talkers were included in the training sequence to give the listener a feeling for the more complex scenes in the experiment. During the entire training phase special attention was paid to the listener's understanding of the procedure and their ability to handle the GUI. Furthermore, a short feedback phase of ten seconds was added after each scene during which the buttons corresponding to the correct source locations were highlighted in the color corresponding to the talker's sex. If the experimenter noticed that a subject had difficulties completing the task properly the training sequence was repeated until the subject had fully understood the task. Unlike during the training phase, the presentation order was randomized in the measurement phase. The subjects were encouraged to request a break whenever they felt fatigued but a break was enforced after 48 of the 96 scenes were completed if the subject did not do so.

5.2.5 Analysis methods

To analyze the localization error in a multi-talker localization task, each source in the response has to be matched with a target source in the presented scene. Due to the design of the current study there was no definite solution for a perfect (true) matching because there was no way of uniquely identifying the target sources. Instead, each source was identified based on only three properties: distance (from the listener), direction (with respect to the front loudspeaker) and sex (male/female). The lack of a unique source identity makes the interpretation of the responses ambiguous, e.g. if there are two target sources at 1 m, 0° and 2 m, 22.5° respectively and the responses are at 2 m, 0° and 1 m, 22.5° then there is no way to find out which response was made for which target source. Therefore, some assumptions have to be made about the way the subjects make their responses in order to eliminate the ambiguity in the interpretation of the results as best as possible.

Based on these assumptions, a matching algorithm can be applied that evaluates all possible ways to match the responses and target sources in a scene and then chooses the optimal solution. What exactly constitutes an optimal solution is determined by the assumption made. For example, the assumption could be made that subjects are extremely accurate in determining the direction of a target source and randomly guess its distance. In this case, the matching algorithm should try to find the match between responses and

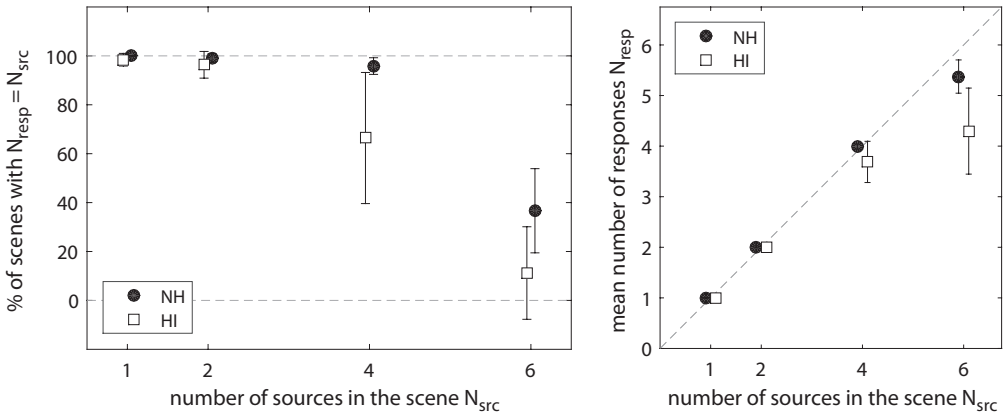


FIGURE 5.4: Performance of the subjects in identifying the number of sources in a scene N_{src} as a function of N_{src} . The left panel shows the percentage of scenes where N_{src} was identified correctly, the right panel shows the mean number of sources identified. Both panels include the group mean of NH listeners (black circles) as well as HI listeners (squares). Error bars represent ± 1 standard deviation from the group mean.

target sources that produces the smallest sum of direction localization errors across all target sources in the scene. However, due to the evenly distributed, symmetrical layout of target source locations, it is possible that there are multiple matches that produce the same minimal error. In this case, the solution can be narrowed down by finding the minimum of a second error criterion, e.g. in the example above for all matches that result in the same minimum direction localization error the optimal solution could be either the one that produces the lowest number of errors in assigning the sex of the talkers or the one that results in the smallest distance localization error.

In a realistic scenario as the one used in the experiment described here, it can be assumed that listeners have different accuracies in localizing the direction and distance of a target source as well as in identifying its sex. Therefore, theoretically, the best way to match responses and target sources is represented by the minimization of the sum of the weighted localization errors in both dimensions and the error in assigning the sex. In practice, however, it would be very difficult to determine how the errors should be weighted. Motivated by results of a pilot study it was assumed here that listeners have a much higher accuracy when localizing the azimuth of a target source than its distance. Hence, for the analysis of the results in the following sections, the responses and target sources of a scene were matched such that the summed azimuth localization error across all sources of the scene was minimized first, before sex and then distance of the talker were considered.

5.3 Results

Figure 5.4 shows the performance in identifying the number of target sources as a function of the number of target sources in the scene (N_{src}). The left panel shows performance as measured by the percentage of scenes where the number of subject responses (N_{resp}) matched N_{src}. NH listeners indicated the correct number of target sources up to $N_{src} = 4$ almost at 100 %, then their performance dropped to 37 % at $N_{src} = 6$. The performance of HI listeners on the other hand already started to break down at $N_{src} = 4$ (66 %) and further decreased to 11 % at $N_{src} = 6$. In both groups, the inter-subject variance was high when performance was substantially lower than 100 %. The right panel of Fig. 5.4 shows performance as measured by the mean N_{resp} as a function of N_{src}. It illustrates that, when the number of responses N_{resp} did not match the number of target sources N_{src}, the number of target sources was generally underestimated. NH listeners gave 5.38 responses on average when $N_{src} = 6$, HI listeners gave 3.68 responses on average when $N_{src} = 4$ and 4.3 responses on average when $N_{src} = 6$.

The effect of the number of sources in a scene on the listeners' localization accuracy is shown in Fig. 5.5. The upper panel displays the root mean square (RMS) error per source of the localization in azimuth as a function of N_{src}, the lower panel displays the mean error per source of the localization in distance as a function of N_{src}. The mean and standard deviation across listeners in the NH group (black symbols) and in the HI group (white symbols) are compared to the results of two groups of simulated listeners. Both simulated groups consisted of 2000 listeners randomly guessing the source locations. The first group of simulated listeners gave the same number of responses per scene as the NH group did on average (as shown in Fig. 5.4, right panel) and the second group of simulated subjects gave the same number of responses per scene as the HI group did on average. The results of these simulations indicate the upper limit for the localization errors that would have been achieved if the listeners in the experiment would have guessed the target source locations completely randomly (i.e the chance level). Figure 5.5 illustrates that localization performance in azimuth was generally worse in the HI group than in the NH group (HI: mean RMS error of 14.4° across all N_{src}, NH: mean RMS error of 5.6° across all N_{src}) and that localization errors (in azimuth) increased with increasing N_{src}. NH listeners were very accurate in localizing the azimuth of a target source if the scene contained only one or two target sources (2° RMS error) whereas HI listeners already made an average RMS error of 9° when there was only one target source in the scene. When $N_{src} = 6$, the mean RMS error in the NH group was 10.6° and 18.9° in the HI group, which means that all listeners still performed much better than chance level.

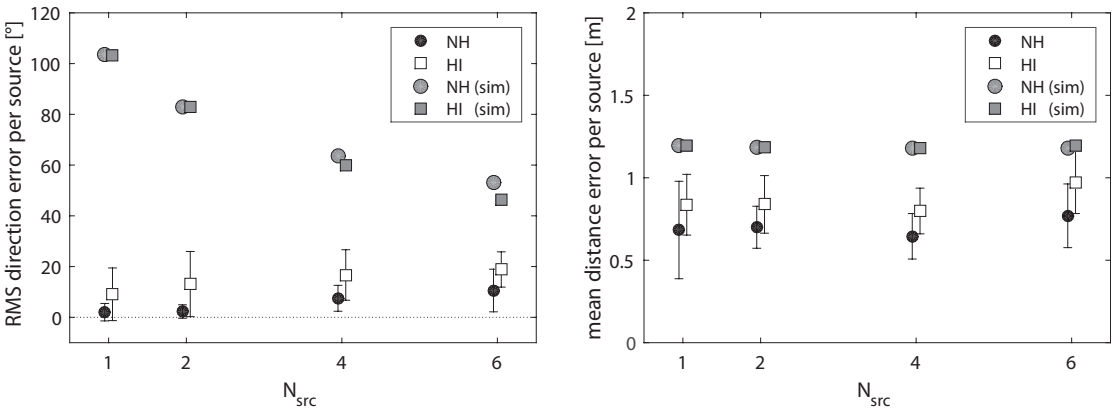


FIGURE 5.5: Localization performance as a function of the number of sources in a scene N_{src}. The left panel shows the mean RMS azimuth localization error per subject response, the right panel shows the mean distance localization error per subject response. Gray symbols indicate chance level.

As opposed to the localization performance in azimuth, the localization performance in distance did not change substantially as a function of the number of target sources present in a scene. The mean distance localization error was slightly increased in both groups for $N_{src} = 6$ but approximately constant across all other values. Overall, HI listeners performed slightly worse (mean error: 0.86 m) than NH listeners (mean error: 0.7 m). It should be noted here that the analysis method is likely to overestimate the distance localization error (see section 5.2.5), however, measured performance was still better than chance level in both groups.

As can be seen in Figures 5.4 and 5.5 the localization performance varied substantially between listeners, especially in the HI group. To analyze the contribution of individual differences in hearing impairment to that variation in the HI group, listeners were divided into three subgroups according to their 4-FAHL. Group 1 consisted of five listeners with a 4-FAHL of less than 20 dB, group 2 consisted of eight listeners with a 4-FAHL of between 20 and 40 dB and group 3 consisted of three subjects with a 4-FAHL of more than 40 dB. Then, the mean RMS azimuth localization error was evaluated within these three subgroups as a function of N_{src}. The results are shown in Fig. 5.6. For all values of N_{src}, the localization error increased with increasing 4-FAHL, however, the difference between the three groups decreased for higher values of N_{src} and was very small at $N_{src} = 6$.

Figure 5.7 shows the localization performance in both azimuth (upper panel) and distance (lower panel) as a function of 4-FAHL for each listener averaged across all responses. A linear regression analysis revealed that there was a weak and marginally significant correlation between localization performance in azimuth and 4-FAHL ($R^2 = 0.27$, $p = 0.039$), but no significant correlation between 4-FAHL and localization performance in distance.

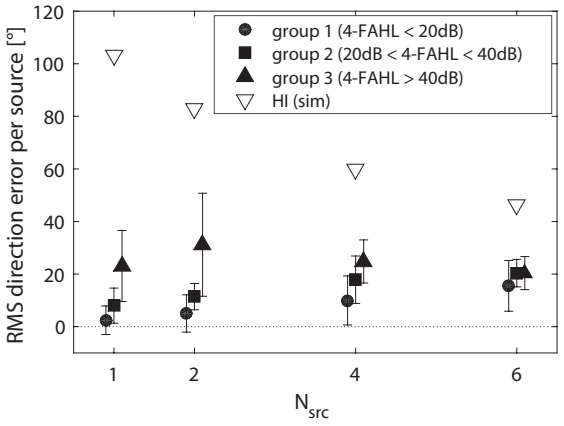


FIGURE 5.6: Localization performance of HI listeners divided into three groups according to their 4-FAHL as measured by the mean RMS azimuth localization error per subject response as a function of the number of sources in a scene N_{src}.

5.4 Discussion

5.4.1 General observations

Spatial awareness, as measured using this new test, appears to be affected by hearing impairment and the complexity of the auditory scene. Even with normal hearing, listeners were not able to correctly analyze an auditory scene when there were six concurrent talkers. Additionally, the localization accuracy decreased with increasing number of concurrent talkers in the scene. Listeners with a hearing impairment generally showed the same trends but performed overall poorer than their normal hearing peers. Their ability to correctly analyze an auditory scene already decreased substantially when the scene contained four concurrent talkers and the RMS azimuth localization error in the HI group was between 7° and 11° higher than in the NH group. While decreased target

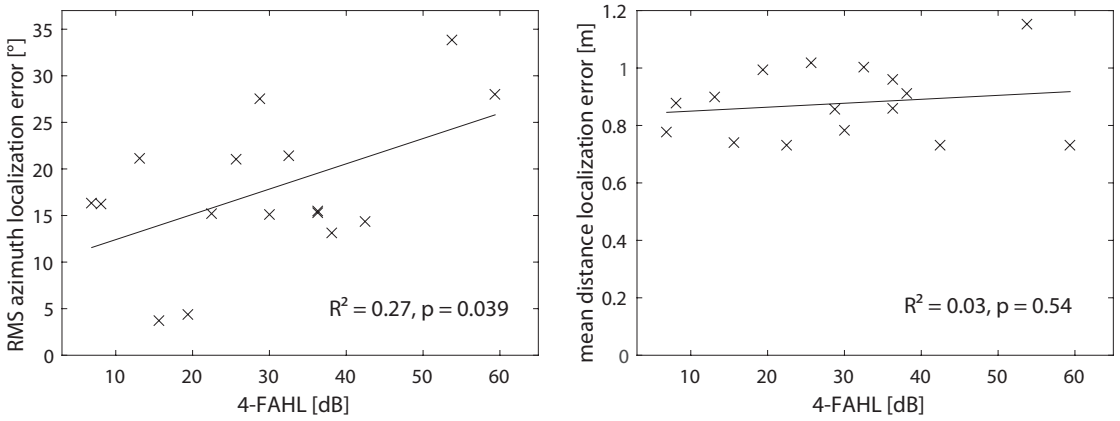


FIGURE 5.7: Scatter plots of the localization performance of subjects as a function of the 4-FAHL. Mean RMS azimuth localization error over all responses (left panel) and mean distance localization error over all responses (right panel).

audibility is likely to have affected the HI listeners' ability to detect all sources in a scene, the localization error was only calculated over the number of responses given, i.e. only for target sources that were detected by the listener. Nevertheless, audibility could still have affected localization performance by reducing the sensation level of the stimuli and the salience of the spatial cues.

Within the HI group, hearing loss proved to be a factor in localizing the direction of a target source. The RMS localization error across all scenes and target sources was significantly correlated with the 4-FAHL. When the scenes were examined separately according to the number of target sources in the scene, it could be observed that hearing loss had the largest effect for scenes containing few sources and was very similar across listeners for $N_{src} = 6$.

The localization performance in distance did not change substantially when the number of source in a scene, N_{src}, was increased and the results in this regard are overall inconclusive. There are several potential reasons for this behavior of the results: First, accuracy in distance perception is generally poor (Gardner, 1969; Loomis *et al.*, 1998). Second, listeners rely to a large extent on the direct-to-reverberant ratio as a cue for distance localization (Bronkhorst and Houtgast, 1999; Zahorik, 2002). In a multi-talker scenario, the interference from other talkers may make it harder to use the direct-to-reverberant ratio of a particular source as an auditory cue. Third, the method that was used to match responses and target sources prioritized the minimization of the localization error in azimuth. Only when there were multiple combinations with the same minimal azimuth localization error was the distance error minimized as well. Hence, the resulting localization error probably tends to overestimate performance of azimuth localization and underestimate performance in distance localization. Fourth, some subjects reported that it was easier to assign a distance to a sound source when multiple sound sources were present in the scene, allowing to use the relative perceived distance between different sources as a cue. This might have (at least partly) compensated for a reduced accuracy with increasing number of target sources. Finally, the design of the experiment did not allow for a very sophisticated analysis of the localization error in distance. There were only four levels of distance and the possible maximum error depended on the source distance: at 1 and 4 m target distance the maximum error was 3 m while at 2 and 3 m target distance the maximum error was 2 m. Although the applied methods were not well suited to measure accuracy in distance perception, including distance as a property of the sound sources also served the purpose of a second error cue and thereby improved the accuracy of the analysis method by resolving ambiguities in the assignment of responses (see also section 5.2.5).

Limiting the time to analyze the scenes to 45 seconds appears not to have affected the subjects' performance. All subjects reported that they did not run out of time in any of the scenes. However, especially for the scenes containing six target sources, subjects also made use of the entire 45 seconds to make their responses.

5.4.2 Analysis method

In the study of Brungart *et al.* (2014) sound sources were realized using different, distinguishable environmental sounds such as a ringing phone or a snoring person, providing the opportunity for subjects to identify the object they were localizing. In contrast, auditory scenes in the current study were composed of competing talkers that were all unfamiliar to the subjects and hence not uniquely identifiable. This made it difficult to analyze the results regarding localization errors as already mentioned in section 5.2.5. Subject responses and target sources were matched by evaluating all possible combinations between all responses and all sources in a scene and choosing the one that resulted in the smallest overall azimuth localization error.

To illustrate the principle effect of the analysis method on the measured localization performance, Fig. 5.8 shows a plot of both the azimuth and distance localization error for the hypothetical case when responses and target sources are matched by minimizing the distance localization error first instead of the azimuth localization error. For $N_{src} = 1$ the results are identical to those presented in Fig. 5.5, because here the matching method has no influence on the results. For a larger number of target sources the azimuth localization error (Fig. 5.8, left panel) increases at a highly unrealistic rate (compared with Fig. 5.5, left panel) and even approaches chance level. The overall distance localization error (Fig. 5.8, right panel) is now reduced and shows a slight improvement when moving from 1 to 2 target talkers instead of a decrease (compared to Fig. 5.5, right panel). The latter aspect is in agreement with the feedback provided by some of the listeners that the presence of a second talker provided an additional relative-distance cue which made the distance judgement easier. In Fig. 5.5, this potentially important detail was hidden by the analysis method that overly focused on minimizing the azimuth localization error. However, given that the azimuth localization error is substantially lower than the distance localization error[2], it seems much more appropriate to focus on the azimuth error in the analysis (as realized throughout this study and described in section 5.2.5) and thereby accept that as a consequence the estimate of the distance localization error

[2]Here it is assumed that the accuracy (or weighting) of a given measure is proportional to the ratio between the error due to guessing and the measured (or expected) error. Using the corresponding localization errors for normal hearing subjects shown in Fig. 5.5 for a single talker (which is independent of the analysis method), this would result in a ratio of $103.7°/2° = 51.85$ for the azimuth and $1.19\,\text{m}/0.95\,\text{m} = 1.25$ for the distance localization error.

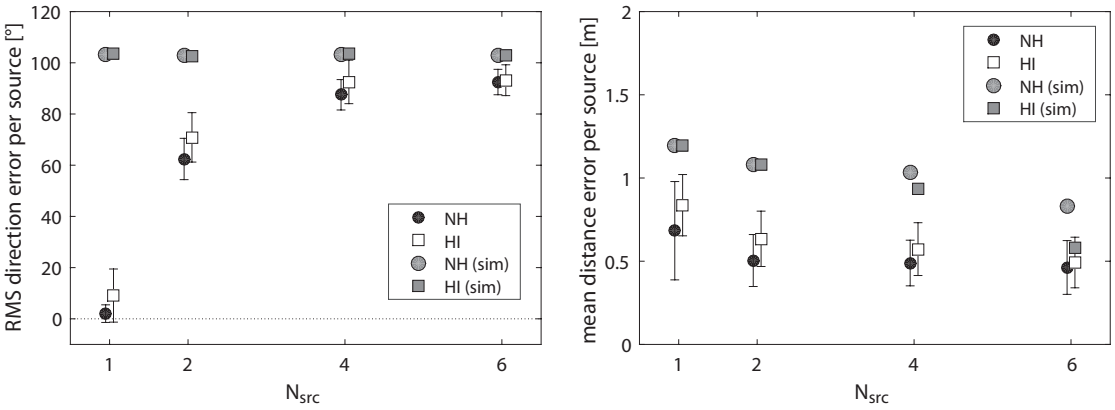


FIGURE 5.8: The mean RMS azimuth localization error per subject response as a function of the number of sources in a scene N_{src} (compare Fig. 5.4, left panel) and the mean distance localization error per subject response (right panel) as a function of the number of sources in a scene N_{src} (compare Fig. 5.4) when subjects responses and target sources are matched by minimizing the distance localization error.

is unreliable. To avoid the need for such source mapping and the errors it inevitably introduces, future studies could use learned or familiar voices to allow test subjects to assign localization responses to the intended source.

5.4.3 Front-back confusions and head movement

While interaural difference cues are useful for localizing a sound source in the left/right dimension, they are ambiguous with respect to resolving if a sound source is in the back or the front of the listener, leading to front-back confusions. In localization experiments, these front-back confusions are normally either avoided by presenting sources only from loudspeakers in the frontal hemisphere (Lorenzi *et al.*, 1999a,b; Kopčo *et al.*, 2010; Best *et al.*, 2011) or accounted for by mapping locations in the rear to the front (Good and Gilkey, 1996; Keidser *et al.*, 2006). In the current study, such a mapping is not possible because it potentially changes the spatial layout of the scene if the scene contains multiple target sources.

However, in the current experiment, listeners were encouraged to rotate their head while assessing the auditory scene to make the task more natural. By observing the listeners during the experiment via a video camera that was installed in the anechoic chamber, it was noted that listeners generally made good use of head movements while completing the task. Head movements have been shown to drastically reduce the probability of front-back confusions in localization tasks (e.g. Wightman and Kistler, 1999 and Bronkhorst, 1995) and thus, it is very unlikely that front-back confusions significantly affected the measured results. This was confirmed by a manual review of the single-source localization performance in the current experiment, which found no front-back confusions at all.

However, occurrences of front-back confusions in scenes with a larger number of sources cannot be ruled out completely.

5.5 Conclusions

In this study, a novel method to measure the spatial hearing ability of NH and HI listeners in a reverberant multi-talker environment was presented. Between $1-6$ talkers were presented simultaneously to the listener for a maximum of 45 seconds, and the task was to identify the sex of all talkers as well as their location in both direction and distance. This localization task is very different from the more common approach where localization of a single brief target signal is measured either in quiet or in (ongoing) background noise. It was assumed that the localization performance measured by such task would better predict the spatial hearing ability experienced by the listeners in a real-life cocktail-party environment. However, applying multiple talkers as targets had the disadvantage that there were no signal properties that allowed the unique identification of the different talkers. For scenes with more than one talker, this made the analysis of the localization results difficult, since no direct one-to-one mapping was possible between the identified (or localized) sound source (talkers) and the actual sources presented in the considered scene. To minimize the potential error introduced by this ambiguity, an analysis method was proposed that minimized the overall error in direction-localization over all identified sources in the scene. In the case that multiple solutions with the same overall error were found, the identified sex and then distance were additionally considered. This method resulted in a reliable measure of the subject's individual localization performance in direction, but at the same time resulted in a localization performance in distance that remained rather unreliable and inconclusive. Furthermore, this method makes it difficult to detect front-back confusions, however, listeners in the presented experiment were encouraged to turn their head which made the occurrence of front-back confusions very unlikely.

Future studies may avoid the disadvantages caused by missing source identity for example by using learned voices. Other studies have used easily distinguishable environmental sounds (e.g. Brungart *et al.*, 2014), but it is difficult to draw conclusions from the results of such studies on localization in multi-talker scenarios, which has often been linked to the ability of listeners to follow conversations (e.g. Hirsh, 1950).

The results of the questionnaire-based study of Noble and Gatehouse (2004) indicated that the ability of the subjects to follow a conversation in a complex environment may also be affected by the movement of sound sources. Future studies assessing spatial hearing in challenging situations should therefore aim to incorporate moving sources

into the sound scenes. This can be achieved for example by using microphone array recording techniques such as Mixed-Order Ambisonics (e.g. Oreinos, 2015).

Chapter 6

Conclusions

This thesis presented a series of three studies analyzing audibility in realistic environments and one study presenting a novel approach to measuring aspects of spatial hearing in realistic environments. The first of the three studies on audibility, presented in Chapter 2, measured the detection of a short speech target in a reverberant multi-talker environment in a group of NH listeners. Masked thresholds were measured adaptively from 15 target locations at different distances and directions in a simulated room. The results showed that in the chosen environment, target audibility depended heavily on both, location and distance of the target to the listener. Across all 15 target locations, the inter-subject mean of the masked thresholds varied in a range of almost 12 dB. This highlights the importance of controlling target audibility in localization tasks. To examine the effects that are responsible for the large spread of masked thresholds across target locations, a functional auditory model proposed by Hant and Alwan (2003) was first extended so that it was applicable to binaural stimuli and then used to predict the masked thresholds measured in the experiment. Despite its relatively simple structure and that no changes to the original model parameters were applied, the model was able to predict the masked thresholds with good accuracy, resulting in an RMS error of 1.4 dB between model predictions and experimental results. An analysis of the decision device of the model showed that for a close target in front of the listener at threshold SNR, most of the detectable signal energy was located in the frequency region above 4 kHz. Due to head shadow effects and the influence of reverberation, detectable energy shifted to lower frequencies at target locations farther away and in other directions. It is suggested that these are the main mechanisms responsible for the variation of masked thresholds across target locations. A practical side effect of the high accuracy of the model is that it can potentially be used to control target audibility in localization experiments also for other speech stimuli and thereby replace time consuming individual measurements of masked thresholds at different target locations.

The study presented in Chapter 3 repeated the previous experiment with a group of HI listeners with different audiometric configurations. The results corresponded very well with the conclusions drawn from the model analysis of the NH experiment. Apart

from an overall decreased sensitivity (i.e. an increase in thresholds), masked thresholds in the HI group also varied less over target location. This may be explained by high frequency cues, which were largely responsible for the variation of masked thresholds over target location in the NH group, being unavailable to the HI subjects due to the increased hearing loss at higher frequencies in the more common sloping audiograms. The HI subject with the lowest high frequency loss (44 dB between 2 and 6 kHz) had masked thresholds that were very similar to the mean thresholds in the NH group. In the auditory detection model, hearing loss was accounted for by adding a threshold simulating noise to the stimuli at the input of the model. Using this approach, the model predicted the average sensitivity in the HI group with good accuracy, but under-predicted individual differences between HI subjects and also under-predicted the variation of the masked thresholds over target location. A comparison of the influence of the masker stimulus and the threshold simulating noise showed that the threshold simulating noise dominated the overall masking signal at frequencies above 1 kHz and suppressed any useful target energy at those frequencies. Due to this low-frequency dominance, the head shadow had only a small effect on the predicted masked thresholds. Consequently, the model predictions across subjects mainly depended on their low-frequency hearing loss, where the smallest differences between subjects were observed. This behaviour of the model needs to be improved to better predict the experimental data.

Chapter 4 examined the effect that uncertainty about the target location (spatial uncertainty), as naturally present in a localization task, has on the masked thresholds in both NH and HI listeners. This was motivated by results from past studies which had reported a significant effect of spatial uncertainty on the outcome of different (mostly speech-related) psychoacoustic metrics. For this study, the environment used in the previous two studies was recreated and all subjects participating in this experiment had also participated in the previous experiment. Using the individual masked thresholds as reference values, psychometric functions were measured for the speech target in a reverberant multi-talker background from eight target directions evenly distributed on a circle with a radius of 2 m around the listener. The measurement was performed in two conditions: A "fixed-direction" condition, where the target direction was fixed in a block of trials and indicated to the subject, and a "random-direction" condition, where the target direction was randomly varied on a trial-to-trial basis and unknown to the subject. An analysis of the psychometric functions with respect to their slope and the threshold showed that the slopes were not affected by the introduction of spatial uncertainty while the thresholds were shifted upwards by 0.8 dB on average. A statistical analysis confirmed that this shift was significant (albeit marginally). In the context of controlling audibility in localization experiments this leads to the conclusion that measurements or model predictions of the masked thresholds are sufficient and the added

effect of spatial uncertainty is negligible.

Localization also plays an important role in dynamic multi-talker conversations, where it helps to focus attention on the active talker. This connection cannot be fully captured in single-source localization tasks. Therefore, a novel method to measure aspects of spatial hearing was presented in Chapter 5. In this paradigm, ten NH subjects and 16 HI subjects had to listen to 96 different acoustic scenes of 45 seconds duration, count the number of talkers in the scene and locate them in direction and distance. This experiment was significantly different from standard localization tests in that the stimuli were continuously active over the span of 45 seconds and multiple sources had to be localized at the same time. The results showed that NH listeners had difficulties to identify the right number of talkers when there were six talkers in a scene whereas performance in the HI group already suffered when there were four talkers in a scene. The RMS direction localization error increased with increasing number of talkers and, on average, was between $2° - 10.5°$ in the NH group and between $7° - 11°$ higher than that in the HI group. The results of the analysis of the localization error in distance remained generally inconclusive due to several reasons: First, the design of the scenes did not allow for a very precise measurement of distance perception. The second reason is connected to a more general aspect of this type of study. While each scene could contain up to six target sources, these sources were all talkers (male or female) that were not uniquely identifiable by the listeners. This makes it impossible to find the true match between subject responses and target sources and independent of the method that is used to match responses and sources, some kind of error will be made. In the study presented here it was decided to match responses and sources by minimizing the direction localization error in a given scene, which is likely to underestimate the subjects' accuracy in distance perception. Despite this shortcoming it was shown that the applied method to assess spatial hearing was feasible and a similar, improved paradigm may prove to be suitable to test different kinds of hearing aid algorithms under realistic conditions.

6.1 Discussion and outlook

The results presented in Chapters 2 and 3 highlight the importance of controlling target audibility in localization experiments in complex acoustic environments. The large range of 12 dB that the masked thresholds covered (not considering the additional variation between subjects) makes clear that calibrating the stimulus level in localization experiments with respect to a single reference measurement may very well cause the stimulus to be inaudible at certain locations. While the applied auditory detection model may be

a useful tool in the design of localization experiments, it should be applied carefully to the prediction of thresholds in hearing impaired listeners. The approach that was chosen here to account for individual hearing loss was not able to fully predict the variation of masked thresholds over target location and individual HI subjects. For a more accurate prediction of the thresholds, future studies may consider applying auditory models that consider other effects of hearing loss besides reduced sensitivity as measured by the audiogram, such as reduced temporal and spectral resolution, loudness recruitment and possibly also cognitive effects. In the context of controlling target audibility in a complex environment, the measured effect of spatial uncertainty on masked thresholds was negligible compared to the overall variation of masked thresholds between subjects and across target locations. However, the results were obtained using only one specific environment and the effect of spatial uncertainty may be larger in environments with a different spatial configuration of the distracters or a higher amount of reverberation. The latter should also be kept in mind for the results presented in Chapters 2 and 3 and should be further explored in future experiments.

Despite the obvious disadvantages caused by the design of the experiment (i.e. the chosen stimuli), the multi-talker localization task presented in Chapter 5 produced sensible results for the localization in the horizontal plane. As expected, HI listeners performed substantially worse than NH listeners and a correlation between hearing loss and localization performance was observed. This is promising for the future development of this kind of test. One possible improvement to such a test could be made by letting the subjects learn a number of target voices so that responses can be uniquely assigned to target sources in a multi-source localization task. It is debatable if a denser layout of target locations would lead to a more precise measurement of localization. In the study presented here, the target directions were chosen to coincide with the loudspeakers in the test facility. This was motivated by the fact that listeners tend to be biased towards loudspeakers even if they are given the opportunity to use target locations between those. Covering the loudspeakers with an acoustically transparent curtain would resolve this issue, but does not necessarily improve the accuracy of the applied localization method.

As Noble and Gatehouse (2004) pointed out, movement of sound sources seems to have a deteriorating effect on the ability of listeners to spatially analyze a sound scene. Future studies should therefore aim to incorporate moving sound sources in tests that assess spatial hearing. This can for example be achieved using microphone array recording techniques like Mixed-Order Ambisonics (MOA). Microphone array recordings can be used to either record an entire scene that can be used in a listening test or to record the background noise of an acoustic environment and then later add sound sources to the scene using impulse response measurements. Due to this flexibility, MOA seems to be a promising approach.

Appendix A

Frequency dependent regularization of a mixed-order Ambisonics encoding system using psychoacoustically motivated metrics[1]

In mixed-order Ambisonics (MOA) encoding the right choice of the regularization parameter λ is crucial to the reproduction quality of the system. The choice of λ is a trade-off between low microphone noise amplification and high directivity. Commonly used performance metrics are not suited to determine an optimum $\lambda(f)$-curve due to their limited psychoacoustical relevance. Therefore, it is proposed here to use two perceptually motivated metrics that consider the reproduced signals at the listener's ear to quantify the reproduction error and the audibility of the total microphone noise. Regularization was then optimized towards previously defined requirements on both of these metrics separately. The results showed that at frequencies below 500 Hz both requirements can be met. Above this frequency one or both of the requirements have to be relaxed in order to find a suitable $\lambda(f)$.

[1] Chapter published as:
Weller, T., Buchholz, J. M. and Oreinos, C. (**2014**), "Frequency dependent regularization of a mixed-order Ambisonics encoding system using psychoacoustically motivated metrics", in *Proceedings of the Audio Engineering Society 55th Conference on Spatial Audio*

A.1 Introduction

Interest in the application of virtual sound environments (VSEs) has increased significantly over the last 15 years. VSEs offer the possibility to objectively and subjectively measure the performance of devices (e.g. hearing aids) in complex acoustic scenarios in a reproducible manner. They also open up opportunities to design previously impossible psychoacoustic paradigms to assess problems of (aided) hearing impaired subjects in everyday situations. In a VSE, subjects are put into a natural listening situation and are able to freely turn their head.

One method used to establish VSEs is higher-order Ambisonics (HOA). The advantage of HOA over other soundfield reproduction techniques is that HOA is easily scalable and very flexible regarding different layouts of microphone and/or loudspeaker arrays. HOA uses a spherical harmonic decomposition of the soundfield and can be implemented either as a planar 2D system or a periphonic 3D system. However, to achieve the same accuracy (i.e. the correct reproduction of a given soundfield, which is directly related to the spatial resolution or directivity of the reproduction system) the number of microphones and loudspeakers needed in a 3D system is much higher than in a 2D system. To find a compromise, mixed-order Ambisonics (MOA) systems have been applied as a combination of high order 2D Ambisonics and low order 3D Ambisonics Travis (2009); Favrot *et al.* (2011). A MOA system provides high directivity of sound sources close to the horizontal plane while still offering the possibility of periphonic recording and playback with a lower directivity.

Different metrics have been applied to assess the performance of microphone arrays used for MOA encoding Favrot and Marschall (2012). These include metrics generally applied to beamformers, such as the white noise gain (WNG) and the directivity index (DI), and the energy vector, a metric introduced to evaluate Ambisonics systems Gerzon (1992). All of these metrics are well understood and useful to compare different systems to each other. It is, however, not straightforward to define goal functions for optimization using these metrics or to use these metrics to judge the quality of a MOA system used for psychoacoustic experiments. Oreinos and Buchholz (2014) recently applied metrics more closely related to psychoacoustics.

Ambisonics encoding is an ill-posed problem that requires regularization. The most common way of regularizing an Ambisonics encoding system is Tikhonov regularization Tikhonov and Arsenin (1977). Tikhonov regularization is controlled by adjusting a regularization parameter λ. The choice of λ is a trade-off between high directivity and low noise amplification, especially at low frequencies. In most cases (e.g. Favrot and Marschall (2012), Moreau *et al.* (2006)) λ is chosen by trial-and-error and it is usually

chosen to be constant over frequency. It is, however, possible to vary λ with frequency and thus to optimize λ with respect to a frequency dependent goal function.

Marschall *et al.* (2012) introduced an alternative encoding method called shape matching (SMM), which showed a better performance than the traditional Ambisonics encoding process (AMB) with respect to WNG and DI. They regularized the encoding matrix in this approach by applying Butterworth filters mimicking the shape of the regularization filters used in AMB. This does, however, not completely solve the ill-posed problem in SMM encoding.

In this study we apply Tikhonov regularization to SMM encoding of microphone array signals. The encoding performance is then evaluated using psychoacoustically motivated metrics and the regularization parameter λ is adjusted for each frequency point individually in order to find the optimal solution.

A.2 Background

This section briefly reiterates the principles of Ambisonics. It provides a summary of MOA encoding using AMB and SMM. The nomenclature uses spherical coordinates that describe a point in space using the radius r, azimuth angle θ ($0 \leq \theta < 2\pi$) and elevation angle δ ($-\pi/2 \leq \delta < \pi/2$) with respect to the coordinate origin O.

A.2.1 Mixed-order Ambisonics

The sound pressure on the surface of a rigid sphere of radius R can be expressed as a Fourier-Bessel series Moreau *et al.* (2006)

$$p(kR, \theta, \delta) = \sum_{m=0}^{\infty} W_m(kR) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \delta), \tag{A.1}$$

where B_{mn}^{σ} are the coefficients of the Fourier-Bessel series (or Ambisonics components) and $Y_{mn}^{\sigma}(\theta, \delta)$ are the real-valued spherical harmonic functions (SHFs) given by

$$Y_{mn}^{\sigma}(\theta, \delta) = \sqrt{(2m+1)(2-\delta_{0,n})\frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\delta) \cdot \begin{cases} \cos n\delta & \text{if } \sigma = +1 \\ \sin n\delta & \text{if } \sigma = -1 \text{ (ignored if } n = 0) \end{cases}, \tag{A.2}$$

where P_{mn} are the associated Legendre functions Moreau *et al.* (2006) and $\delta_{0,n} = 1$ if $n = 0$ and 0 otherwise. For a rigid sphere, the radial function $W_m(kR)$ is given by

$$W_m(kR) = i^m \left(j_m(kR) - \frac{j'_m(kR)}{h'_m(kR)} h_m(kR) \right) \tag{A.3}$$

with j_m being the spherical Bessel functions and h_m the spherical Hankel functions of the second kind, and the prime indicating their derivatives with respect to the argument.

Ambisonics encoding is the process of estimating the Ambisonics components B^{σ}_{mn} using the pressure signals from a set of Q sensors distributed on the surface of a rigid sphere. Due to the discrete sampling the Fourier-Bessel series in eq. (A.1) has to be truncated at an order M. The maximum possible value for M is limited by the number of pressure sensors Q according to

$$2M + 1 < Q \qquad \text{(2D)} \tag{A.4}$$

$$(M + 1)^2 < Q \qquad \text{(3D)}. \tag{A.5}$$

It becomes immediately apparent that the number of required sensors for a certain order M is much higher in a 3D system than in a 2D system. Mixed-order Ambisonics was suggested Daniel (2000); Travis (2009) as a method to combine high order (and thus, high directivity) 2D encoding with low order 3D encoding. Favrot *et al.* (2011) then applied MOA to a microphone array and found that it provides improved directivity in the horizontal plane with only few additional microphones. In a MOA system 2D order M_{2D} and a 3D order M_{3D} the overall number of pressure sensors has to satisfy eq. (A.5) for M_{3D}. The number of sensors in the horizontal plane has to satisfy eq. (A.4) for M_{2D}. Due to the discretization and the application of MOA, the pressure on the surface of a rigid sphere would be estimated by

$$p(kR, \theta, \delta) \approx \sum_{m=0}^{M_{3D}} W_m(kR) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B^{\sigma}_{mn} Y^{\sigma}_{mn}(\theta, \delta) + \sum_{m=M_{3D}+1}^{M_{2D}} W_m(kR) \sum_{\sigma=\pm 1} B^{\sigma}_{mm} Y^{\sigma}_{mm}(\theta, \delta). \tag{A.6}$$

Eq. (A.6) in matrix notation is given by

$$\mathbf{p} = \mathbf{Y} \cdot \mathrm{diag}\left[W_m(kr)\right] \cdot \mathbf{b}, \tag{A.7}$$

where $\mathbf{p}$ is a column vector containing Q measured pressure values, $\mathbf{Y}$ is a $Q \times K$ matrix with the q-th row containing K spherical harmonic functions sampled at the

position (θ_q, δ_q) of the q-th microphone, and diag $[W_m(kR)]$ is a diagonal matrix containing the approriate values of the radial function $W_m(kR)$ for the corresponding spherical harmonic function. In MOA the number of spherical harmonic functions is $K = (M_{3D} + 1)^2 + 2(M_{2D} - M_{3D})$. Then, $\mathbf{b}$ is a column vector containing the K Ambisonics components obtained by a least-squares solution of eq. (A.7):

$$\mathbf{b}(f) = \mathbf{E}(f) \cdot \mathbf{p}(f), \tag{A.8}$$

with

$$\mathbf{E}(f) = \text{diag}\left[\frac{1}{W_m(kR)}\right] \cdot \mathbf{Y}^+ \tag{A.9}$$

where $\mathbf{Y}^+$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{Y}$. $\mathbf{E}$ is called the encoding matrix.

A.2.2 Shape matching

As an alternative method to traditional Ambisonics encoding as introduced above (AMB), a strategy called the shape matching method (SMM) was proposed Marschall *et al.* (2012). SMM makes use of the knowledge that for a perfect plane wave hitting the microphone array from angle (θ_l, δ_l) the Ambisonic components are given by

$$B^\sigma_{mn} = Y^\sigma_{mn}(\theta_l, \delta_l). \tag{A.10}$$

This means that encoding a set of L plane waves from incident angles forming a regular grid around the microphone array should result in a $K \times L$ matrix $\mathbf{Y}_L$:

$$\mathbf{E} \cdot \mathbf{P} = \mathbf{Y}_L \tag{A.11}$$

where $\mathbf{E}$ is the $K \times Q$ encoding matrix and $\mathbf{P}$ is a $Q \times L$ matrix containing the pressure values of L plane waves measured at the Q microphones. For every point on the frequency scale of the system, the encoding matrix will then be given by

$$\mathbf{E}(f) = \mathbf{Y}_L \cdot \mathbf{P}^+(f), \tag{A.12}$$

where $\mathbf{Y}_L$ can be computed using eq. (A.2). The avantage of using SMM instead of AMB encoding is that the pressure matrix $\mathbf{P}$ can not only be obtained through simulation using eq. (A.1) but also by measurement using a real microphone array. In the latter case, SMM will additionally perform a calibration for the given microphone array by comparing the measured signals to the response of an ideal system.

A.3 Methods

In order to measure the performance of the MOA encoding stage and to optimize the regularization, an entire MOA processing chain was simulated. This includes the MOA encoding of virtual sound sources picked up by a microphone array, the decoding of the Ambisonics components using a loudspeaker array and the application of head-related transfer functions (HRTFs) to obtain the signals at the ears of a listener located in the loudspeaker array. Due to practical considerations a MOA system with orders $M_{2D} = 11$ and $M_{3D} = 5$ was analyzed.

The choice of the regularization parameter λ is a trade-off between two properties of a MOA system: For a given order, low microphone noise amplification will result in a less directive system and a highly directive system will have high microphone noise amplification. It is therefore necessary to find measures for both of these properties in order to optimize λ for a given microphone array.

Popular metrics to measure the performance of microphone arrays are the white noise gain (WNG) and the directivity index (DI). It is, however, difficult to make statements about the perceptual quality of a MOA system using WNG and DI. Since one of the applications of MOA are psychoacoustical experiments, it is desirable to find metrics that are easier relatable to the perceptual effects of MOA processing.

In the present study we therefore used two psychoacoustically motivated metrics as proposed by Oreinos and Buchholz (2014): (i) The reconstruction error E_{HOA}, which compares the magnitude spectrum of a reproduced virtual sound source with an unprocessed reference and (ii) the audibility of microphone noise after processing by the HOA system.

A.3.1 Microphone array

Both AMB and SMM encoding involve the solution of an overdetermined system using a least-squares approach that constitutes an ill-posed problem. It is therefore advisable to choose the microphone array layout such that the condition number $\kappa(\mathbf{Y})$ of the SHF matrix (see section A.2.1) of the microphone array is as small as possible Favrot and Marschall (2012). In a periphonic HOA system this would be typically achieved by a quasi-regular layout. MOA, however, requires an increased density of microphones around the horizontal plane for the high order 2D part of the system. Therefore, a non-regular layout has to be considered.

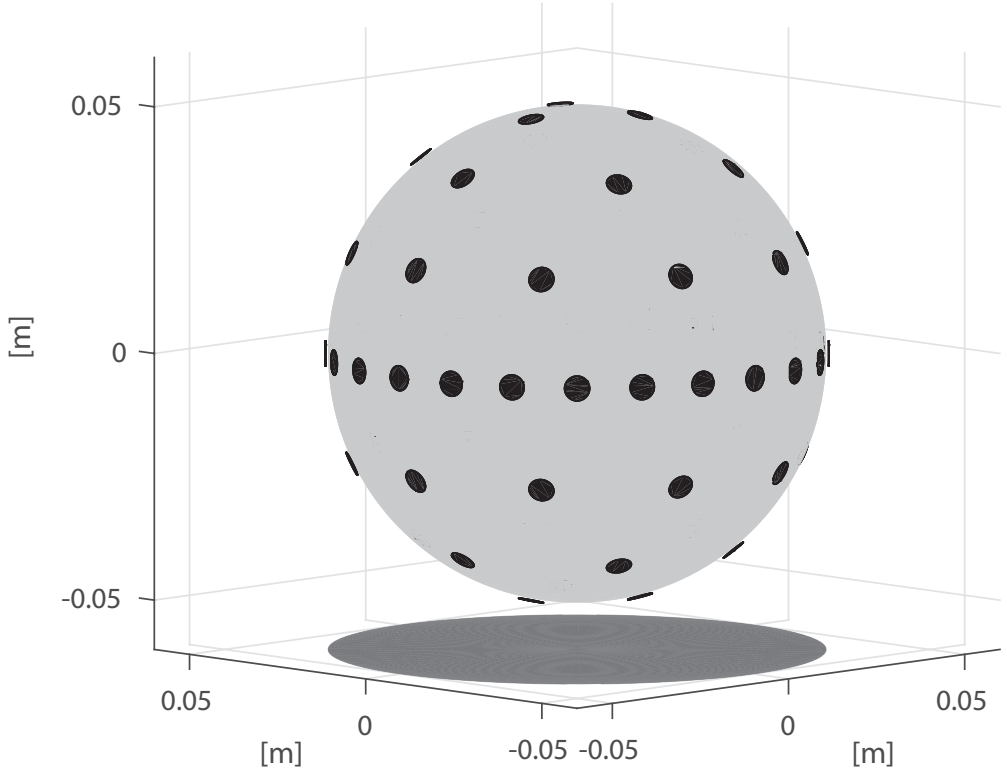


FIGURE A.1: Picture of the microphone array layout. It consists of a ring of 24 microphones on the horizontal plane, two rings with 10 microphones each at $\theta = \pm 25°$, two rings with 6 microphones each at $\theta = \pm 50°$ and two rings with 3 microphones each at $\theta = \pm 75°$, making a total of $Q = 62$ microphones.

To accomodate for this property a layout of multiple horizontal rings at different elevation angles was proposed Favrot and Marschall (2012). An algorithm was provided in the same publication that finds the ring layout with the smallest condition number for a certain combination of M_{2D} and M_{3D}. The final layout used in the simulation is a slightly modified version of the layout obtained using this algorithm with the MOA orders indicated above and is shown in Fig. A.1. This layout consists of $Q = 62$ microphones and leads to a condition number of $\kappa(\mathbf{Y}) = 2.542$. The radius of the microphone array was chosen to be 5 cm.

A.3.2 Regularization

As mentioned before, both AMB and SMM encoding are ill-posed problems. In general, this means that they can become unstable and are sensitive to errors. To avoid excessive microphone noise amplification at low frequencies and high orders the encoding process has to be regularized. The most popular regularization method in Ambisonics is known as Tikhonov regularization.

In both considered approaches to Ambisonics encoding the pseudo-inverse of a matrix has to be found. One method to determine the pseudo-inverse is the singular value decomposition (SVD). The SVD of a $m \times n$ complex matrix $\mathbf{A}$ ($m \geq n$) is given by

$$\mathrm{svd}(\mathbf{A}) = \mathbf{U}\Sigma\mathbf{V}^*, \tag{A.13}$$

where $\mathbf{U}$ is a $m \times m$ complex unitary matrix, $\mathbf{V}^*$ is the conjugate transpose of the $n \times n$ complex unitary matrix $\mathbf{V}$ and Σ is an $m \times n$ diagonal matrix with nonnegative real numbers $\sigma = (\sigma_1, \ldots, \sigma_n)$ on the diagonal. These numbers are called the singular values of matrix $\mathbf{A}$. The pseudo-inverse of matrix $\mathbf{A}$ is then given by

$$\mathbf{A}^+ = \mathbf{V}\Sigma^{-1}\mathbf{U}^*, \tag{A.14}$$

with Σ^{-1} being a $n \times m$ diagonal matrix with values $\sigma^{-1} = (\sigma_1^{-1}, \ldots, \sigma_n^{-1})$ on the diagonal.

In order to limit the effect of very small singular values Tikhonov regularization applies filter factors $\frac{\sigma^2}{\sigma^2+\lambda^2}$ such that

$$\sigma_{\mathrm{reg}}^{-1} = \left(\frac{\sigma_1}{\sigma_1^2 + \lambda^2}, \ldots, \frac{\sigma_n}{\sigma_n^2 + \lambda^2}\right) \tag{A.15}$$

and

$$\mathbf{A}_{\mathrm{reg}}^+ = \mathbf{V}\Sigma_{\mathrm{reg}}^{-1}\mathbf{U}^*. \tag{A.16}$$

The amount of regularization is controlled by the regularization parameter λ. For $\lambda = 0$ there is no regularization at all and then the effect of regularization increases with increasing λ.

In the case of AMB encoding the pseudo-inverse of matrix $\mathbf{Y}$ has to be found. However, the conditioning of matrix $\mathbf{Y}$ can be easily controlled by an appropriate layout of the microphone array. As mentioned before, the layout used in this study leads to a condition number of $\kappa(\mathbf{Y}) = 2.542$ and no regularization is needed. In contrast, the inversion of the filter coefficients $W_m(kR)$ will lead to very high factors at low frequencies and high orders. Regularization is therefore applied to the diagonal matrix diag $\left[\frac{1}{W_m(kR)}\right]$ and the resulting encoding matrix is

$$\mathbf{E}_{\mathrm{reg}}(f) = \mathrm{diag}\left[\frac{W_m^*(kR)}{|W_m(kR)|^2 + \lambda^2}\right] \cdot \mathbf{Y}^+. \tag{A.17}$$

In SMM encoding the pressure matrix $\mathbf{P}$ is inverted. Here, regularization has to be applied as described by eqs. (A.13) - (A.16) and the resulting encoding matrix is

$$\mathbf{E}_{\text{reg}}(f) = \mathbf{Y}_L \cdot \mathbf{P}^{+}_{\text{reg}}(f). \tag{A.18}$$

For all of the presented simulations, the pressure matrix $\mathbf{P}$ was generated evaluating eq. (A.1) for $L = 162$ virtual plane wave sources distributed regularly around the microphone array.

A.3.3 Loudspeaker array

For a perceptual evalutation of the MOA system, a loudspeaker array has to be simulated. First, the Ambisonics components are decoded into loudspeaker signals. HRTFs are then applied to these loudspeaker signals to find an appropriate representation of the signals at the eardrum of a listener inside the loudspeaker array.

The MOA decoding equation is given by

$$\mathbf{s}(f) = \mathbf{D} \cdot \mathbf{b}_{\text{reg}}(f), \tag{A.19}$$

where $\mathbf{s}(f)$ is the vector containing the signals of N loudspeakers at frequency f, $\mathbf{b}_{\text{reg}}(f)$ is the vector containing the regularized Ambisonics components and $\mathbf{D}$ is the $N \times K$ decoding matrix Moreau *et al.* (2006). To limit the errors introduced in the decoding process, a large loudspeaker array with $N = 252$ loudspeakers in a regular layout was simulated.

For each of the N loudspeakers there is a set of two HRTFs describing the signal propagation between the loudspeaker and the eardrum of the left and right ear. The spectrum of the reproduced MOA signal at the eardrums is then determined as

$$H^{l/r}_{\text{MOA}}(f) = \sum_{n=1}^{N} s_n(f) \cdot H^{l/r}_{\text{TF,n}}(f), \tag{A.20}$$

with $H^{l/r}_{\text{TF,n}}(f)$ being the HRTF between the n-th loudspeaker and the left and right ear respectively. The HRTFs were measured using a head and torso simulator with in-ear microphones Oreinos and Buchholz (2013).

A.3.4 Reproduction error

The reproduction error is defined as

$$E_{\mathrm{MOA}}(f) = \frac{|H_{\mathrm{MOA}}(f)|}{|H_{\mathrm{ref}}(f)|}, \tag{A.21}$$

where H_{MOA} is the spectrum resulting from MOA processing and H_{ref} is the reference spectrum. For a virtual plain wave source from direction $(\theta_{\mathrm{vs}}, \delta_{\mathrm{vs}})$ the reference was chosen to be the HRTF measured at that angle:

$$H_{\mathrm{ref}}(f) = H_{\mathrm{TF}}(f, \theta_{\mathrm{vs}}, \delta_{\mathrm{vs}}) \tag{A.22}$$

We defined a virtual source to be reproduced correctly at frequency f if $|E_{\mathrm{MOA}}(f)| \leq 2\,\mathrm{dB}$, i.e. the reproduction error was less than 2 dB. Eq. (A.21) was then evaluated for 400 virtual sources equally distributed around the microphone array and the percentage of correctly reproduced sources $p_c(f)$ was determined.

A.3.5 Audibility of microphone noise

The influence of the regularization parameter λ on the amplification of system inherent microphone noise was analyzed by simulating the equivalent input noise of a typical hearing aid microphone (Buchholz (2013)) as input to all 62 virtual microphones. Assuming that the microphone noise signals are fully uncorrelated the power spectral density (PSD)

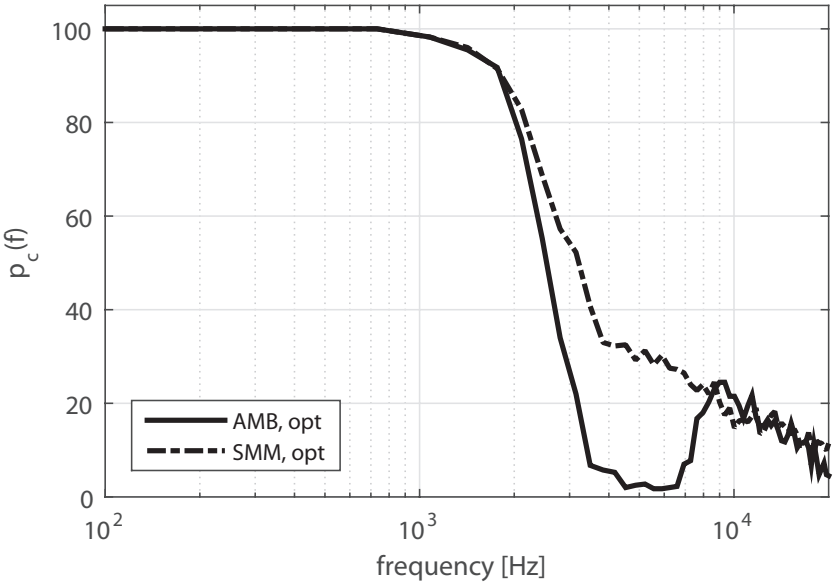
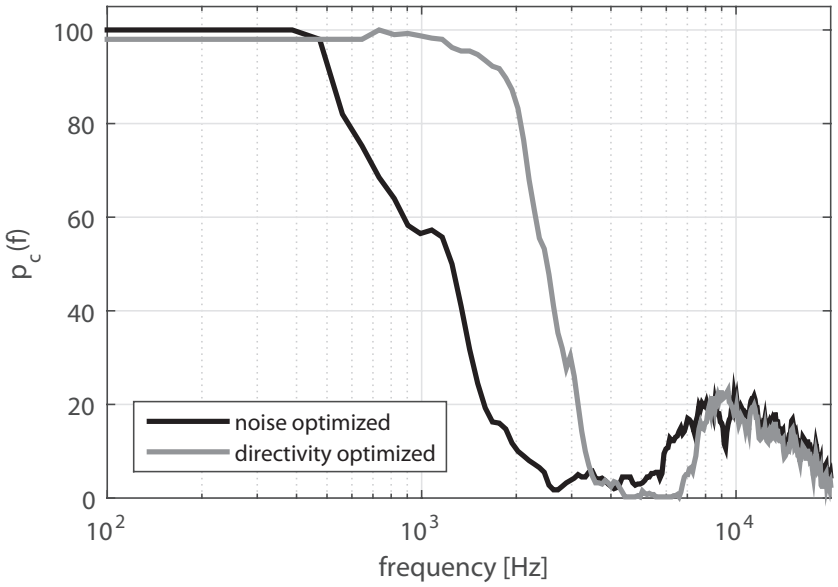
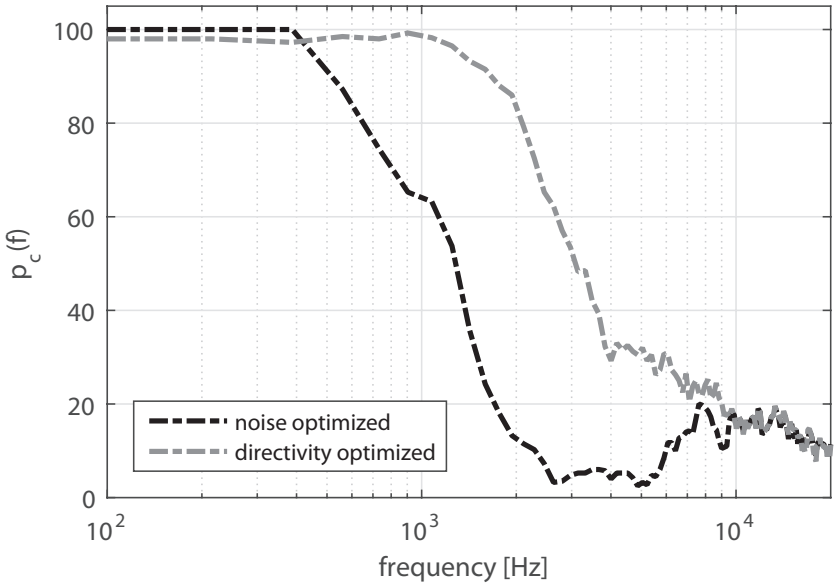


FIGURE A.2: Percentage of correctly reproduced sound sources over frequency for AMB and SMM encoding if no regularization is applied (simulating the microphone array as described in section A.3.1)



(A) AMB encoding



(B) SMM encoding

FIGURE A.3: Percentage of correctly reproduced sound sources. The red curves show the percentage when $\lambda(f)$ is noise-optimized and the blue curves show the percentage when $\lambda(f)$ is directivity-optimized.

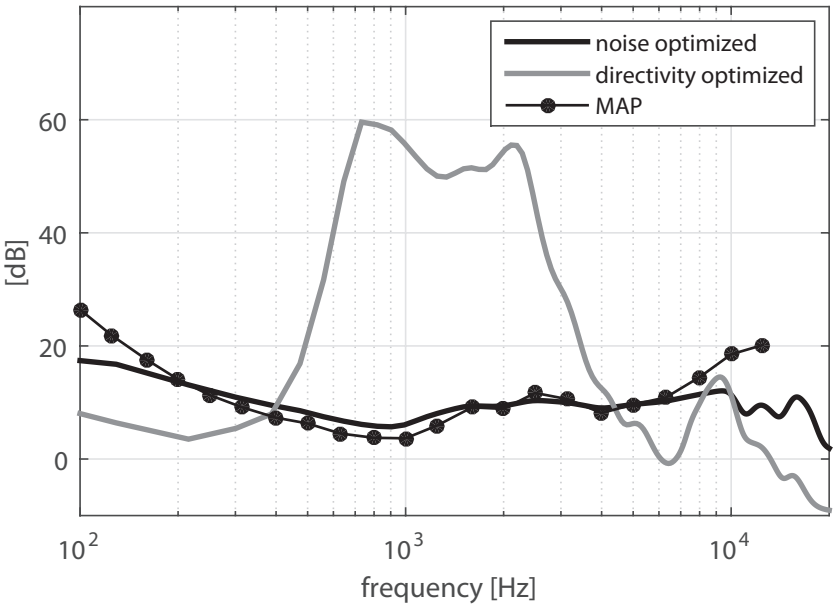
of the noise produced at the eardrum of a listener in the center of the loudspeaker array is then

$$N_{\mathrm{MOA}}(f) = \sum_{q=1}^{Q} |H_{\mathrm{MOA,q}}(f)|^2 \cdot B(f), \tag{A.23}$$

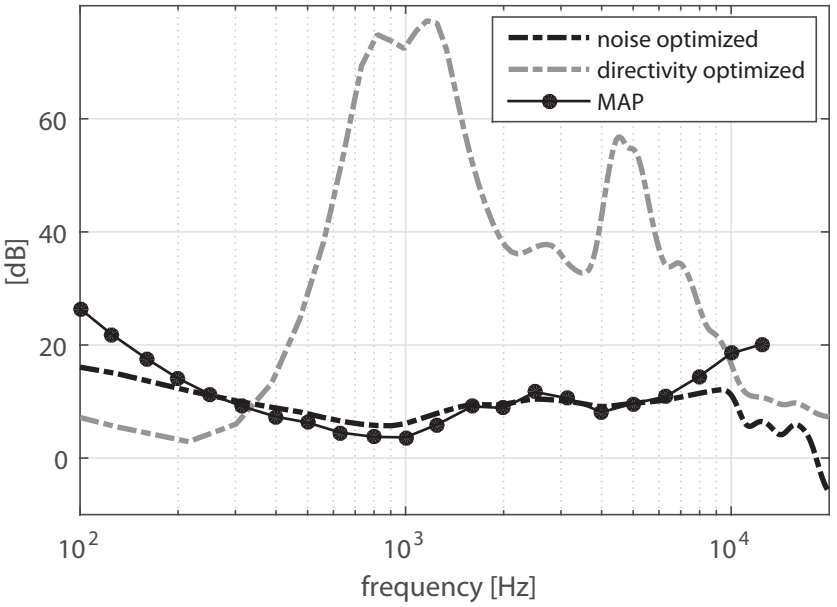
where $H_{\mathrm{MOA,q}}(f)$ is the MOA transfer function between the q-th microphone and the eardrum and $B(f)$ is the PSD of the equivalent input noise after integration in equivalent rectangular bandwidth (ERB) bands Moore (2003).

A.3.6 Optimization

If there is no microphone noise present, the highest achievable percentage of correctly reproduced sources $p_c(f)$ will be theoretically obtained by applying no regularization, i.e. $\lambda = 0$. In practice, the limited numerical range in simulation software introduces an error that makes the system unstable. Thus, the best source reproduction will be achieved by choosing a sufficiently small value for the regularization parameter ($\lambda = 10^{-6}$



(A) AMB encoding



(B) SMM encoding

FIGURE A.4: PSD of the signal at the eardrum produced by microphone noise. The red curves show the noise power when λ is noise-optimized and the blue curves show the noise power when λ is directivity-optimized. The black curve indicates the MAP.

was used). Figure A.2 shows the percentage of correctly reproduced sources in this optimal case for both AMB and SMM encoding. For frequencies above 2 kHz p_c drops significantly for both encoding methods. This initial drop can be attributed to the truncation error caused by the low 3D order of the system. It can be assumed that for frequencies above 2 kHz only sources close to the horizontal plane will be reproduced correctly. This is also indicated by the slow decay of the percentage of correct sources at frequencies above about 4 kHz for SMM in Figure A.2 (dashed line) that follows after the initial drop.

In order to find a regularization scheme that produces high directivity, $\lambda(f)$ was then iteratively increased until $p_c(f)$ was 2 % lower than in the optimal case. In terms of microphone noise amplification, the aim was to adjust the regularization parameter $\lambda(f)$ such that $N_{\mathrm{MOA}}(f)$ matches the minimal audible pressure (MAP) Oreinos and Buchholz (2014) and thus, can be considered just inaudible. The MAP is the pressure of a diffuse noise sound field (mapped to the eardrum) that produces a masking level equivalent to the hearing threshold.

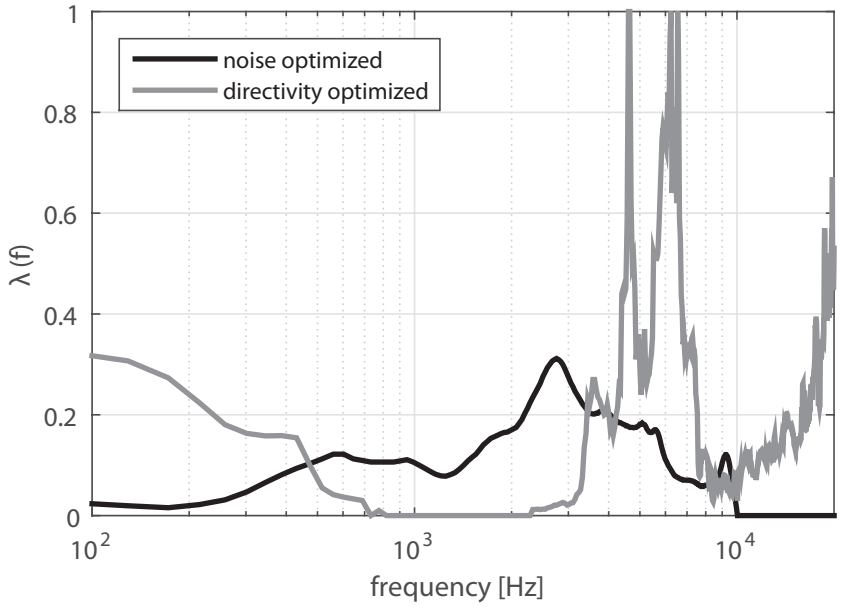
A.4 Results

Figure A.3 shows the percentage of correctly reproduced sources p_c over frequency for the cases where regularization is optimized for high directivity and where regularization is optimized towards inaudible noise for both AMB and SMM encoding. For both encoding methods, optimizing regularization for inaudible noise instead of high directivity shifts the curves significantly towards lower frequencies. The 100 %-cutoff is shifted by 1/2 octave, the 90 %-point is shifted by almost 2 octaves and the 60 %-point is shifted by about 1.5 octaves.

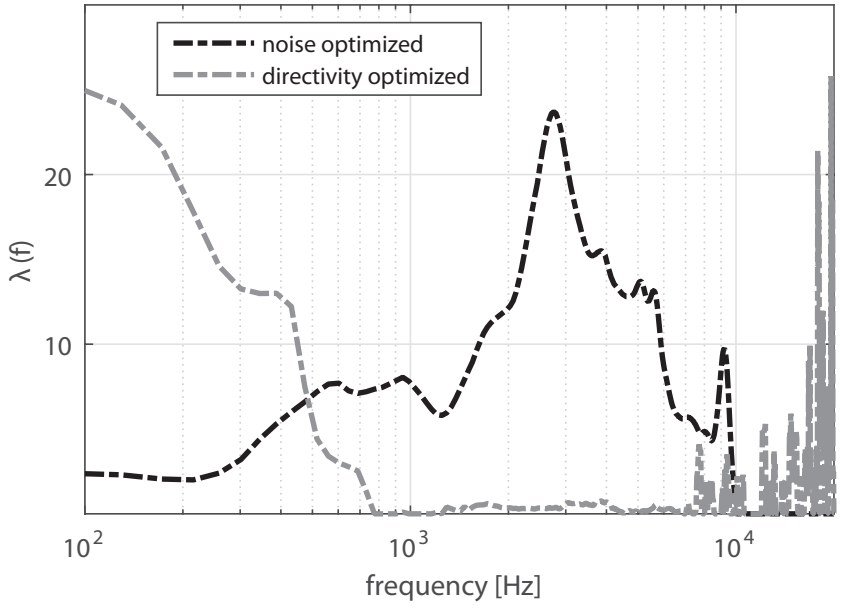
Figure A.4 shows the PSD of the signal at the eardrum of a listener as the result of inherent microphone noise for both optimization cases and both encoding methods. If regularization is optimized for low noise, it is possible to match the MAP curve reasonably well with both encoding methods. Optimizing regularization for high directivity results in a much higher noise PSD between 0.4 kHz and 4 kHz (AMB) and between 0.35 kHz and 10 kHz (SMM).

Figure A.5 shows the resulting λ-characteristics after optimization. Up to 0.5 kHz the maximum regularization parameter allowed to still achieve correct reproduction of the sources is higher than the minimum regularization parameter needed to make the microphone noise inaudible. This means that both requirements can be satisfied. Above this frequency however, the upper threshold for $\lambda(f)$ allowed for correct reproduction quickly

drops, while the lower threshold for $\lambda(f)$ needed to satisfy the noise amplification requirement increases. In this frequency region, a compromise between noise amplification and directivity of the system has to be made as not both requirements can be satisfied at the same time.



(A) AMB encoding



(B) SMM encoding

FIGURE A.5: Resulting $\lambda(f)$-characteristics after optimization. The red curves show $\lambda(f)$ optimized for low noise and the blue curves show $\lambda(f)$ optimized for high directivity. Note the different range of λ values, which is caused by the different encoding methods used.

A.5 Discussion

In this study, a mixed-order Ambisonics system was simulated with focus on two different encoding methods and their regularization. To account for the trade-off between microphone noise amplification and directivity inherent to all Ambisonics systems, two different metrics were used to measure the performance of the system. For the first metric, the measured microphone noise of a commercially available hearing aid microphone was simulated as the input to all microphones. The resulting signal after MOA processing was then evaluated in terms of audibility by comparing it to the minimal audible pressure. In order to evaluate the directivity of the system, the reproduction error of 400 plane wave sources was calculated and the percentage of correctly reproduced sources was measured.

Regularization was then adjusted for both encoding methods such that previously defined requirements were met. This was done for both metrics seperately. An optimization with respect to the audibility of microphone noise was carried out by matching the signal produced by MOA processing of the microphone noise to the MAP curve. Optimization with respect to the directivity of the system was realized by defining a deviation of 2 % from the theoretical maximum of correctly reproduced as the goal function.

The results show that at low frequencies ($f < 500\,\mathrm{Hz}$) both requirements can be satisfied by a suitable choice of the regularization parameter $\lambda(f)$. Above this frequency however, both requirements cannot be met at the same time and the choice of $\lambda(f)$ only determines the compromise that has to be made. This means that one or both of the requirements have to be relaxed. However, the microphone noise might be masked by the recorded (desired) signal and thus elevating the minimal audible pressure when the desired signal is present. In quiet intervals it could be possible to estimate the microphone noise and apply noise reduction techniques to render the microphone noise inaudible.

Figure A.2 gives a good estimate of the frequency at which the truncation error of the lower order 3D system starts to affect the reproduction quality (between 1 and 2 kHz). Above this frequency the choice of $\lambda(f)$ should only be determined by the performance of the higher order 2D part of the MOA system. At frequencies where the truncation error of the 2D system effects the reproduction quality, the choice of $\lambda(f)$ will have little influence on the directivity of the system and regularization should be determined by microphone noise amplification.

Even though both metrics used in this study are perceptually motivated and better suited to assess the quality of a system used for psychoacoustical experiments, it is still difficult to draw a conclusion about the actual perceived sound quality in the real system. It

would therefore be useful to carry out a subjective evaluation, which includes a thorough analysis of the system in the time domain. This could be realized by either simulating the MOA encoding stage or by using recordings from a real microphone array.

Appendix B

Validation of the loudspeaker-based room auralization system (LoRA)

A large part of the work presented in this thesis relies on the use of a loudspeaker-based virtual acoustic environment (VAE) that was realized using the loudspeaker-based room auralization (*LoRA*) toolbox for MATLAB. Since a validation of this VAE in the context of the specific experiments presented here was beyond the scope of this work, this appendix serves to summarize previous work validating the use of the *LoRA* toolbox in auditory research.

B.1 Favrot and Buchholz (2010)

The *LoRA* toolbox was first proposed by Favrot and Buchholz (2010). They evaluated *LoRA* objectively by comparing different room acoustic parameters extracted from the room impulse responses (RIRs) in a virtual loudspeaker array simulated with *LoRA* to the same parameters extracted from the RIRs provided by the room acoustic simulation software *Odeon*. Figure B.1 shows the error of the speech transmission index (STI) introduced by the *LoRA* signal processing measured at different distances from the optimal listening position in the center of the simulated loudspeaker array for two different simulated environments: a moderately reverberant classroom (left) and a highly reverberant concert hall (right). A comparison with the head movement limen (shaded area, representing the error range caused by head movement) shows that variations introduced by the *LoRA* signal processing were in a similar range to variations introduced by small location changes in the simulated room. Other monaural parameters considered by Favrot and Buchholz (2010) were the reverberation time T_{30}, the early decay time (EDT), the clarity C_{80} and the strength G, all of which yielded similar errors as the STI. An analysis of the interaural cross correlation coefficient (IACC) also showed that the error introduced by *LoRA* was comparatively small but that the IACC was consistently

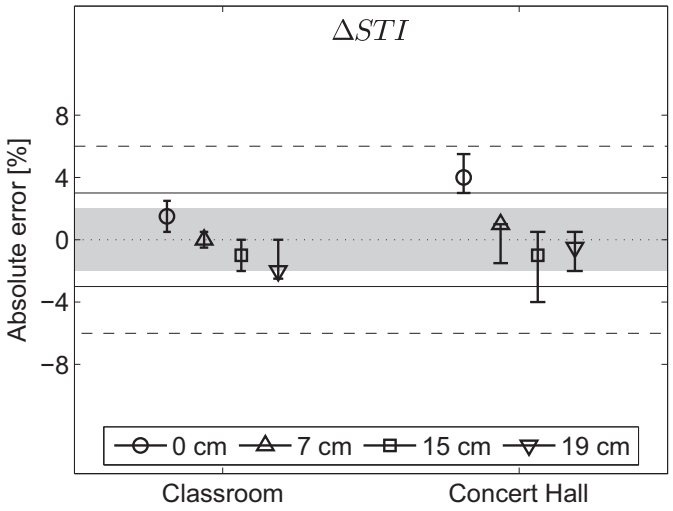


FIGURE B.1: Absolute error of the speech transmission index (STI) introduced by *LoRA* signal processing at different distances from the optimal listening positions in the simulated loudspeaker array. Results are shown for a simulated classroom (left) and a simulated concert hall (right). The shaded area denotes the head movement limen (representing the error range caused by head movement) and the solid line and dashed line denote the single and double subjective limen respectively (roughly corresponding to the just noticeable difference for the given parameter). [from Favrot and Buchholz (2010), printed with permission]

larger in the late part of the RIRs generated using *LoRA* than of the RIRs generated by *Odeon*.

B.2 Oreinos (2015)

Oreinos (2015) conducted a study where they measured the benefit of a binaural beamformer in bilateral hearing aids both objectively and subjectively in a reverberant environment. The beamformer benefit was measured and compared in three conditions: (a)

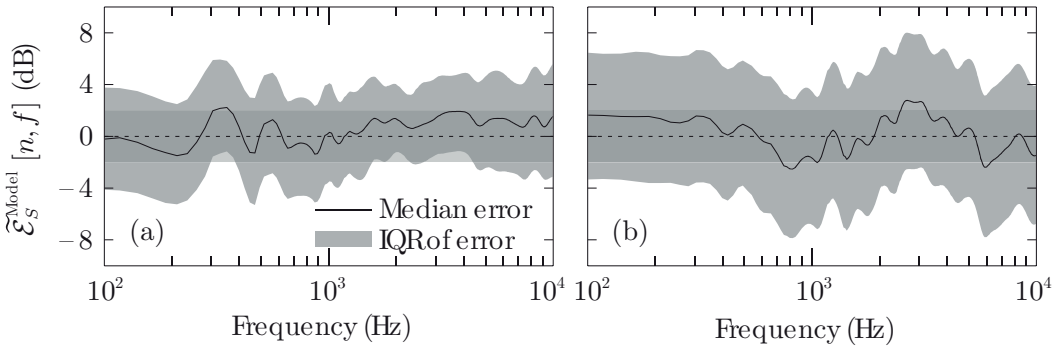


FIGURE B.2: Median and IQR of the short-time pressure errors incurred by modeling the real room using *LoRA* and measured at the output of the beamformer processing for: (a) the target, and (b) all (summed) distractors. The dark shaded area shows the ±2 dB error region. [from Oreinos (2015), printed with permission]

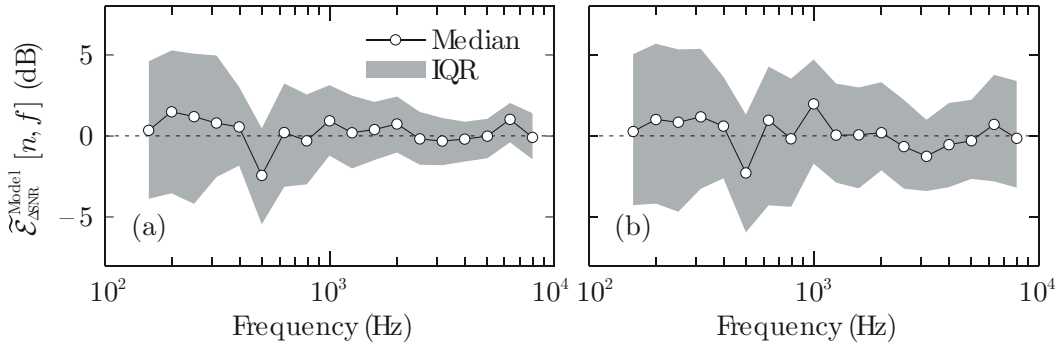


FIGURE B.3: Median and IQR of the SNR benefit error incurred by modeling the real room using *LoRA* and measured at the output of: (a) a cardiod heraing aid microphone, and (b) binaural beamformer processing. [from Oreinos (2015), printed with permission]

the real room, (b) a virtual reproduction of the real room using *LoRA*, and (c) a virtual reproduction of the real room using microphone array recordings.

The objective evaluation was performed using the sound pressure reconstruction error and the SNR benefit error at the output of the beamformer. The sound reproduction error was measured as the difference between the short-time power spectral density (PSD) in the real room and the virtual representations of the real room. The SNR benefit was measured as the difference in SNR between the signal at the frontal omnidirectional hearing aid microphone and the signal at the output of the beamformer. Figure B.2 shows the median and interquartile range of the sound pressure reconstruction error introduced by modeling the room using *LoRA* measured at the output of for (a) a speech target and (b) the sum of 14 speech distractors. The SNR benefit error measured at

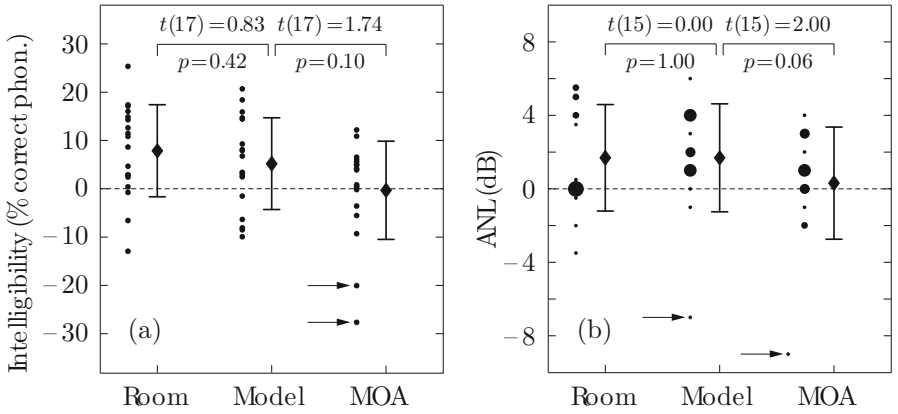


FIGURE B.4: Mean and scatter plots illustrating the benefit of the binaural beamformer, compared to the cardioid hearing aid microphone, on: (a) intelligibility (additional % correct phonemes), and (b) ANL (in dB). The whiskers show ±1 standard deviation about the mean, which is represented by the diamond. The scattered dots show all benefit observations, while the outlying values are indicated with the right-pointing arrows. The size of the dots in plot (b) increases with the number of underlying observations. [from Oreinos (2015), printed with permission]

the output of (a) a caridiod hearing aid microphone and (b) the binaural beamformer is shown in Fig. B.3.

A subjective evaluation of the beamformer benefit was carried out in terms of speech intelligibility (using BKB sentences) and acceptable noise level (ANL) in a group of 18 aided hearing impaired subjects. The ANL was the maximum level of a background noise signal tolerated by the subjects while listening to a target signal at the most comfortable sound level that was individually set by the subjects themselves. The measured benefit of the beamformer on these two metrics in all three conditions is shown in Fig. B.4.

Both the objective and the subjective evaluation show that results obtained in a virtual representation of a room generated using the *LoRA* toolbox are very similar to the results obtained in the real room. This consistent increase in the IACC was mainly due to the finite number of playback loudspeakers that limited the diffuseness of the reproduced sound field (i.e. reverberation).

Appendix C

Experimental Data

This appendix contains the detailed experimental data acquired in the experiments described in chapters 2 - 4. The multidimensional data obtained in the experiment described in chapter 5 has been made available online at http://www.nal.gov.au/download/MultisourceLocalizationData_TW.zip.

ID	$d = 1\,\mathrm{m}$				$d = 2\,\mathrm{m}$								$d = 4\,\mathrm{m}$		
	0°	90°	180°	270°	0°	45°	90°	135°	180°	225°	270°	315°	0°	180°	270°
NH1	−10.25	−12.5	−7.5	−12.5	−9	−11.5	−11	−8.75	−3	−7.75	−8.5	−9	−5	−5	−7
NH2	−10.25	−15.75	−6.75	−8.75	−6.5	−9	−10.25	−8	0	−3.5	−8	−9.25	−3.75	0.5	−3.5
NH3	−9.75	−14.25	−3.75	−11	−6.5	−8	−8	−3.75	1.5	−4.75	−6.5	−7	−1.75	1.5	−5.25
NH4	−10	−13.5	−6.75	−10.25	−7.25	−11	−8	−4.75	−1.5	−6.25	−5	−9	−5.25	−1.75	−4.75
NH5	−9.25	−9.25	−6.75	−10.25	−7.25	−8.25	−8	−2.5	0	−5.75	−8.25	−9.75	−2.25	−1.25	−2.25
NH6	−13.25	−13.75	−6.5	−11.5	−9.5	−10.75	−10.25	−6.5	−3.25	−7.75	−7	−13	−8	−3.25	−2.5
NH7	−8.5	−13	−5.5	−11.5	−7	−10.5	−10.75	−5	−3.5	−6.75	−6.5	−10.75	−3.5	−7.5	−7
NH8	−10.25	−11.75	−4.5	−11.75	−6.5	−12.25	−6.75	−3.25	0	−1.5	−3.75	−10	−5.75	−0.75	−3.5
NH9	−8	−14.5	−4.75	−9.25	−5.5	−11.5	−10.75	−3.5	−4	−5.25	−7.5	−12	−5	−2.75	−3.25
HI1	0.25	−0.75	2.25	−0.75	−1	−2	3	4	5.75	4	3	0.75	−0.75	3.25	2.75
HI2	−7.75	−8.5	−5.5	−13	−8.25	−7	−1.75	−4.5	−3.5	−7.5	−7.25	−7.5	−4.75	−2.25	−4
HI3	−2.25	−0.75	−0.75	−1.75	−0.5	−1.25	−0.75	1.25	1.25	1.75	1	1.75	2.5	1.25	2
HI4	−7.25	−3	−0.75	−1.75	−1.25	−2.25	−0.75	3.75	2	2.5	−1.75	−4.5	0.5	3.25	0
HI5	−6	−1	3	−2.75	−3.75	−4.25	0	0.25	0	1.25	0.5	−5.5	−0.75	4.75	0.25
HI6	−4.25	−3.75	−2.25	−2.75	−4	−2	−2.25	1.25	1.5	−2	−3.25	−2	0	−1	−1.25°
HI7	0.5	0.75	2.75	−0.5	1.25	2	3.25	5.75	5.5	2.5	4.25	1.75	4.25	6.5	2.75
HI8	−9	−4.5	−2.25	−6.25	−4.75	−4.25	−2.5	−2.25	0.25	−0.75	−2.25	−4.5	−3	1	−1.5
HI9	−6.5	−8	−3.5	−5.75	−3.75	−5	−2.5	−2.25	−3.75	−3.75	−5.75	−3.75	−1.25	−0.25	−1

TABLE C.1: Individual masked thresholds measured for the normal hearing (NH) and hearing impaired (HI) subjects participating in the experiments described in chapters 2 and 3. All values in [dB SNR]

ID	cond.	0°	45°	90°	135°	180°	225°	270°	315°
NH1	*fixed*	-0.7	-2.4	0.26	-2.48	-2.1	-0.08	-5.91	0.77
	random	-3.16	-2.55	0.41	-4.88	-1.22	-1.06	-1.95	-0.08
NH2	*fixed*	2.82	1.14	-0.33	-0.75	-5.76	-1.2	-0.25	-2.26
	random	1.58	2.6	1.17	0.29	-1.75	-0.59	1.33	0.17
NH3	*fixed*	-2.48	-2.51	0.28	0.19	-2.45	-0.28	4.78	0.4
	random	-2.02	-1.78	0.16	1.34	-2.93	-3.04	2.28	2.65
NH4	*fixed*	0.07	-0.15	3.35	1.3	1.19	1.16	0.32	2.12
	random	-0.08	1.28	3.28	0.95	1.73	-0.02	1.58	3.37
NH5	*fixed*	0.15	-0.44	-3.5	-1.28	-4.22	1.89	-3.11	-1.71
	random	-0.05	-1.3	0.35	1.11	-3.04	0.3	-2.25	-0.57
NH6	*fixed*	-0.9	0.8	-4.4	-0.77	-1.85	-1.89	-3.05	1.59
	random	-0.17	5.03	0.73	2	-1.96	-1.65	1.46	-1.14
NH7	*fixed*	-5.03	-2.88	-1.26	-4.07	-2.09	-2.5	-3.06	-1.55
	random	0.79	1.11	-2.25	-2.92	-2.94	0.19	-3.14	-1.51
NH8	*fixed*	-0.53	-2.34	-0.25	-2.75	0.35	-0.35	-3.54	0.18
	random	0.95	0.08	4.25	0.74	-2.62	-1.82	0.89	2.71
NH9	*fixed*	-2.04	-2.47	-3.97	-3.48	-1.22	2.1	-1.82	1.81
	random	-3.57	-3.91	-1.7	-2.68	-0.67	0.17	-2.79	-2.68
HI1	*fixed*	4.51	1.1	-1.51	-1.24	-2.03	-4.64	-0.94	2.5
	random	2.1	3.1	1.2	-1.92	-3.21	-2.79	-2.77	1.43
HI2	*fixed*	1.59	-4.47	-4.91	-0.28	0.86	4.54	6.95	-2.37
	random	0.91	-1603	-0.18	-0.7	3.64	-0.84	-0.87	-0.04
HI3	*fixed*	4.99	0.83	-0.2	-2.63	-0.99	-1.93	-0.18	0.79
	random	4.88	2.3	3.24	0.62	1.56	1.43	0.31	-1.23
HI4	*fixed*	-0.22	-0.18	-1.66	-3.14	-0.71	-4.47	-0.06	2.12
	random	1.53	0.98	1.26	-1.14	-1.45	-2.77	1.96	2.79
HI5	*fixed*	2.72	0.28	0.57	11.63	7.82	12.32	4.94	17.34
	random	3.11	4.28	-1.14	0.02	4.01	-1.48	2.89	7.18
HI6	*fixed*	3.37	-3.96	-0.92	-1.36	-0.14	-0.97	-0.8	0.32
	random	1.47	-1.2	-2.41	-3.36	-3.89	-2.34	-0.25	-3.55
HI7	*fixed*	0.97	-3.05	-2.86	-1.88	-1.37	-1.88	-1.51	0.3
	random	0.98	-0.25	0.89	-1.91	-1.62	0.33	0.08	-1.2
HI8	*fixed*	4.67	1.24	-0.58	-2.06	0.15	-5.34	0.86	-1.8
	random	0.78	0.07	2.12	0.08	-1.48	-3.36	-3.55	1.11

TABLE C.2: Individual masked thresholds derived from psychometric functions measured in two different conditions (*fixed* and *random*) for the normal hearing (NH) and hearing impaired (HI) subjects that participated in the experiment described in chapter 4. All values are in [dB] re the masked thresholds at $d = 2\,\mathrm{m}$ in table C.1.

Appendix D

Ethics approval

126 Greville Street
Chatswood NSW 2067
Australia
T (02) 9412 6872
F (02) 9411 8273
www.nal.gov.au

National Acoustic Laboratories
A division of Australian Hearing

Australian Hearing Human Research Ethics Committee *APPROVAL FOR RESEARCH INVOLVING HUMAN SUBJECTS*	
APPROVAL NUMBER: AHHREC2012-3	
PROJECT NUMBER	CRC R4.3.3b
PROJECT TITLE	Spatial hearing tests in loudspeaker-based real-world environments
CLASSIFICATION	*Class 2: Project with low risk*
PRINCIPAL INVESTIGATORS	Jörg Buchholz, Gitte Keidser, Virginia Best, Tobias Weller, Adam Westermann, Chris Orinos
DATE APPROVED/RATIFIED	13 March 2012
APPROVAL METHOD	Approved at the meeting of the full Committee on 13 March 2012.

This approval is based on the information contained in the ethics application that was presented to the Committee on 13 March 2012. A duplicate set of the documents is enclosed for your record.

In compliance with the "National Statement on Ethical Conduct in Human Research" (2007), Annual reporting to the Committee on progress of the project is required. This will happen in March or September of each year and you will be reminded near the time.

The Committee expects to be notified of any changes to the approved protocol or other issues that may have an impact on the ethics of the project either by means of the annual progress reports (checklists) or as an application for variation.

All future correspondence relating to the ethical aspects of this project must quote the above Approval Number.

Dr Tim Gainsford
Operations & Finance Manager, NAL
and AHHREC Secretary

® Registered Trademark of Australian Hearing Services
ABN 80 308 797 003

Australian Hearing Hub, Level 5
16 University Avenue, Macquarie University,
NSW 2109, Australia
T +61 2 9412 6872
F +61 2 9412 6769
www.nal.gov.au

National Acoustic Laboratories

Australian Hearing Human Research Ethics Committee *APPROVAL FOR RESEARCH INVOLVING HUMAN SUBJECTS*	
APPROVAL NUMBER: AHHREC2014-23	
Project Number	PhD14.1
Project Title	Assessment of spatial hearing in complex real-world environments
Classification	*Class 1: Project with negligible risk*
Principal Investigators authorized to conduct research	Tobias Weller, Jörg Buchholz, Virginia Best, Gitte Keidser
Date Approved	7/10/2014
Approval Method	Approved by the Research Director plus one other uninvolved senior NAL researcher as a Class 1 project with negligible risk.

This approval is based on the information contained in the ethics application that was presented to the Research Director on 1/10/2014 and is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007) available at: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72_national_statement_130624.pdf.

A duplicate set of the documents is enclosed for your record.

Annual reporting to the Committee on progress of the project is required including a final report when the work is completed or discontinued for any reason. Reminders will be sent when progress reports are due.

The Committee expects to be notified of any changes to the approved protocol or other issues that may have an impact on the ethics of the project either by means of the annual progress reports (checklists) or as an application for variation. Adverse or unforeseen events that affect the continued ethical acceptability of the project should be reported to the Chairman immediately.

All future correspondence relating to the ethical aspects of this project must quote the above Approval Number.

Dr Tim Gainsford
Operations & Finance Manager, NAL
and AHHREC Secretary

® Registered Trademark of Australian Hearing Services
ABN 80 308 797 003

Bibliography

Abouchacra, K. S., Emanuel, D. C., Blood, I. M., and Letowski, T. R. (**1998**). "Spatial percpetion of speech in various signal to noise ratios", Ear Hearing **19**, 298–309.

Akeroyd, M. A., Gatehouse, S., and Blaschke, J. (**2007**). "The detection of differences in the cues to distance by elderly hearing-impaired listeners", J. Acoust. Soc. Am. **121**, 1077–1089.

Akeroyd, M. A. and Guy, F. H. (**2011**). "The effect of hearing impairment on localization dominance for single-word stimuli", J. Acoust. Soc. Am. **130**, 312–323.

Allen, K., Alais, D., and Carlile, S. (**2009**). "Speech intelligibility reduces over distance from an attended location: Evidence for an auditory spatial gradient of attention", Atten. Percept. Psychophys. **71**, 164–173.

Baer, T. and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise", J. Acoust. Soc. Am. **94**, 1229–1241.

Baer, T. and Moore, B. C. J. (**1994**). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech", J. Acoust. Soc. Am. **95**, 2277–2280.

Best, V., Carlile, S., Kopčo, N., and van Schaik, A. (**2011**). "Localization in speech mixtures by listeners with hearing loss", J. Acoust. Soc. Am. **129**, 210–215.

Best, V., Marrone, N., Mason, C. R., Kidd, Jr., G., and Shinn-Cunningham, B. G. (**2008**a). "Effects of sensorineural hearing loss on visually guided attention in a multitalker environment", J. Assoc. Res. Otolaryngol. **10**, 142–149.

Best, V., Ozmeral, E., and Shinn-Cunningham, B. (**2007**). "Visually-guided attention enhances target identification in a complex auditory scene", J. Assoc. Res. Otolaryngol. **8**, 294–304.

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (**2008**b). "Object continuitiy enhances selective auditory attention", Proc. Natl. Acad. Sci. **105**, 13174–313178.

Beutelmann, R. and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners", J. Acoust. Soc. Am. **120**, 331–342.

Blauert, J. (**1997**). *Spatial Hearing*, 372–392 (MIT press, Cambridge).

Bolia, R. S., Nelson, W. T., and Morley, R. M. (**2001**). "Asymmetric performance in the cocktail party effect: implications for the design of spatial audio displays", Hum. Factors **43**, 208–216.

Bonino, A. and Leibold, L. (**2008**). "The effect of signal-temporal uncertainty on detection in bursts of noise or a random-frequency complex", J. Acoust. Soc. Am. **124**, 321–327.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural processing model based on contralateral inhibition. I. Model structure", J. Acoust. Soc. Am. **110**, 1074–1088.

Bronkhorst, A. W. (**1995**). "Localization of real and virtual sound sources", J. Acoust. Soc. Am. **98**, 2542–2553.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions", Acta Acust. united Ac. **86**, 117–128.

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

Bronkhorst, A. W. and Plomp, R. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing", J. Acoust. Soc. Am. **92**, 3132–3139.

Brungart, D. S., Cohen, J., Cord, M., Zion, D., and Kalluri, S. (**2014**). "Assessment of auditory spatial awareness in complex listening environments", J. Acoust. Soc. Am. **136**, 1808–1820.

Brungart, D. S. and Simpson, B. D. (**2007**). "Cocktail party listening in a dynamic multitalker environment", Percept. Psychophys. **69**, 79–91.

Buchholz, J. M. (**2013**). "A real-time hearing-aid research platform (HARP): Realization, calibration, and evaluation", Acta Acust. united Ac. **99**, 477–492.

Butler, R. A. (**1987**). "An analysis of the monaural displacement of sound in space", Percept. Psychophys. **41**, 1–7.

Buus, S., Florentine, M., and Poulsen, T. (**1997**). "Temporal integration of loudness, loudness discrimination, and the form of the loudness function", J. Acoust. Soc. Am. **101**, 669–680.

Buus, S., Schorer, E., Florentine, M., and Zwicker, E. (**1986**). "Decision rules in detection of simple and complex tones", J. Acoust. Soc. Am. **80**, 1646–1657.

Byrne, D. and Noble, W. (**1998**). "Optimizing sound localization with hearing aids", Trends Amplif. **3**, 51–73.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears", J. Acoust. Soc. Am. **25**, 975–979.

Colburn, H. S. (**1996**). "Computational models of binaural processing", in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay, volume 6, 332–400 (Springer, New York).

Culling, J. F. (**2007**). "Evidence specifically favoring the equalization-cancellation theory of binaural unmasking", J. Acoust. Soc. Am. **122**, 2803–2813.

Daniel, J. (**2000**). "Représentation des champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia", Ph.D. thesis, Université Paris 6.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the effective signal processing in the auditory system. I. Model structure", J. Acoust. Soc. Am. **99**, 3615–3622.

Dean, R. B. and Dixon, W. J. (**1951**). "Simplified statistics for small numbers of observations", Anal. Chem. **23**, 636–638.

Dillon, H. (**2012**). *Hearing Aids*, 370–403, 2nd edition (Thieme, New York).

Drullman, R. and Bronkhorst, A. W. (**2000**). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation", J. Acoust. Soc. Am. **107**, 2224–2235.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences", J. Acoust. Soc. Am. **35**, 1206–1218.

Durlach, N. I., Mason, C. R., Gallun, F. J., Shinn-Cunningham, B., Colburn, H. S., and Kidd, Jr., G. (**2005**). "Informational masking for simultaneous nonspeech stimuli: Psychometric functions for fixed and randomly mixed maskers", J. Acoust. Soc. Am. **118**, 2482–2497.

Durlach, N. I., Thompson, C. L., and Colburn, H. S. (**1981**). "Binaural interaction in impaired listeners: A review of past research", Int. J. Audiol. **20**, 181–211.

Egan, J. P., Greenberg, G. Z., and Schulman, A. I. (**1961**). "Interval of time uncertainty in auditory detection", J. Acoust. Soc. Am. **33**, 771–778.

Ericson, M. A., Brungart, D. S., and Simpson, B. D. (**2004**). "Factors that influence intelligibility in multitalker speech displays", Int J. Aviat. Psychol. **14**, 313 – 334.

Fan, W. L., Streeter, T. M., and Durlach, N. I. (**2008**). "Effect of spatial uncertainty of masker on masked detection for nonspeech stimuli", J. Acoust. Soc. Am. **124**, 36–39.

Fastl, H. and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models*, 203–238 (Springer, Berlin).

Favrot, S. and Buchholz, J. (**2010**). "LoRA: A loudspeaker-based room auralization system", Acta Acust. united Ac. **96**, 364–375.

Favrot, S. and Marschall, M. (**2012**). "Metrics for performance assessment of mixed-order Ambisonics spherical microphone arrays", in *Audio Engineering Society UK Conference 25*.

Favrot, S., Marschall, M., Käsbach, J., Buchholz, J. M., and Weller, T. (**2011**). "Mixed-order Ambisonics recording and playback for improving horizontal directionality", in *Audio Engineering Society Convention 131*.

Fitzgibbons, P. J. and Wightman, F. L. (**1982**). "Gap detection in normal and hearing-impaired listeners", J. Acoust. Soc. Am. **72**, 761–765.

Florentine, M. and Buus, S. (**1981**). "An excitation-pattern model for intensity discrimination", J. Acoust. Soc. Am. **70**, 1646–1654.

Florentine, M., Buus, S., and Geng, W. (**1999**). "Psychometric functions for gap detection in a yes-no procedure", J. Acoust. Soc. Am. **106**, 3512–3520.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech", J. Acoust. Soc. Am. **106**, 3578–3588.

Fründ, I., Haenel, N., and Wichmann, F. (**2011**). "Inference for psychometric functions in the presence of nonstationary behavior", J. Vision **11**.

Gardner, M. B. (**1969**). "Distance estimation of 0° or apparent 0° oriented speech signals in anechoic space", J. Acoust. Soc. Am. **45**, 47–53.

Gatehouse, S. and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (SSQ)", Int. J. Audiol. **43**, 85–99.

George, E. L. J., Festen, J. M., and Houtgast, T. (**2006**). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners", J. Acoust. Soc. Am. **120**, 2295–2311.

Gerzon, M. A. (**1992**). "General metatheory of auditory localisation", in *Audio Engineering Society Convention 92*.

Good, M. D. and Gilkey, R. H. (**1996**). "Sound localization in noise: The effect of signal-to-noise ratio", J. Acoust. Soc. Am. **99**, 1108–1117.

Green, D. M. (**1961**). "Detection of auditory sinusoids of uncertain frequency", J. Acoust. Soc. Am. **33**, 897–903.

Green, D. M. and Weber, D. L. (**1980**). "Detection of temporally uncertain signals", J. Acoust. Soc. Am. **67**, 1304–1311.

Grubbs, F. E. (**1950**). "Sample criteria for testing outlying observations", Ann. Math. Stat. **21**, 27–58.

Hacker, M. J. and Ratcliff, R. (**1979**). "A revised table of d' for M-alternative forced choice", Percept. Psychophys. **26**, 168–170.

Hammershøi, D. and Møller, H. (**2005**). "Binaural technique - basic methods for recording, synthesis, and reproduction", in *Communication Acoustics*, edited by J. Blauert, 223–254 (Springer, Berlin).

Hant, J. J. and Alwan, A. (**2003**). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise", Speech Commun. **40**, 291–313.

Hant, J. J., Strope, B. P., and Alwan, A. A. (**1997**). "A psychoacoustic model for the noise masking of plosive bursts", J. Acoust. Soc. Am. **101**, 2789–2802.

Häusler, R., Colburn, S., and Marr, E. (**1983**). "Sound localization in subjects with impaired hearing", Acta Oto-Laryngol. **96**, 1–62.

Hawkins, D. B. and Wightman, F. L. (**1980**). "Interaural time discrimination ability of listeners with sensorineural hearing loss", Int. J. Audiol. **19**, 495–507.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). "Speech intelligibility and localization in a multi-source environment", J. Acoust. Soc. Am. **105**, 3436–3448.

Hirsh, I. J. (**1950**). "The relation between localization and intelligibility", J. Acoust. Soc. Am. **22**, 196–200.

Jepsen, M. L. and Dau, T. (**2011**). "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss", J. Acoust. Soc. Am. **129**, 262–281.

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception", J. Acoust. Soc. Am. **124**, 422–438.

Jones, G. L. and Litovsky, R. Y. (**2008**). "Role of masker predictability in the cocktail party problem", J. Acoust. Soc. Am. **124**, 3818–3830.

Keidser, G., Rohrseitz, K., Dillon, H., Hamacher, V., Carter, L., Rass, U., and Convery, E. (**2006**). "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers", Int. J. Audiol. **45**, 563–579.

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

Kopčo, N., Best, V., and Carlile, S. (**2010**). "Speech localization in a multitalker mixture", J. Acoust. Soc. Am. **127**, 1450–1457.

Le Goff, N., Buchholz, J. M., and Dau, T. (**2013**). "Modeling horizontal localization of complex sounds in the impaired and aided impaired auditory system", in *The Technology of Binaural Listening*, edited by J. Blauert, 121–144 (Springer, Berlin).

Leibold, L. J. and Werner, L. A. (**2006**). "Effect of masker-frequency variability on the detection performance of infants and adults", J. Acoust. Soc. Am. **119**, 3960–3970.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics", J. Acoust. Soc. Am. **49**, 467–477.

Loomis, J. M., Klatzky, R. L., Philbeck, J. W., and Golledge, R. G. (**1998**). "Assessing auditory distance perception using perceptually directed action", Percept. Psychophys. **60**, 966–980.

Lorenzi, C., Gatehouse, S., and Lever, C. (**1999**a). "Sound localization in noise in hearing-impaired listeners", J. Acoust. Soc. Am. **105**, 3454–3463.

Lorenzi, C., Gatehouse, S., and Lever, C. (**1999**b). "Sound localization in noise in normal-hearing listeners", J. Acoust. Soc. Am. **105**, 1810–1820.

Marschall, M., Favrot, S., and Buchholz, J. (**2012**). "Robustness of a mixed-order Ambisonics microphone array for sound field reproduction", in *Audio Engineering Society Convention 132*.

Minaar, P., Albeck, S. F., Simonsen, C. S., Søndersted, B., Oakley, S. A. D., and Bennedbæk, J. (**2013**). "Reproducing real-life listening situations in the laboratory for testing hearing aids", in *Audio Engineering Society Convention 135*.

Minaar, P., Favrot, S., and Buchholz, J. (**2010**). "Improving hearing aids through listening tests in a virtual sound environment", Hear. J. **63**, 40–44.

Mondor, T. A. and Zatorre, R. J. (**1995**). "Shifting focusing auditory spatial attention", J. Exp. Psychol. Hum. Percept. Perform. **21**, 387–409.

Moore, B. C. and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss", Hearing Research **188**, 70–88.

Moore, B. C. J. (**2003**). *An introduction to the psychology of hearing*, 65–126, 5th edition (Academic Press, London, UK).

Moore, B. C. J. (**2007**). *Cochlear hearing loss*, 2nd edition (John Wiley & Sons).

Moore, B. C. J. and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acoust. Soc. Am. **74**, 750–753.

Moreau, S., Daniel, J., and Bertet, S. (**2006**). "3D sound field recording with higher order Ambisonics - objective measurements and validation of spherical microphone", in *Audio Engineering Society Convention 120*.

Mueller, M. F., Kegel, A., Schimmel, S. M., Dillier, N., and Hofbauer, M. (**2012**). "Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes", J. Acoust. Soc. Am. **131**, 4732–4742.

Neff, D. L. and Green, D. M. (**1987**). "Masking produced by spectral uncertainty with multicomponent maskers", Percept. Psychophys. **41**, 409–415.

Noble, W., Byrne, D., and Lepage, B. (**1994**). "Effects on sound localization of configuration and type of hearing impairment", J. Acoust. Soc. Am. **95**, 992–1005.

Noble, W. and Gatehouse, S. (**2004**). "Interaural asymmetry of hearing loss, speech, spatial and qualities of hearing scale (SSQ) disabilities, and handicap", Int. J. Audiol. **43**, 100–114.

Noble, W. and Gatehouse, S. (**2006**). "Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the speech, spatial, and qualities of hearing scale (ssq)", Int. J. Audiol. **45**, 172–181.

Noble, W., Ter-Horst, K., and Byrne, D. (**1995**). "Disabilities and handicaps associated with impaired auditory localization", J. Am. Acad. Audiol. **6**, 129–140.

Oreinos, C. (**2015**). "Virtual acoustic environments for the evaluation of hearing devices", Ph.D. thesis, Macquarie University.

Oreinos, C. and Buchholz, J. (**2014**). "Objective analysis of higher-order Ambisonics for hearing aid applications", Submitted.

Oreinos, C. and Buchholz, J. M. (**2013**). "Measurement of a full 3D set of HRTFs for in-ear and hearing aid microphones on a Head and Torso Simulator (HATS)", Acta Acust. united Ac. **99**, 836–844.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold", J. Acoust. Soc. Am. **72**, 1788–1803.

Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)", Acta Acust. united Ac. **34**, 200–211.

Rindel, J. (**2000**). "The use of computer modeling in room acoustics", J. Vibroeng. **3**, 41–72.

Rorabacher, D. B. (**1991**). "Statistical treatment for rejection of deviant values: critical values of Dixon's Q parameter and related subrange ratios at the 95% confidence level", Anal. Chem. **63**, 139–146.

Rosenhall, U. (**1985**). "The influence of hearing loss on directional hearing", Scandinavian Audiol. **14**, 187–189.

Scharf, B. (**1961**). "Loudness summation and spectrum shape", J. Acoust. Soc. Am. **33**, 838–839.

Seeber, B. U., Kerber, S., and Hafter, E. R. (**2010**). "A system to simulate and reproduce audio-visual environments for spatial hearing research", Hearing Res. **260**, 1–10.

Shailer, M. J. and Moore, B. C. J. (**1987**). "Gap detection and the auditory filter: Phase effects using sinusoidal stimuli", J. Acoust. Soc. Am. **81**, 1110–1117.

Teder-Sälejärvi, W. A. and Hillyard, S. A. (**1998**). "The gradient of spatial auditory attention in free field: An event-related potential study", Percept. Psychophys. **60**, 1228–1242.

Tikhonov, A. N. and Arsenin, V. Y. (**1977**). *Solutions of ill-posed problems* (V. H. Winston & Sons).

Travis, C. (**2009**). "A new mixed-order scheme for Ambisonics signals", in *Ambisonics Symposium*.

van den Bogaert, T., Carette, E., and Wouters, J. (**2011**). "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna", Int. J. Audiol. **50**, 164–176.

van den Bogaert, T., Klasen, T. J., Moonen, M., van Deun, L., and Wouters, J. (**2006**). "Horizontal localization with bilateral hearing aids: Without is better than with", J. Acoust. Soc. Am. **119**, 515–526.

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks", J. Acoust. Soc. Am. **90**, 858–865.

Völk, F. and Fastl, H. (**2010**). "Advantages of binaural room synthesis for research and fitting of hearing aids, cochlear implants, electro-acoustical stimulation, and combined systems", in *Proceedings of the 20th International Congress on Acoustics*.

Weller, T. and Buchholz, J. M. (**2015**). "Auditory masking of speech in reverberant multi-talker environments", J. Acoust. Soc. Am. submitted.

Westermann, A. and Buchholz, J. M. (**2015**). "The influence of informational masking in reverberant, multi-talker environments", J. Acoust. Soc. Am. In print.

Wiggins, I. M. and Seeber, B. U. (**2012**). "Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners", Ear Hearing **33**, 399–410.

Wightman, F. L. and Kistler, D. J. (**1999**). "Resolution of front-back ambiguity in spatial hearing by listener and source movement", J. Acoust. Soc. Am. **105**, 2841–2853.

Wright, B. A. and Saberi, K. (**1999**). "Strategies used to detect auditory signals in small sets of random maskers", J. Acoust. Soc. Am. **105**, 1765–1775.

Zahorik, P. (**2002**). "Assessing auditory distance perception using virtual acoustics", J. Acoust. Soc. Am. **111**, 1832–1846.

Zurek, P. M., Freyman, R. L., and Balakrishnan, U. (**2004**). "Auditory target detection in reverberation", J. Acoust. Soc. Am. **115**, 1609–1620.

Zwicker, E. (**1956**). "Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs", Acustica **6**, 365–381.

Zwislocki, J. J. (**1969**). "Temporal summation of loudness: An analysis", J. Acoust. Soc. Am. **46**, 431–441.