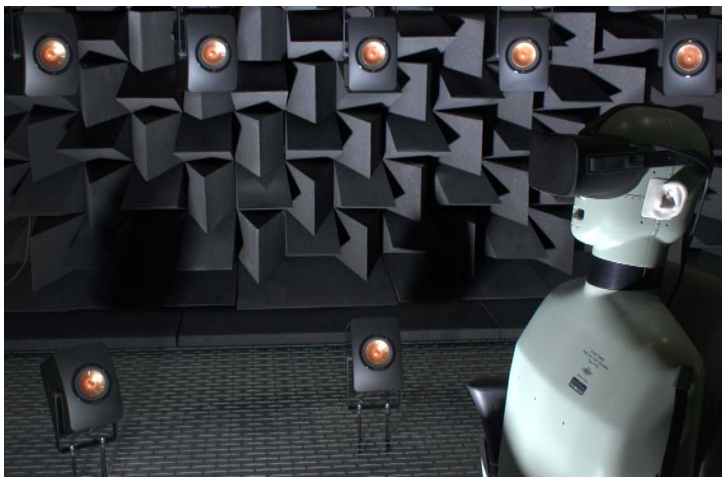
CONTRIBUTIONS TO
HEARING RESEARCH

Volume 39

*Axel Ahrens*

# Characterizing auditory and audio-visual perception in virtual environments



**DTU Health Tech**
Department of Health Technology

# Characterizing auditory and audio-visual perception in virtual environments

PhD thesis by
Axel Ahrens

Preliminary version: July 4, 2019


Technical University of Denmark

2019

This PhD dissertation is the result of a research project carried out at the Hearing Systems Section, Department of Health Technology (formerly Electrical Engineering), Technical University of Denmark.

The project was partly financed by the Oticon Centre of Excellence for Hearing and Speech Sciences (2/3) and by the Technical University of Denmark (1/3).

## Supervisors

**Prof. Torsten Dau**
**Dr. Marton Marschall**
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

# Abstract

One of the challenges in hearing research is to explain the human ability to understand speech in complex, noisy environments, commonly referred to as a cocktail-party scenario. To gain a better understanding of how the auditory system performs in complex acoustic environments, one approach is to reproduce such listening situations in the laboratory. By applying spatial audio reproduction techniques, sound fields can be reproduced, which may be well-suited for bringing more realistic sound scenes into the laboratory. However, physical limitations affect the reproduction methods and might also affect perception. In addition to acoustic information, auditory perception can be influenced by visual information. Virtual reality glasses might be a promising tool to add visual information to virtual acoustic scenarios. However, a perceptual characterization of virtual audio-visual reproductions is lacking.

This thesis focused on three aspects related to the perception in virtual auditory and audio-visual environments: (i) The accuracy of the reproduction of a virtual acoustic room in terms of speech intelligibility, (ii) the relation between the source size and speech intelligibility, and (iii) the role of visual information and the impact of virtual reality glasses on sound localization. It is demonstrated that the acoustic reproduction based on impulse responses measured with a microphone array provides the closest match to a reverberant reference room in terms of speech intelligibility, while a reproduction based on room acoustic simulations shows significantly different results as compared to a reference room. The differences in speech intelligibility can be accounted for by using a computational speech intelligibility model. Furthermore, it is shown that speech intelligibility is worse in conditions where the energy of a target and an interfering speech is spatially spread in comparison to point-like sources. The relationship between the energy spread and speech intelligibility can be described with a computational model that utilizes a better-ear listening strategy. Finally, it is demonstrated that virtual reality glasses disturb the acoustic field around the head which can decrease the sound localization accuracy. When virtual visual information is presented, the sound source localization accuracy improves to a comparable extent as it has been shown in realistic environments.

Overall, throughout this thesis, it is shown that virtual reality glasses and loudspeaker-based virtual sound environments represent powerful tools for the reproduction of realistic scenarios and contribute to a better understanding of auditory processing and perception in cocktail party-like scenarios.

# Resumé

En af udfordringerne i høreforskning er at forklare menneskets evne til at forstå tale i komplekse og støjende miljøer, ofte refereret til som "cocktail party'"-scenariet. For bedre at forstå, hvordan menneskets auditoriske system fungerer i komplekse akustiske miljøer, er en fremgangsmåde at reproducere sådanne lyttesituationer i et laboratorie. Ved at benytte rumlige lydreproduktionsteknikker kan man gengive optagede eller syntetiserede lydfelter, hvilket kan være velegnet til at bringe mere realistiske lydscenarier ind i laboratoriet. Dog påvirker fysiske begrænsninger forskellige reproduktionsmetoder og kan også påvirke lydopfattelsen. Ud over akustisk information kan auditorisk opfattelse også påvirkes af visuel information. Virtual reality-briller kan være et lovende værktøj til at tilføje visuel information til virtuelle akustiske scenarier. En perceptuel karakterisering af virtuelle audiovisuelle reproduktioner findes dog ikke.

Denne afhandling fokuserede på tre aspekter relateret til opfattelse i virtuelle audio- og audiovisuelle miljøer: (i) Nøjagtigheden af reproduktionen af et virtuelt akustisk rum i forhold til taleforståelse, (ii) forholdet mellem kildestørrelse og taleforståelse, og (iii) rollen af visuel information og effekten af virtual reality-briller på lydlokalisering. Det påvises, at den akustiske reproduktion baseret på impulsresponser, målt med et mikrofon-array, giver det tætteste match på et referencerum med efterklang i forhold til taleforståelse, mens en gengivelse baseret på rumakustiske simuleringer viser signifikant forskellige resultater sammenlignet med et referencerum. Forskellene kan redegøres for ved hjælp af en beregningsmodel for taleforståelse. Forholdet mellem størrelse af virtuelle lydkilder og taleforståelse er også fundet betydningsfuld. Det påvises, at taleforståelse er dårligere i situationer, hvor energien fra en taler og en forstyrrende taler spredes rumligt i forhold til punktlignende kilder. Forholdet mellem energispredningen og taleforståelsen kan beskrives med en beregningsmodel, der anvender en "bedre-øre"-lyttestrategi. Endelig er det påvist, at virtual reality-briller forstyrrer det akustiske felt omkring hovedet, hvilket kan reducere nøjagtigheden af lydlokaliseringen. Når virtuel visuel information tilføjes, forbedres nøjagtigheden af lydkildelokaliseringen til at være sammenlignelig med realistiske miljøer.

Overordnet gennem denne afhandling påvises det at virtual reality-briller og højttalerbaserede virtuelle lydmiljøer er kraftfulde værktøjer til at reproducere realistiske scenarier og at de bidrager til en bedre forståelse af auditorisk processering og opfattelse i cocktailparty-lignende scenarier.

# Acknowledgments

First, I would like to thank my supervisors, Torsten and Marton, for their support. Also, a huge thanks to all my colleagues from the Hearing Systems group, who I had so much fun with, in particular all the fantastic people I shared an office with. A special hug goes to Caroline for all of her support over the years.

A big part of this project took place in and around the audio-visual immersion lab (AVIL). Thank you to all the people who helped designing, building and developing it. I deeply enjoyed the process. I also want to thank Pauli Minaar for the mentoring throughout the years as well as the Oticon Fonden for funding the project.

Thanks to the three examiners, Pavel Zahorik, Piotr Majdak and Ewen MacDonald. You made the defense a challenge and great experience. I never thought that the defense could actually be fun!

Finally, I want to thank my family and Fede for keeping up with me and supporting me during this long process.

# Related publications

## Journal papers

- Ahrens, A., Marschall, M., and Dau, T. (**2019a**). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments", Hearing Research 377. 10.1016/j.heares.2019.02.003

- Ahrens, A., Duemose Lund, K., Marschall, M., and Dau, T. (**2019b**). "Sound source localization with varying amount of visual information in virtual reality", PLOS ONE 14(3): e0214603. 10.1371/journal.pone.0214603

- Ahrens, A., Marschall, M., and Dau, T. (**submitted**). "The effect of sound source width on speech intelligibility in anechoic and reverberant environments".

## Conference papers

- Ahrens, A., Joshi, S.N., and Epp, B. (**2015**). "Spektrale Gewichtung von interauralen Zeit- und Pegelunterschieden zur Lateralisierung von Breitbandsignalen", Proceedings of the Deutsche Gesellschaft für Akustik, 41st German Convention on Acoustics, Nuremberg, Germany, March 2015.

- Ahrens, A., Marschall, M., and Dau, T. (**2017**). "Evaluating a Loudspeaker-Based Virtual Sound Environment using Speech-on-Speech Masking", Proceedings of the Deutsche Gesellschaft für Akustik, 43th German Convention on Acoustics, Kiel, Germany, March 2017.

- Ahrens, A., Dau, T., and Marschall, M. (**2017**). "Effect of 2D ambisonics order on speech intelligibility with closely spaced talkers", Proceedings of the 4th International Conference on Spatial Audio, Graz, Austria, September 2017.

## Published abstracts

- Ahrens, A., Joshi, S.N., and Epp, B. (**2015**). "Spectral Weighting of Binaural Cues: Effect of Bandwidth and Stream Segregation", 38th Annual MidWinter Meeting of the Association for Research in Otolaryngology, Baltimore, MD, United States, February 2015.

- Ahrens, A., Marschall, M., and Dau, T. (**2016**). "Evaluating the auralization of a small room in a virtual sound environment using objective room acoustic measures", J. Acoust. Soc. Am. 140, 3177, Honolulu, HI, United States, December 2016.

- Ahrens, A., Marschall, M., and Dau, T. (**2017**). "Measuring speech intelligibility with speech and noise interferers in a loudspeaker-based virtual sound environment", J. Acoust. Soc. Am. 141, 3510, Boston, MA, United States, June 2017.

- Epp, B., Ahrens, A., and Joshi, S.N. (**2017**). "Spectral weighting of interaural time- and level differences for broadband signals", J. Acoust. Soc. Am. 141, 3891, Boston, MA, United States, June 2017.

- Ahrens, A., Marschall, M., and Dau, T. (**2018**). "The Relation between Source Width Perception and Speech Intelligibility with Virtual Sound Sources", 41st Annual MidWinter Meeting of the Association for Research in Otolaryngology, San Diego, CA, United States, February 2018.

## Datasets

- Ahrens, A. (**2018**). "Binaural Impulse Responses with and without HTC Vive HMD", Zenodo. 10.5281/zenodo.1185335

- Ahrens, A. (**2018**). "Room acoustics model of a listening room", Zenodo. 10.5281/zenodo.1232317

- Ahrens, A., Duemose Lund, K., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality", Zenodo. 10.5281/zenodo.1293059

# Contents

# 1 General introduction

The perception of an acoustic stimulus in the presence of interfering stimuli is arguably one of the most studied areas within hearing research. Colin Cherry (1953) coined the term "cocktail-party problem" and showed that listeners can selectively focus on a speech signal presented in one ear while another interfering talker is simultaneously presented to the other ear. Although Cherry's cocktail party was rather simplistic, with only two talkers presented over headphones (Middlebrooks et al., 2017, chapter 1), the study nevertheless triggered the birth of a whole field of research on speech perception in the presence of interfering stimuli (see Bronkhorst, 2000; Middlebrooks et al., 2017, for reviews).

## 1.1 Perception of sounds in space

The auditory system relies on multiple cues in the process of perceiving a single source in space. To localize a sound in the horizontal plane, i.e. in the azimuth direction, two cues are mainly used by the auditory system, interaural time and level differences (ITDs and ILDs). ITDs have been shown to be of main importance at low frequencies whereas ILDs dominate at high frequencies, which is also referred to as the duplex theory of sound localization (Macpherson and Middlebrooks, 2002; Rayleigh, 1907).

The cues used for localization in the median plane, i.e. elevation, are less well understood. The head and the pinna act as direction-dependent filters, also known as HRTFs. The resulting colorations are a cue to identify the elevation of a source (Batteau, 1967, 1968; Fisher and Freedman, 1968). It has been shown that the HRTFs, or features of the HRTFs, are learned and stored in the auditory system (Hofman et al., 1998; Van Wanrooij, 2005). However, to utilize the HRTF information, the auditory system needs to make assumptions about the generally unknown source spectrum. Thus, other cues are needed to solve this ill-posed problem, for example head motion or the assumptions of locally constant spectra (Middlebrooks, 1992). In fact, elevation perception has been

proposed to be based on multi-feature, template-based matching (Baumgartner et al., 2014; Macpherson and Sabin, 2013; Van Opstal et al., 2017).

In addition to azimuth and elevation, source distance and size are essential to entirely describe a sound source in space. The main auditory cues for distance perception are the intensity of the sound and the direct-to-reverberant ratio (DRR) in enclosed or semi-enclosed spaces (Zahorik et al., 2005). For the intensity cue, the auditory system needs to estimate the source level, which is not a reliable cue if the source is not familiar to the listener (Coleman, 1962; McGregor et al., 1985; Zahorik, 2002). The DRR is the ratio between the energy reaching the listener directly and the energy that is reflected from surfaces (reverberant energy) before arriving at the listener. The DRR decreases with increasing distance between source and receiver because the reverberant energy remains approximately constant if the room is sufficiently reverberant, while the direct energy decreases with increasing distance, according to the inverse-square law (Zahorik, 2002).

The size of a source and its underlying cues have been defined in multiple ways in previous studies. Blauert (1997) defined the acoustic size of an object as a localization blur or the smallest possible change of position that the auditory system can detect. However, a sound can be accurately localizable and perceived as being large at the same time, as for example in concert halls (Griesinger, 1997). The interaural cross-correlation (IACC) (Ando, 2007; Schroeder et al., 1974) as well as the lateral energy fraction (LF) (Bradley, 2011) have been proposed as physical correlates to the source size percept.

## 1.2 Auditory information for cocktail-party listening

When considering multiple sound sources in an acoustic scene, the auditory system needs to segregate these sources to process the information. Speech intelligibility in the presence of background noise is generally improved when the target and the interfering signals are spatially separated compared to a colocated configuration of the sources (Bronkhorst, 2000). It has been shown that the separation in azimuth (Duquesnoy, 1983), elevation (Martin et al., 2012) and distance (Westermann and Buchholz, 2015) improves the ability to segregate sources. However, the effect of differences in source size on speech intelligibility has not been systematically investigated.

Two mechanisms have been shown to help speech understanding in sit-

uations with spatially separated sources: binaural unmasking and better-ear listening. The binaural unmasking component arises from the interaural phase differences between a target and an interferer (Durlach, 1963; Durlach, 1972; Hirsh, 1948; Licklider, 1948). The difference in detection thresholds between a condition where target and interferer have equal interaural phase and a condition where target and interferer have different interaural phases is referred to as the binaural masking level difference (BMLD). The BMLD was initially modelled as an equalization-cancellation process only including the phase differences (Durlach, 1963). Later it was shown that the interaural coherence, i.e. the similarity between the ear signals, of the masker also affects the BMLD as the equalization-cancellation process cannot fully cancel the masking signal (Culling et al., 2004). In the presence of reverberation, the interaural coherence is reduced compared to an anechoic condition, and thus, the BMLD is also lower (Lavandier and Culling, 2007, 2010; Monaghan et al., 2013).

The better-ear listening component arises from the ability of the auditory system to use the information at the ear with the better signal-to-noise ratio (SNR). The head acts as an obstacle, i.e. a direction-dependent filter, for the sound waves which can be measured as an ILD. Thus, when the signal and the interferer are at different locations, the SNR at the two ears can be different. The SNR difference between the ears is particularly large when target and interferer are located at opposite sides of the head. The auditory system has been shown to be able to take advantage of the better-ear SNR across frequency and time (Brungart and Iyer, 2012; Culling and Mansell, 2013; Glyde et al., 2013). In reverberant environments, the long-term better-ear advantage is reduced as the ear signals tend to become more similar.

Some computational auditory models have been developed that predict binaural speech intelligibility based on the two components, better-ear listening and binaural unmasking (Beutelmann and Brand, 2006; Beutelmann et al., 2010; Chabot-Leclerc et al., 2016; Jelfs et al., 2011; Lavandier and Culling, 2010; Lavandier et al., 2012; Rennies et al., 2011; Wan et al., 2010, 2014). These models have been shown to predict speech intelligibility in numerous conditions, for example in environments with varying degrees of reverberation and with different interferer configurations. The application of these models allows a quantification of how the auditory system may be utilizing the better-ear listening and binaural unmasking components in various listening conditions.

## 1.3 Visual information for cocktail-party listening

In addition to auditory cues, visual information can affect how auditory stimuli are perceived. Visual cues can improve speech intelligibility (Sumby and Pollack, 1954) but can also alter the perception of auditory cues for speech (McGurk effect; McGurk and MacDonald, 1976) and for spatial location (ventriloquism effect; Howard and Templeton, 1966). Thus, when aiming to reproduce a cocktail-party scenario, visual information needs to be considered in addition to the auditory information.

## 1.4 Towards creating a virtual cocktail party

Generally, studies that aim to investigate perception in realistic multi-talker environments have used simplified setups to reproduce a cocktail-party scenario. These simplifications typically involve a reduced number of sound sources, the absence of reverberation as well as the exclusion of head movements. Thus, these setups might not be ecologically valid and may not reflect effects that would normally occur in real listening scenarios.

Various factors in a real-world scenario need to be captured in a virtual representation, such that the percept is natural or, ideally, indistinguishable from the real world. To create an auditory virtual scene that is perceptually indistinguishable from a real one, the sensory nervous system needs to receive inputs containing the relevant features. The reproduction of these input signals is still limited by technology. Although several studies have addressed this, it is still unclear what features are needed to create a realistic virtual auditory scene.

Previous volumes in this series of theses *"Contributions to Hearing Research"*, have shown significant progress in creating and evaluating virtual auditory scenes. In Vol. 9 *(Sylvain Favrot: A loudspeaker-based room auralization system for auditory research)*, a method for the auralization of rooms was developed and evaluated (Favrot and Buchholz, 2010). The acoustic scenes were reproduced with an array of loudspeakers and were based on computational room acoustic simulations. The advantage of loudspeaker-based reproduction is that it allows for head-movements without the need for head-tracking, which is otherwise necessary for the reproduction over headphones. Room acoustic simulations allow the creation of environments that do not physically exist or enable the modification of environments to investigate certain room acoustical features.

The playback in Favrot and Buchholz (2010) was done using either higher-order ambisonics (HOA) or a nearest-loudspeaker mapping (NLM). HOA is a method based on the spherical harmonics decomposition of a three-dimensional sound field. The reproduction accuracy increases with the number of spherical harmonics, i.e. the order (M), applied. The number of transducers (N) needed to capture/reproduce HOA signals of order M is $N \geq (M+1)^2$ (Ward and Abhayapala, 2001). With decreasing order, pressure and phase errors in the reproduced sound field increase. For a given order, the errors increase with increasing frequency as well as with distance from the centre of the loudspeaker array.

When applying the NLM approach, such position and frequency dependent errors are avoided by mapping the direct sound and each of the early reflections to the geometrically closest loudspeaker. In essence, virtual sources are replaced by real sources (the closest loudspeaker) at the expense of accuracy in terms of source location and distance. The impact of this simplification may be negligible when many loudspeakers are available, and when the considered sound sources are at a similar distance as the loudspeakers. For the late reflections, Favrot and Buchholz (2010) proposed a reproduction method based on energy envelopes represented in 1$^{st}$ order ambisonics and multiplied with uncorrelated noise for each loudspeaker in an attempt to create a diffuse sound field.

The disadvantage of room acoustic simulations is that the acoustic properties of the surfaces need to be known or estimated. Thus, it can be challenging to model an existing room with high accuracy. Furthermore, modeling complex environments with many sources or moving sources is cumbersome. To capture such complex scenes, microphone array recordings have been shown to be a valuable tool. In Vol. 18 of this collection *(Marton Marschall: Capturing and reproducing realistic acoustic scenes for hearing research)*, the development of such a microphone array was carried out. The array consists of 52 microphone capsules on a rigid sphere with a radius of 5 cm. To record and play back an acoustic scene, HOA coding and decoding is used. However, similarly to HOA playback, for HOA capture the physical limitations of the number and spacing between microphone capsules result in a limitation of the ambisonics order, and thus, additional errors in the capture process. Thus, when using microphone arrays for capture and loudspeaker arrays for playback, the errors from both ambisonics capture and reproduction processes contribute to the overall error (Oreinos, 2015).

In other volumes of this series, the perception in virtual environments was

further investigated, as in Vol. 29 *(Jens Cubick: Investigating distance perception, externalization and speech intelligibility in complex acoustic environments)*, where the influence of visual information on auditory distance perception was examined (Gil-Carvajal et al., 2016). To vary the visual information presented to the listeners, Gil-Carvajal et al. (2016) physically placed the listeners in different rooms. Virtual visual reproductions of environments and scenarios would allow greater flexibility in conducting such experiments. However, unnatural visual and auditory information can lead to altered head- and body-motion in relation to real-world behavior (Hendrikse et al., 2018).

Other studies investigated the reproduction of realistic scenarios in the laboratory for hearing research using similar approaches. Seeber et al. (2010) reported the development and evaluation of loudspeaker arrays for spatial hearing research. Grimm et al. (2016), Minnaar et al. (2010), and Oreinos and Buchholz (2016) described loudspeaker array setups particularly designed for hearing aid evaluations and hearing research in general. Pausch et al. (2018) presented a cross-talk cancellation system for the same purpose.

## 1.5 Perception in virtual environments

Previous research on the evaluation of virtual sound environments has focused on physical parameters, on quality of experience, or on psychoacoustic measures. As described above, HOA leads to pressure and phase errors above certain frequencies depending on the ambisonics order. Numerous studies have investigated these limitations using physical measures such as spectral errors (Daniel, 2001; Epain et al., 2010; Favrot and Marschall, 2012; Marschall et al., 2012; Oreinos, 2015; Poletti, 2005; Solvang, 2008; Ward and Abhayapala, 2001). Other studies investigated the accuracy of reproduced room acoustic parameters such as reverberation time or clarity (Cubick and Dau, 2016; Favrot and Buchholz, 2010).

However, for perceptual research, a physically accurate sound field might not always be necessary. The quality of experience is related to the authenticity and plausibility of spatial sound reproductions (Wierstorf, 2014) and is often evaluated using descriptive attributes or comparisons between multiple stimuli. This subjective evaluation of the quality of spatial sound reproduction is an important measure. However, it does not necessarily correlate with a physically accurate sound field reproduction.

When the aim is to conduct hearing research, a perceptual quality measure might not be the right tool for the evaluation of the accuracy of a virtual sound scenario. Instead, certain measures need to match the listeners' performance in a reference or real environment. For example, HOA has been shown to influence sound source localization accuracy (Bertet et al., 2013; Stitt et al., 2014). Larger localization errors were found for low ambisonics orders than for higher orders. In speech intelligibility experiments, similar but not exactly matching results have been found in virtual and in real environments (Cubick and Dau, 2016; Favrot and Buchholz, 2009; Oreinos and Buchholz, 2016). Thus, further investigations on speech perception in virtual rooms are needed.

While perception in virtual sound environments has been widely investigated, the research in virtual audio-visual environments is still at its early stages. Hendrikse et al. (2018) showed that virtual talkers presented on a screen influenced head- and eye-movement behavior relative to a condition without visual information. Stecker et al. (2018) investigated the ability to detect an acoustic "odd-ball" in an virtual sound environment with additional visual source location information presented on virtual reality glasses. They showed that listeners are able to detect a single talker with incoherent room acoustic properties in a multi-talker scene. In other studies, perturbations of virtual reality glasses on the HRTFs were investigated (Genovese et al., 2018; Gupta et al., 2018). However, the effect of virtual reality glasses on perception, such as sound source localization accuracy, remains unclear.

## 1.6 Overview of the thesis

The goal of the present thesis was to reduce the gap between laboratory testing and perception in the real world, and thus to move closer towards realizing a realistic virtual cocktail party in a laboratory setting. This work focused on three main questions, which also form the three main chapters of the thesis:

- How well can an acoustic virtual room, created with state-of-the-art techniques, match a real room in terms of speech intelligibility?
- What is the relationship between the source size and speech intelligibility in spatial conditions?
- What is the role of visual information, and the impact of virtual reality glasses, on sound localization?

To address the first question, the study presented in chapter 2 set out to investigate several approaches for creating an acoustic virtual version of a room. Speech intelligibility was measured in a real room and in multiple virtual versions of the room to investigate the perceptual effects of the limitations of the reproduction techniques. The virtual rooms were captured using either room acoustic simulations or microphone array recordings, and played back over a loudspeaker array using HOA and NLM methods. To better understand the differences in speech intelligibility across environments, a computational auditory model was applied.

In chapter 3, the effect of source size of virtual sources on speech intelligibility is addressed. The physical source width is varied by applying multiple ambisonics orders. Subjects were asked to localize and rate the perceived size of the reproduced sources, and also performed a speech intelligibility experiment with both fixed and adaptive separation angles.

Chapter 4 takes a first step towards the realization of an audio-visual virtual environment. This study investigated the influence of visual information and the virtual reality hardware itself on sound source localization. First, the physical impact of the virtual reality glasses on localization cues, such as ITDs and ILDs, as well as spectral cues, was characterized. Then, localization accuracy in a real and a virtual loudspeaker environment was measured, with varying amounts of visual information presented to the listeners.

The thesis concludes with chapter 5, where a general summary of the findings, a discussion of the implications, as well as perspectives on future research on virtual environments is provided.

# 2

# Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments[a]


## Abstract

Loudspeaker-based virtual sound environments provide a valuable tool for studying speech perception in realistic, but controllable and reproducible acoustic environments. The evaluation of different loudspeaker reproduction methods with respect to perceptual measures has been rather limited. This study focused on comparing speech intelligibility as measured in a reverberant reference room with virtual versions of that room. Two reproduction methods were based on room acoustic simulations, presented either using mixed-order ambisonics or nearest loudspeaker mapping playback. The third method utilized impulse responses measured with a spherical microphone array and mixed-order ambisonics. Three factors that affect speech intelligibility were varied: reverberation, the spatial configuration and the type of the interferers (speech or noise). Two interferers were placed either colocated with the target, or were symmetrically or asymmetrically separated. The results showed differences between the reference room and the simulation-based reproductions when the target and the interferers were spatially separated but not when they were colocated. The reproduction utilizing the microphone array was most similar to the reference room in terms of measured speech intelligibility. Differences in speech


[a] This chapter is based on Ahrens A, Marschall M, Dau T (2019); Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments. Hearing Research.

intelligibility could be accounted for using a binaural speech intelligibility model which considers better-ear signal-to-noise ratio differences and binaural unmasking effects. Thus, auditory modeling might be a fast and efficient way to evaluate virtual sound environments.

## 2.1 Introduction

One of the challenges in hearing research is to understand the mechanisms involved in speech perception in complex acoustic scenarios, such as in a restaurant or at a social gathering, commonly referred to as a cocktail-party scenario (Bronkhorst, 2000; Cherry, 1953). To study the factors influencing speech perception in a given acoustic environment in a controllable and reproducible manner, virtual sound environments (VSEs) provide a valuable tool. For example, loudspeaker-based VSEs can reproduce acoustic scenes in a laboratory to investigate how the auditory system functions in realistic listening scenarios.

Using such a system, Koski et al. (2013) ccompared speech reception thresholds (SRTs) in a multi-talker scenario measured in a reference room, with corresponding SRTs measured in virtual room reproductions using microphone array recordings and directional audio coding (Pulkki, 2007). An increase of the SRT of up to 2 dB (i.e. decreased speech intelligibility) was found in some of the virtual conditions, but no significant differences appeared in the highest fidelity reproduction setup, which used up to nine loudspeakers and an anechoic reproduction room. Instead of microphone array recordings, Cubick and Dau (2016) used room acoustic simulations and a combination of higher-order ambisonics (HOA; Gerzon, 1973) and an approach to map early sound reflections to the nearest loudspeakers (NLM; Favrot and Buchholz, 2010). The setup included a target talker in the front direction and three speech-shaped noise interferers behind the listener. Speech intelligibility measurements revealed a 2 dB higher SRT in the virtual room, relative to the reference room, when using the NLM approach, and a 4 dB higher SRT when the reproduction was based on HOA. In contrast to the speech intelligibility results, classical room acoustic measures, i.e. reverberation time, clarity and interaural cross-correlation, were found to be similar in the virtual room and in the reference room, showing that these parameters are not sensitive enough to reveal differences in certain conditions. Whereas the studies of Koski et al. (2013) and Cubick and Dau (2016) used noise

as interfering signals, Oreinos and Buchholz (2016) employed seven conversational pairs of talkers distributed in a reverberant reference room. The reference room was reproduced either using a simulation-based NLM approach, as in Cubick and Dau (2016), or a HOA microphone array recording and reproduction technique. High correlations between the SRTs measured in the real and the virtual rooms were obtained between the SRTs measured in the real and the virtual rooms. However, the simulation-based NLM approach led to lower SRTs and the microphone array-based HOA approach to higher SRTs than those obtained in the reference room.

Overall, the studies of Cubick and Dau (2016), Koski et al. (2013), and Oreinos and Buchholz (2016) demonstrated that, while speech intelligibility measures in VSEs provide a reasonable correlation with corresponding measurements in the real environment, deviations remained which have not yet been resolved. The goal of the current study was to further analyze these discrepancies between real and virtual environments, as well as the differences observed across the different reproduction methods, in relation to several main factors influencing speech intelligibility. Specifically, the effects of (i) masking of different types of interferers, (ii) their spatial positions relative to the target speech signal as well as (iii) the amount of reverberation in the environment on the intelligibility of a target speech were investigated. This was done by measuring SRTs in multiple conditions. In terms of the effects of speech masking, both speech interferers with a high similarity to the target speech and speech-modulated, spectrally-matched noise interferers were considered. While the speech masker was assumed to produce some amount of informational masking (IM; Brungart et al., 2001; Watson, 2005), the noise masker was considered to produce mainly energetic masking and only little IM.

The influence of the spatial separation of the interferers was examined by considering three spatial conditions: a “colocated” condition, where the target and two interferers were presented from the frontal direction; a condition with “symmetrically separated interferers”, where the target was in the front and the interferers at ±30° azimuth; and an “asymmetric interferer condition”, where the two interferers were presented from the same location at 30° azimuth. Finally, the effect of reverberation was investigated by considering SRTs in an anechoic control condition, a reverberant reference room and virtual versions of the reference room. Three reproduction methods were considered in the present study. The reference room was either reproduced based on room acoustic

simulations and rendered using NLM or HOA, similarly to Cubick and Dau (2016), or based on impulse response measurements obtained with a HOA microphone array, as in Oreinos and Buchholz (2016). The stimuli were played back using a spherical loudspeaker array installed in an anechoic chamber.

To characterize the virtual rooms objectively, classical room acoustic measures were employed, such as the reverberation time. Furthermore, the computational speech intelligibility model of Jelfs et al. (2011) was considered to compare the predicted SRTs in the different conditions and to analyze the differences between the cues underlying speech intelligibility in the framework of the model.

## 2.2 Methods

### 2.2.1 Reference room

A standard listening room (IEC 268-13, 1985), reflecting the acoustics of a living room, with a volume of 100 $m^3$ (7.52 m*4.75 m*2.8 m) and an average reverberation time of 0.4 s, was chosen as the reference environment for this study. The wooden floor of the room is covered with a carpet, the plastered walls are partly covered with different acoustic panels and diffusors and the ceiling is fully covered with acoustic panels. The acoustical properties of the room are unknown and were estimated (find the room model estimates in the accompanying dataset, zenodo.org/record/1232317). The listening position was centered along the longest dimension of the room and 1.35 m from the back wall (see Figure 2.1). The talkers were imitated using Dynaudio BM6P (Dynaudio A/S, Skanderborg, Denmark) loudspeakers, driven by custom-made amplifiers and a RME FIREFACE 800 (Audio AG, Haimhausen, Germany) sound card. The loudspeakers were located at 2.4 m distance from the listener at 0° and at ±30° azimuth and were placed approximately at ear level (h = 1.17 m, from the floor to the center of the woofer of the loudspeaker).

### 2.2.2 Acoustic scene generation and recording

The reference room was reproduced using two alternative approaches. Room acoustics were either simulated using a commercially available acoustic simulation software, or captured by recording impulse responses using a spherical microphone array.

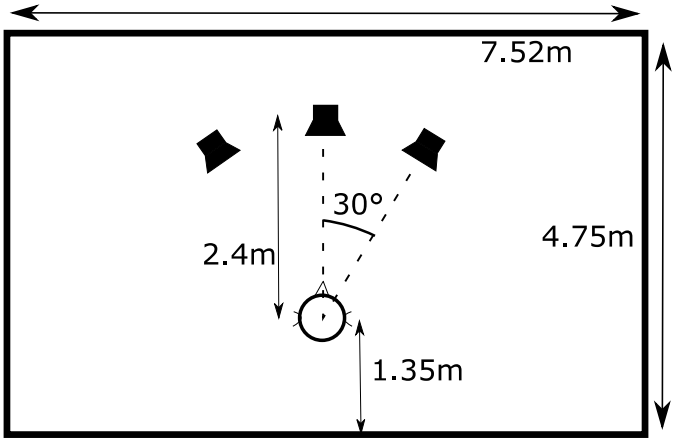



Figure 2.1: Sketch of the loudspeaker-listener configuration in the reference room. The height of the room is 2.8 m. The loudspeaker height is 1.17 m.

To simulate the acoustics of the listening room, a geometrical model of the room was constructed in the room acoustics software Odeon version 13.04 (Odeon A/S, Kgs. Lyngby, Denmark), including the same source and receiver/listener positions as in the reference room. The directivity and frequency response of the loudspeakers were incorporated in the model as in Cubick and Dau (2016). The absorption coefficients of the room surfaces were optimized from initial estimates of the surface materials, using the Odeon genetic material optimizer (Christensen et al., 2014). The optimization was performed by employing measured reverberation times (T20, T30), as well as early decay time and clarity (C7, C50, C80) parameters as calculated (ITA-toolbox; Berzborn et al., 2017) from impulse responses measured in the reference room. The details of the impulse response measurement procedure are described below. The optimized absorption coefficients did improve the room acoustics model with respect to the measurements, however the error remained larger than previously reported by Christensen et al. (2014). The reason for the larger error is likely due to the size of the room used in the current study, which is small in comparison to rooms generally modelled using Odeon. From the optimized room acoustics model, direct sound, early reflections and energy decay curves were exported in eight octave bands from 63 Hz to 8 kHz and processed using the Loudspeaker-Based Room Auralization toolbox (LoRA; Favrot and Buchholz, 2010) to obtain impulse responses for each loudspeaker in the VSE. The optimized room acoustics model can be found in a dataset (Ahrens, 2018). Two processing strategies implemented in LoRA were applied: a nearest-loudspeaker mapping (NLM) and

a mixed-order ambisonics (MOA) coding strategy. The NLM approach maps the direct sound and each of the early reflections to the geometrically closest loudspeaker. Late reflections were reproduced with energy envelopes represented in 1st order ambisonics and multiplied with uncorrelated noise for each loudspeaker (Favrot and Buchholz, 2010). For MOA, the same strategy was used for the late reflections as for the NLM. The direct sound and the early reflections were encoded using 7$^{\text{th}}$ order horizontal and 5$^{\text{th}}$ order periphonic ambisonics. The loudspeaker signals were obtained from the MOA signals using a dual-band mode matching / "max-$r_E$" decoder (Daniel, 2001), with a crossover frequency of 4 kHz.

Measurements in the reference room were undertaken with a 52-channel spherical microphone array with a radius of 5 cm (Marschall et al., 2012). Impulse responses (IRs) were recorded between the three source positions with the Dynaudio BM6P loudspeakers and the microphone array placed at the listening position. The IRs were measured using eight 16 s long logarithmic sweep signals (Müller and Massarani, 2001). The same MOA orders were used for encoding the array signals as for the simulations (7$^{\text{th}}$ order horizontal, 5$^{\text{th}}$ order periphonic). From the ambisonics components the loudspeaker signals were obtained using a dual-band mode matching / "max-$r_E$" decoder (Daniel, 2001; Marschall, 2014) as for the simulation-based reproduction and a regularization parameter of $\lambda = 0.01$ (Marschall et al., 2012).

The two room acoustic simulation-based reproduction strategies are termed "simulated NLM" and "simulated MOA" and the microphone array recording-based reproduction is termed "recorded MOA" throughout the article.

### 2.2.3 Virtual sound environment (VSE)

The virtual sound environment consists of a spherical array of 64 loudspeakers located in an anechoic chamber (7 m*8 m*6 m), with the listener's head positioned in the center of the sphere of 2.4 m radius. A depiction of the loudspeaker array can be seen in Figure 2.2. The empty anechoic chamber is considered anechoic above 100 Hz according to ISO 26101 (ISO26101, 2012). The loudspeakers are mounted on seven rings elevated by ±80°, ±56°, ±28° and 0° with respect to the head position, with 2, 6, 12 and 24 loudspeakers uniformly distributed on the respective rings.

The loudspeakers are of type KEF LS50 (KEF Audio, Maidstone, UK) and driven by three sonible d:24 amplifiers (sonible GmbH, Graz, Austria) and con-

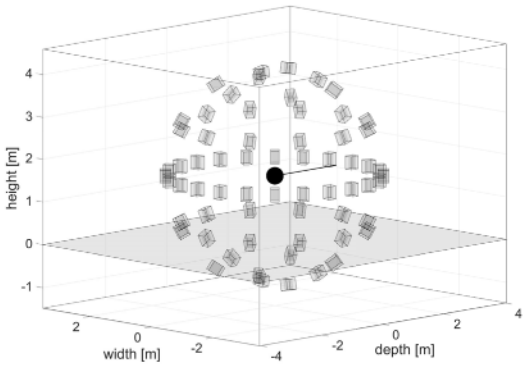



Figure 2.2: Depiction of the virtual sound environment consisting of 64 loudspeakers. The gray surface represents the wire-mesh floor and the black sphere the listening position with the facing direction indicated by the line.

trolled via two biamp TESIRA Server digital signal processing (DSP) units and sixteen TESIRA SOC-4 digital-to-analog converters (biamp Systems Inc., Beaverton, USA). Level, time, and frequency response corrections were applied using the DSP units, based on IR measurements at the midpoint of the loudspeaker array.

### 2.2.4 Room acoustic measures

Three objective room acoustic parameters were investigated and compared between the reference room and its virtual versions created with the three reproduction techniques (simulated NLM, simulated MOA and recorded MOA). Three energy parameters, reverberation time (T30), early decay time (EDT) and speech clarity (C50) were calculated from the impulse responses (IRs) measured between the three source positions and the listener position (as shown in Figure 2.1). These parameters have been shown to correlate with speech intelligibility (Bradley, 1986). An omni-directional ½inch pressure-field microphone (Type 4192, Brüel & Kjær, Nærum, Denmark) was used to acquire the room impulse responses (RIRs) to calculate the energy parameters. In the reference room, the RIRs were directly measured at the listening position using the three loudspeakers corresponding to the three source positions. In the VSE, IRs were measured from each of the 64 loudspeakers to the omni-directional microphone positioned at the center of the array, pointing upwards. Subsequently, these 64 IRs were convolved with the impulse responses generated for each loudspeaker by one of the three reproduction methods, and summed to obtain the repro-

duced RIRs. All IRs were truncated to 0.7 s. T30, EDT and C50 were calculated from the RIRs using the ITA-toolbox (Berzborn et al., 2017).

### 2.2.5 Speech intelligibility experiment

The speech material for the experiment was taken from the multi-talker version of the Dantale II matrix sentence test (Behrens et al., 2007; Wagener et al., 2003). The sentences have a five-word structure (Name, Verb, Numeral, Adjective, Noun) with low context information and ten words per word-category. The word-category "name" was presented as a call-sign and subjects were asked to identify the remaining four words on a user-interface displayed on an iPad Air 2 screen (Apple Inc., Cupertino, USA). The responses were scored on a word basis and speech reception thresholds (SRT) were measured with an adaptive procedure at 70% correct intelligibility (Brand et al., 2002). The presentation level of each of the maskers was kept constant at a sound pressure level (SPL) of 60 dB, while the level of the target speech was adjusted adaptively, starting at 70 dB SPL. The multi-talker version of the Dantale II contains five female talkers with similar voice pitch. Three of the five talkers with the closest average root-mean-square levels were selected to reduce level differences in the test (talkers 1, 4 and 5).

SRTs were measured in three spatial conditions as shown in Figure 2.3: a co-located condition with target and two interferers presented from the front, a symmetrically separated condition with the target from the front but the interferers at ±30°, and an asymmetrically separated condition with the two interferers at -30°. The difference between the colocated and the given non-colocated spatial sound source configuration is commonly considered to reflect a spatial benefit (SB). In the present study, the difference between the colocated and the symmetrical interferer configuration was defined as the SB. The difference between the symmetric and asymmetric interferer locations was considered to reflect the effect of long-term better-ear listening. The long-term better-ear listening advantage is in the current paper termed "asymmetry benefit" (AB) to clearly distinguish from short-term better-ear listening effects. Benefits for both symmetric and asymmetric interferers as compared to the colocated case are referred to as spatial release from masking in this study.

Two kinds of interfering signals were used: speech interferers using sentences spoken by different talkers from the Dantale II database, and noise interferers. To create the noise interferers, for each sentence, the broadband Hilbert

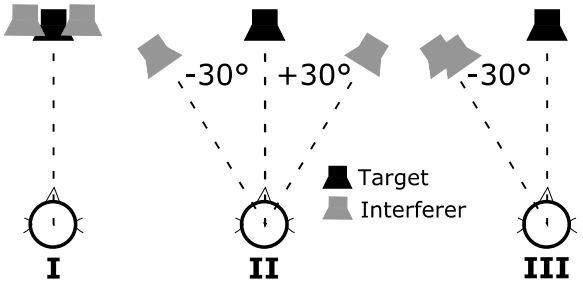



Figure 2.3: The three spatial configurations with two interfering sources colocated (I), symmetrically separated (II) and asymmtrically separated (III) with respect to the target.

envelope was extracted and low-pass filtered at 40 Hz as in Best et al. (2013) and Westermann and Buchholz (2015). Subsequently, the envelope was multiplied with a speech-shaped noise having the same long-term magnitude spectrum as the particular sentence. The speech interferer is contextually similar to the target and can be expected to produce a high amount of informational masking (IM), while the noise interferer is expected to produce less IM but has similar envelope statistics and spectral content as the speech masker (Agus et al., 2009; Best et al., 2013; Ewert et al., 2017; Westermann and Buchholz, 2015). For each SRT measurement, the call-sign (name) for the target sentence was chosen randomly and kept for the following sentences, while the three target and interfering talkers were randomly permutated for each sentence. The call-sign was shown on the user interface to the listener before the start, and continuously throughout each condition. The interfering sentences did not contain the same words as the target.

The speech intelligibility experiment was performed with ten young, normal-hearing listeners with an average age of 24.7 years ($\sigma$=4.5y) and pure-tone audiogram thresholds below 20 dB HL at the octave band frequencies between 250 Hz and 8 kHz. In addition to the previously presented reproduction conditions, a control condition was also included where the three spatial conditions (co-located, symmetrically and asymmetrically separated) were reproduced in the loudspeaker environment without reverberation (i.e. anechoic presentation using single loudspeakers). The interferers were either speech or noise. Thus, two interferer types, three spatial conditions, and five reproduction methods were tested, leading to a total of 30 conditions, with the 5 reproduction methods being: (1) reference room, (2) simulation-based NLM, (3) simulation-based MOA, (4) recording-based MOA, (5) anechoic control. The conditions were

presented in random order. Five of the ten subjects started the experiments in the reference room whereas the other five started in the VSE. Each of the conditions was repeated three times in the reference room and once in the VSE. In total, the experiments lasted about 4h for each listener. All listeners were financially compensated on an hourly basis and provided informed consent. The experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

### 2.2.6 Speech intelligibility modeling

The binaural speech intelligibility model of Jelfs et al. (2011) was used to predict the speech intelligibility data in the conditions considered in the present study. The model uses binaural room impulse responses (BRIRs) measured between the target and interferer locations and the listening position as input signals and computes the target-to-interferer ratio. The target-to-interferer ratio comprises a binaural masking level difference or binaural unmasking (BU) component and a long-term better-ear signal-to-noise ratio (BE-SNR) component. The implementation of the model was taken from the auditory modeling toolbox (Soendergaard and Majdak, 2013). The BRIRs were obtained as described above, but using a head and torso simulator (HATS, Type 4100, Brüel & Kjær, Nærum, Denmark) instead of an omni-directional microphone as for the room acoustic measures. The BRIRs were presented to the model at 0 dB SNR, i.e. with the BRIR at the target location having the same energy as the BRIRs at the interferer locations.

## 2.3 Results

### 2.3.1 Room acoustic measures

The obtained objective room acoustic measures for the reference room and the three reproduction methods are shown in Figure 2.4 for octave frequency bands. Panels A-C show the energy parameters T30, EDT, and C50, respectively. The results represent averages over the three source positions. The gray shaded area represents just-noticeable differences (JNDs) for the results obtained in the reference room. The reported JNDs for T30 and EDT are 5% (Vorländer, 1995) and 1.1 dB for C50 (Bradley et al., 1999).

The reverberation times in the simulation- and recording-based reproductions were found to match the reference room well. The results were within or close to the JNDs at most frequencies. However, at 125 Hz, the reverberation time was slightly overestimated with the two simulation-based methods whereas the recording-based reproduction led to a slight underestimation. The EDT and C50 were reproduced accurately with the recorded-MOA method, whereas differences beyond the corresponding JNDs were found with the simulation-based reproductions.

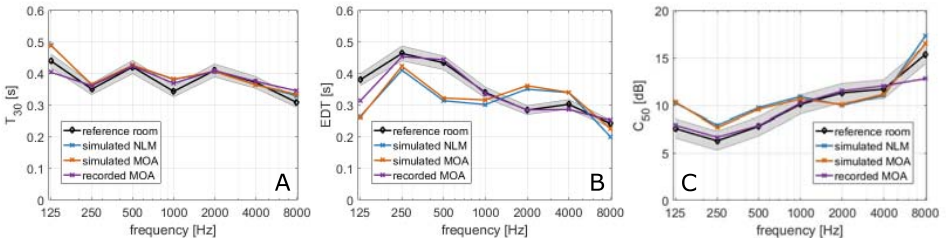



Figure 2.4: Reverberation time (T30), early decay time (EDT) and clarity (C50) in octave bands measured in the reference room and in the VSEs. The grey shaded area represents the just-noticable differences relative to the results obtained in the reference room.

### 2.3.2 Speech intelligibility

Figure 2.5 shows speech reception thresholds (SRTs, SNR at 70% correct words) in dB target-to-masker ratio (TMR). The results with the speech interferers are shown in panel A. The results obtained with the noise interferers are shown in panel B. The white, light blue and dark blue boxes represent the spatial locations of the two interfering signals: colocated, symmetrically separated and asymmetrically separated from the target, respectively. The various reproduction methods, i.e. the reference room, the three virtual rooms and the anechoic condition, are indicated on the abscissa.

To analyze the outcomes of the speech intelligibility experiment, a linear mixed effects model was fitted to the SRTs and analyzed employing an analysis of variance, using the statistics software 'R' and the step function included in the lmerTest package (Kuznetsova et al., 2014). The factors interferer location, interferer type, repetitions and reproduction method were treated as fixed effects. The factor listener was treated as a random effect, including its interactions with the fixed effects. The factor repetitions was not found to have a significant effect on the SRT [$F(2,382)=2.38$, $p=0.09$] and was removed from the final

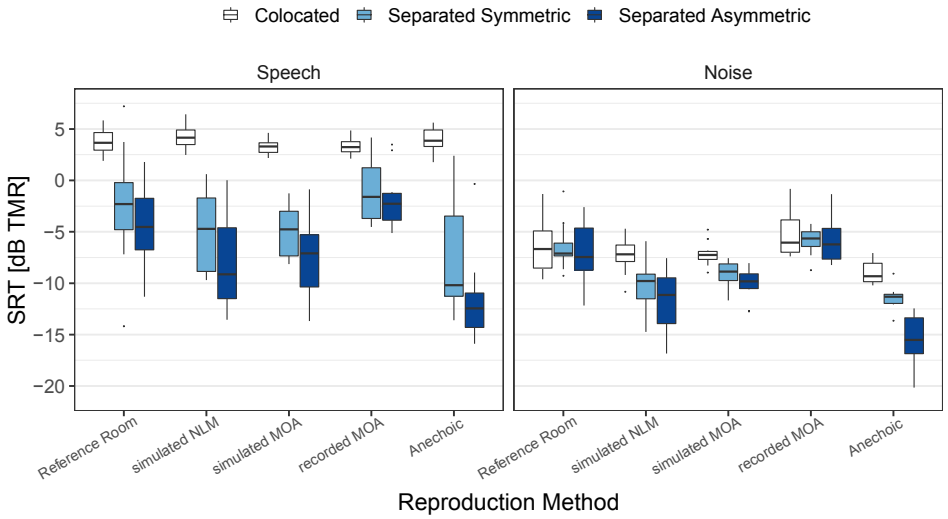



Figure 2.5: Boxplots (median and 1st/3rd quartile) of speech reception thresholds (SRTs) in dB TMR (target-to-masker ratio) with speech (left) and noise (right) interferers in the reference room (IEC listening room), the two room acoustic simulation based reproductions, the microphone array based reproduction and the anechoic condition. The results are split according to the spatial configuration of the interferers: white represents the colocated condition, light gray the symmetric and dark gray the asymmetric distribution of the two interfering talkers. (The whiskers include 1.5 times the interquartile range.)

model. The factors interferer location [$F(2,14.81)=59.4$, $p<0.0001$], interferer type [$F(1,9.02)=115.23$, $p<0.0001$] and reproduction method [$F(4,384)=91.97$, $p<0.0001$], as well as the interactions between interferer location and interferer type [$F(2,384)=146.53$, $p<0.0001$], and between interferer location and reproduction method [$F(8,384)=16.09$, $p<0.0001$], were found to be significant. The interaction of interferer type and reproduction method [$F(4,378)=1.93$, $p=0.11$] and the 3-way interaction [$F(8,370)=1.56$, $p=0.13$] were not found to be significant, but were nevertheless kept in the model because interactions on a level basis were suspected. To analyze differences between levels, a post-hoc multiple comparison analysis was performed. The post-hoc analysis was performed by contrasting least-square means using the “lsmeans” library (Lenth, 2016). Resulting p-values were corrected for multiple comparisons using the Tukey method.

### 2.3.3 Training effect and test-retest variability

The conditions measured in the reference room were repeated three times to investigate a possible training effect and the test-retest variability of the Dantale

II-based speech test. A training effect over the three repetitions could not be found [$F(2,382)=2.38$, $p=0.09$]. The test-retest variability was estimated as the standard deviation of the repetitions and averaged over conditions and subjects. It was found to be 1.5 dB and comparable to other speech intelligibility tests (Plomp and Mimpen, 1979).

### 2.3.4 Effect of reverberation on speech intelligibility

To investigate the effect of reverberation on speech intelligibility, differences between the reference room and the anechoic condition were investigated. The speech intelligibility results are shown in Figure 2.5 and the significance values of the pairwise comparisons are shown in Table 2.1. In the colocated configuration (white boxes), no influence of reverberation was found for speech while a significant effect was found for the noise interferers. In the case of the symmetrically separated interferers, reverberation resulted in an average increase of SRT by 5.4 dB for speech, and by 4.9 dB for the noise interferers. For the asymmetric interferers, the effect of reverberation was 6.9 dB for speech and 8.4 dB for the noise interferers.

### 2.3.5 Effect of reproduction methods on speech intelligibility

In the colocated configuration, no difference was found between the reproduction methods and the reference condition, neither for speech nor for the noise interferers. The significance values of all pairwise comparisons are shown in Table 2.1.

For the symmetrically separated interferers, the simulation-based reproduction methods showed statistically significant differences to the reference condition. For the speech interferers, 2.7 dB lower SRTs (better speech intelligibility) were found for both the simulated NLM and the simulated MOA methods. For the noise interferers, the SRTs decreased by 3.6 dB for the simulated NLM, and by 2.5 dB for the simulated MOA method relative to the reference condition. The SRTs obtained with the recorded MOA method were not significantly different from the reference condition, neither for the speech, nor for the noise interferers.

When comparing the two simulation-based approaches using NLM and MOA, no significant effect was observed with symmetric interferers. However, when comparing the simulation-based to the recording-based approach, signif-

Table 2.1: Statistical overview of comparisons between reproduction methods for speech reception thresholds.

| | Colocated | | Separated Symmetric | | Separated Asymmetric | |
|---|---|---|---|---|---|---|
| | Speech Interferers | Noise Interferers | Speech Interferers | Noise Interferers | Speech Interferers | Noise Interferers |
| Reference Room vs Simulated NLM | t(372)=-0.67, p=0.96 | t(372)=1.13, p=0.79 | t(372)=3.82, p=0.0015 | t(372)=5.11, p<0.0001 | t(372)=4.63, p=0.0001 | t(372)=6.51, p<0.0001 |
| Reference Room vs Simulated MOA | t(372)=0.68, p=0.96 | t(372)=0.91, p=0.89 | t(372)=3.8, p=0.0016 | t(372)=3.47, p=0.0052 | t(372)=3.97, p=0.0008 | t(372)=4.21, p=0.0003 |
| Reference Room vs Recorded MOA | t(372)=0.49, p=0.99 | t(372)=-1.74, p=0.41 | t(372)=-1.73, p=0.42 | t(372)=-1.04, p=0.84 | t(372)=-3.96, p=0.0008 | t(372)=-1.83, p=0.36 |
| Reference Room vs Anechoic | t(372)=-0.36, p=1.0 | t(372)=3.4, p=0.0066 | t(372)=7.59, p<0.0001 | t(372)=6.83, p<0.0001 | t(372)=9.68, p<0.0001 | t(372)=11.84, p<0.0001 |
| Simulated NLM vs Simulated MOA | - | - | t(372)=-0.02, p=1.0 | t(372)=-1.88, p=0.33 | t(372)=-0.54, p=0.98 | t(372)=-1.88, p=0.33 |
| Simulated NLM vs Recorded MOA | - | - | t(372)=-4.53, p=0.0001 | t(372)=-6.81, p<0.0001 | t(372)=-7.01, p<0.0001 | t(372)=-6.81, p<0.0001 |
| Simulated MOA vs Recorded MOA | - | - | t(372)=-4.51, p=0.0001 | t(372)=-4.93, p<0.0001 | t(372)=-6.47, p<0.0001 | t(372)=-4.93, p<0.0001 |

icantly higher SRTs were observed in the microphone array-recording condition. These differences were found to be 3.9 dB for the speech interferers, both in the case of the NLM and MOA reproduction. For the noise interferers, the corresponding SRT differences were 4.4 dB in the case of NLM reproduction and 3.2 dB in the case of MOA reproduction.

For the asymmetrical interferers, the simulation-based reproduction methods again showed significant differences from the reference condition. For the speech interferers, the SRTs decreased by 3.3 dB for the simulated NLM, and by 2.8 dB for the simulated MOA method, relative to the reference room. For the noise interferers, SRTs were 4.6 dB lower for the simulated NLM and 3 dB lower for the simulated MOA method than obtained in the reference room. The recording-based reproduction method did not show a significant difference to the reference with noise interferers, but with the speech interferers the SRTs increased by 2.8 dB in relation to the reference room.

The two simulation-based reproduction methods, using NLM and MOA, showed no significantly different SRTs with asymmetric interferers, with both the speech and the noise interferers. However, lower SRTs were found in the two simulation-based methods in relation to the recording-based method. The difference was about 6 dB for the simulated NLM method with both interferer types. Differences in SRT of 5.6 dB with speech and 4.3 dB with noise interferers were obtained between the simulation- and recording-based MOA methods.

### 2.3.6 Effect of spatial separation on speech intelligibility

Figure 2.6 shows the SB (light blue), i.e. the difference between the colocated and the symmetrically separated interferer condition, and the AB values (dark blue), i.e. the difference between the symmetrically and asymmetrically separated interferer conditions. The significance values of the pairwise comparisons are shown in Table 2.2. For the speech interferers, a significant SB was found in all reproduction conditions. For the noise interferers, no significant SB was found in the reference room nor for the MOA reproductions. However, a significant SB of 2.9 dB was found for the NLM reproduction and in the anechoic condition (2.5 dB).

A significant AB of 2.4 dB was found in the reference room for the speech interferers, but not for the noise interferers. Similarly, the AB effect was significant for the speech but not the noise interferers in the case of the simulated NLM and the simulated MOA methods. For the recording-based reproduction,

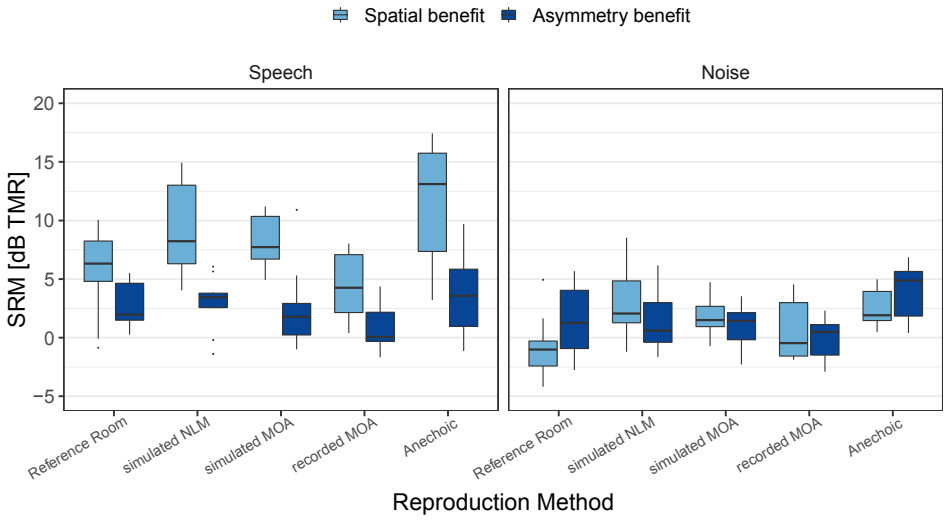



Figure 2.6: The spatial release from masking (SRM) due to separating target and interfering talkers (left) and the benefit due to asymmetric versus symmetric interferers (right) for the different reproduction methods with speech and noise interferers. (The boxes represent the median and the 1st/3rd quartile. The whiskers include 1.5 times the interquartile range.)

no AB was found for either the speech or the noise interferers. In the anechoic condition, the AB effect was significant for both speech and noise interferers.

### 2.3.7 Speech intelligibility modeling

Figure 2.7A shows the results from the simulations obtained with the Jelfs et al. (2011) model in the conditions with the symmetrically (left panel) and asymmetrically (right panel) separated noise interferers. The colocated condition is omitted because the model takes only impulse responses into consideration, thus no model outcome is seen when all sources are presented from the same location. The model outcome (squares) is shown as the sum of the two contributors, the BU (circles) and the BE-SNR (triangles). Since the BE-SNR contribution can be below zero, the total model outcome can be lower than the BU contribution. In the configuration with the interferers placed symmetrically left and right, the BE-SNR is close to zero for all reproduction methods, as expected. The model predicts the highest BU in the anechoic condition. The contribution of BU is similar, about 1 dB, in the reference room and with the recorded MOA method. The simulated NLM and simulated MOA methods show a predicted BU contribution of about 1.7 dB. For the asymmetric interferer configuration, the contribution of BU to the model output is smaller than for the symmetric

Table 2.2: Statistical overview of comparisons between reproduction methods for spatial benefit and asymmetry benefit.

| | Spatial Benefit | | Asymmetry Benefit | |
|---|---|---|---|---|
| | Speech Interferers | Noise Interferers | Speech Interferers | Noise Interferers |
| Reference Room | t(28.18)=8.83, p<0.0001 | t(28.18)=0.12, p=0.99 | t(165.96)=4.62, p<0.0001 | t(165.96)=0.91, p=0.63 |
| Simulated NLM | t(104.76)=9.34, p<0.0001 | t(104.76)=2.96, p=0.0104 | t(336.27)=3.38, p=0.0023 | t(336.27)=1.66, p=0.22 |
| Simulated MOA | t(104.76)=8.35, p<0.0001 | t(104.76)=1.93, p=0.13 | t(336.27)=2.87, p=0.0122 | t(336.27)=1.13, p=0.5 |
| Recorded MOA | t(104.76)=4.49, p=0.0001 | t(104.76)=0.59, p=0.83 | t(336.27)=0.93, p=0.62 | t(336.27)=-0.1, p=0.99 |
| Anechoic | t(104.76)=11.84, p<0.0001 | t(104.76)=2.56, p=0.0316 | t(336.27)=4.42, p<0.0001 | t(336.27)=4.58, p<0.0001 |

interferers, with values between 0.5 and 1 dB. Overall, the modeled BU is similar between the reproduction methods, except for the anechoic control condition where a contribution of 2.5 dB is predicted. The asymmetric interferer configuration was expected to result in a SNR advantage in one ear. However, the simulated BE-SNR shows values close to zero for both the reference and the recording-based MOA reproductions. The simulation-based NLM and MOA reproductions, on the other hand, show a 2 dB and 1.1 dB higher BE-SNR than the reference, respectively. The highest predicted BE-SNR of 5.5 dB was found in the anechoic condition.

Figure 2.7B shows the total model outcome together with the corresponding speech intelligibility data from the present study with noise interferers. The comparison was limited to the noise interferers over the speech interferers because the model is not able to incorporate IM. The model was fitted to the median SRT obtained in the reference room for each spatial configuration. The model captures the differences between the reproduction methods in relation to the reference room fairly well. Nevertheless, the symmetric interferer configuration (Figure 2.7, left) is not captured as well as the asymmetric interferer configuration (Figure 2.7, right).

## 2.4 Discussion

The present study investigated the discrepancies that appear between speech intelligibility tests in real and virtual environments, and the effect of various reproduction methods on these differences. Several common factors influencing speech intelligibility were varied: the spatial position and type of interferers, as well as the presence of reverberation.

### 2.4.1 The role of spatial configuration

The three spatial configurations of the interferers provided different levels of separation between the target and the interfering signals in terms of spatial cues. In the colocated condition, no such differences were available to the listener. Previous studies suggested that in a situation with similar target and interferer and no spatial separation, a positive TMR is needed for segregation, implying a level cue for selecting the target (Best et al., 2012; Brungart et al., 2001). Consequently, the reproduction method must mainly capture the sound

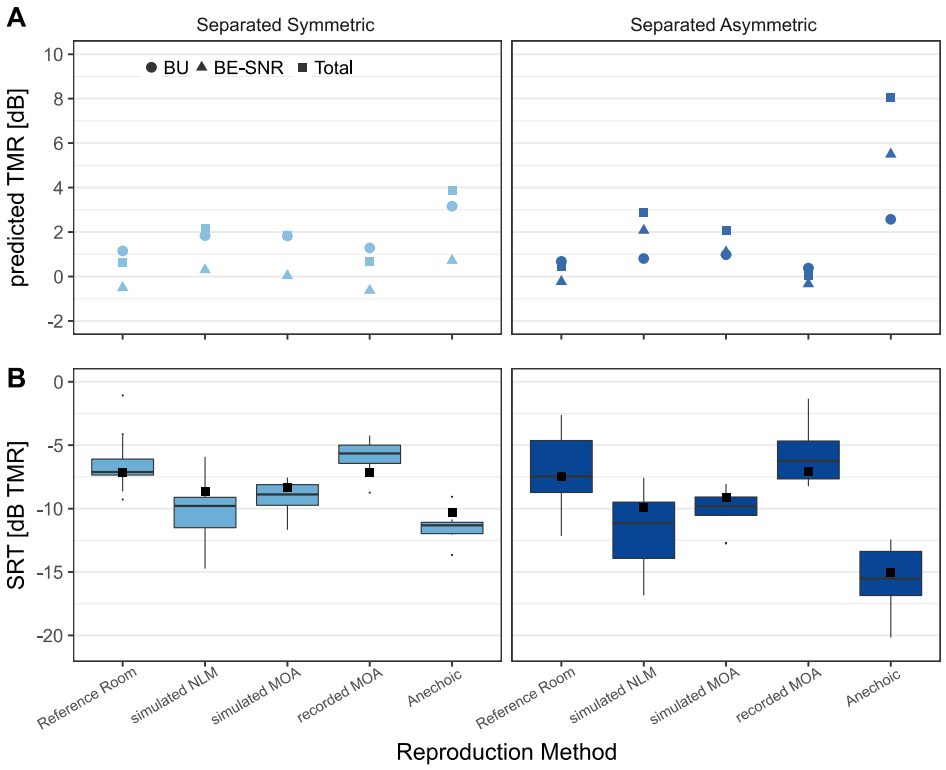



Figure 2.7: A: Model result (squares) split into binaural unmasking (BU, circles) and better-ear signal-to-noise ratio (BE-SNR, triangles) benefit for the symmetric (left) and asymmetric (right) interferer conditions. B: Speech reception thresholds (SRTs in dB TMR) measured (boxplots) and modeled (black squares). The SRTs were obtained with noise interferers. The model is fitted to the median SRT of the reference room. (The boxes represent the median and the 1st/3rd quartile. The whiskers include 1.5 times the interquartile range.)

levels of the sources to reflect speech intelligibility correctly when the interferers are colocated. Indeed, no differences between any of the reproduction methods for speech interferers were found, as the SNR was correctly reproduced. With noise interferers, reverberation does play a role, as reflected by the lower SRTs obtained in the anechoic condition compared to the reverberant reference room. However, no differences were observed between the reference and the reproduction methods for the colocated noise interferers either.

When the target and interferers are symmetrically separated, spatial location differentiates the source signals. However, due to the left-right symmetry of the interferer positions, and as long as no head movement occurs, there is no long-term SNR benefit at either ear, and the auditory system must rely on binaural cues, i.e. interaural-time differences, or short-term better-ear listening

(Brungart and Iyer, 2012; Glyde et al., 2013). This is supported by the predictions obtained with the model by Jelfs et al. (2011) (see Figure 2.7), showing a close to zero BE-SNR advantage and a main contribution of BU across all reproduction methods. Note that the model only considers a long-term better-ear advantage, and would not reflect any short-term advantage that may exist. Furthermore, the model does not take head motion into account, which might have led to intelligibility advantages during the experiments, where subjects were explicitly allowed to move their heads.

When the target and interferers are asymmetrically positioned, a long-term SNR benefit may be available at one ear. The asymmetric configuration resulted in the largest spatial release from masking overall. However, contrary to the expectation, no spatial release from masking was observed in the reference room for either of the separated spatial configurations with a noise masker, suggesting that a long-term better-ear advantage was not, in fact, available. This is in line with the model predictions (Figure 2.7), which showed a BE-SNR advantage of about 0 dB for the reference room also for the asymmetric configuration. Thus, the low amount of reverberation was sufficient to negate the effect of asymmetric positioning in terms of long-term SNR at the ears, as it was found in the anechoic condition.

In the symmetrically and asymmetrically separated configurations, differences emerge between the reproduction methods. Results from the recording-based reproduction compared favourably to the reference and a significant difference only appeared for one condition, with asymmetric speech interferers. The role of the interferer type is discussed further below. In contrast, the simulation-based reproductions led to consistently lower SRTs, or, in other words, a larger amount of spatial release from masking than in the reference room. Oreinos and Buchholz (2016) investigated speech intelligibility in VSEs in aided hearing-impaired listeners using a similar setup, but with seven conversational sources as interferers. They also found lower SRTs in their simulation-based virtual room than in the reference environment. However, the differences in that study were small and comparable to the test-retest variability of the speech test. In the current study, these differences were found to be somewhat larger, on the order of 2-3 dB, compared with the estimated test-retest variability of 1.5 dB. Despite using the same simulation framework as in Oreinos and Buchholz (2016), they considered a spatially more distributed masker configuration with a larger number of talkers, as well as the longer reverberation time in a

larger room, which might have contributed to reduced reproduction errors. The fact that lower SRTs were observed for both ambisonics and nearest-loudspeaker presentation in the present study suggests that the deviations likely originate from the room acoustics modeling rather than the playback method, as also indicated by the room acoustic measures.

### 2.4.2 The role of reverberation

Reverberation is known to reduce speech intelligibility (Duquesnoy and Plomp, 1980; Houtgast et al., 1980; Plomp, 1976), which was the case for all conditions when compared to the anechoic control, except for the condition with colocated speech interferers. Thus, the acoustics of the room had an effect on the resulting SRTs in all but one case. It follows that an accurate reproduction of the acoustics is necessary to obtain SRTs that match those measured in the reference room. The simulation-based reproduction methods resulted in lower SRTs compared to the reference and the recording-based method when the target and interferers were separated. This suggests that some aspects of the room's acoustics were not correctly captured with these methods. Indeed, the deviations apparent for the two simulation-based methods in terms of clarity, and especially early decay time (see Figure 2.4), which has been shown to be negatively correlated with speech intelligibility (Grimm et al., 2016), indicate that early reflections are not correctly reproduced by the room model. Early reflections have been shown to improve speech intelligibility (Arweiler and Buchholz, 2011; Bradley et al., 2003; Lochner and Burger, 1964; Soulodre et al., 1989), thus, it is not surprising that it is insufficient to just correctly simulate the overall reverberation time in a room. The early reflection pattern also needs to be correct in order to obtain SRTs that closely correspond to the reference room. A general challenge with the room modeling approach is that it may be difficult to obtain detailed enough information about the room (geometry, material properties, etc.) to enable such an accurate simulation. In contrast, the recording-based approach captured the detailed acoustic response of the room, at least for the measured source-receiver positions, leading to a closer match to the reference room both in terms of room acoustic parameters, as well as measured SRTs. Favrot and Buchholz (2010) showed that the changes of the room acoustic parameters due to the reproduction system itself are within the listeners' perceptual difference limens. Thus, the differences observed in the current study most likely result from inaccuracies in the room acoustic simulation.

For both simulation-based reproduction methods, the same late reverberation is reproduced using 1$^{st}$ order ambisonics. This method aims to create perceptually reasonable, but not physically accurate late reverberation. It has been shown that room acoustics parameters (e.g. EDT, T30, C50) are only affected marginally by this method (Favrot and Buchholz, 2010). Thus, it is more likely that the inaccuracies of the early reflections have the largest effect on the speech intelligibility.

### 2.4.3 The role of interferer type

Two interferer types, speech and noise, were applied to investigate any differences in the reproduction methods with respect to IM. As expected, lower SRTs were found for the noise interferers than for the speech interferers. The high SRTs with speech interferers were due to the high similarity (same sentence structure, same gender) of the speech interferers with the target speech, which leads to a high probability that target and interferes are confused. An SRM with speech interferers was found in both conditions with and without reverberation. With noise interferers, a spatial release from masking was found in the anechoic condition, but in the reverberant reference room, the spatial release from masking disappeared, both in the symmetric and the asymmetric configurations. Comparable results of a reduced or diminishing release from masking in reverberant conditions were found in previous studies (Freyman et al., 1999; Westermann and Buchholz, 2015), arguing for a spatial release from IM, which only occurs when the amount of IM is high, as in the speech interferer condition of the present study. With the noise interferers, an SB was only found in the NLM condition (and in the anechoic condition), which further suggests that the early reflections but also the diffuseness of the late reverberation in these conditions are not correctly reproduced.

### 2.4.4 The role of ambisonics reproduction

One defining feature of sound sources reproduced using ambisonics is that they have a higher spatial energy spread, i.e. a higher number of loudspeakers playing simultaneously, than the single loudspeaker used in the reference room (Gerzon, 1992; Stitt et al., 2016; Zotter and Frank, 2012). It was hypothesized that the larger energy spread could lead to reduced interaural level differences, and thus a reduced spatial release from masking, especially for the asymmetric

condition. A comparison between the two simulation-based methods, employing ambisonics versus the mapping to single loudspeakers, should reflect this effect. SRTs for NLM reproduction were indeed lower by 0.5 to 1.6 dB in the asymmetric configuration, but these differences were not statistically significant. Thus, it is unclear whether ambisonics reproduction led to a reduced AB. However, reverberation also reduces the opportunity for better-ear listening and, as discussed above, no contribution of long-term better-ear listening was found in the reference room. The fact that better-ear listening did occur for the simulation-based methods, as also predicted by the model, again indicates insufficient reverberation in these cases. Therefore, in realistic situations, where multiple sources in reverberant environments are reproduced, a reduction of a better-ear advantage due to ambisonics coding, at least at the high orders as employed in this study, is expected to be minimal, as also argued by Oreinos (2015).

The larger energy spread may explain the results in the only condition in which the recording-based reproduction differed significantly from the reference: a higher SRT was obtained with asymmetric speech interferers. Microphone array recordings suffer from low directivity at low frequencies due to physical limitations imposed by the array size (Marschall et al., 2012; Meyer and Elko, 2004), increasing the energy spread at low frequencies in the reproduced sound field. It is unclear from the current study whether the energy spread introduced by the array processing (encoding of the spherical microphone array signals, and decoding to the loudspeaker array) had a significant effect on the measured SRTs, or whether these effects were negligible considering the amount of reverberation in the room.

### 2.4.5 Choice of reproduction method

Based on the results of the study, the virtual room reproduced using microphone array recordings provided the closest overall match to the reference room in terms of measured SRTs as well as objective room acoustic parameters. Therefore, microphone array recordings appear to be the method of choice if the goal is the precise reproduction of a specific room. In contrast to the findings obtained here, Oreinos and Buchholz (2016) found slightly larger errors for their recording-based reproduction method in terms of SRTs and a beamformer benefit for aided-impaired listeners. Their conclusion was that both simulation and recording-based methods could be applied in practice, as the errors

introduced were generally smaller than the size of the effects tested. In the present study, room modeling errors appeared to be the source of the discrepancies observed with the simulation-based methods. It is possible that with further optimization of the room model, better results can be obtained for the simulation-based reproduction methods. In general, the simulations provide more control over the generated acoustic signals, and with the NLM method, some of the frequency-range limitations present in ambisonics reproduction can be circumvented (Daniel, 2001; Favrot and Buchholz, 2010; Gerzon, 1992). Thus, the simulation-based approaches may be better suited for cases where a larger degree of control is desired, and where a close matching of a particular room is not of high importance.

### 2.4.6 Limitations and perspectives

One of the limitations of this study is that only a single room was considered. Since the room acoustic parameters of the simulated virtual rooms did not match those of the real room, conclusions regarding the applicability of room acoustic simulations for the reproduction of rooms need to be taken with care. Furthermore, the considered room was small in relation to the general room size, for which the room acoustics software has been developed. Thus, future work should include various rooms with different levels of early reflections and reverberation to provide a more complete picture of the advantages and disadvantages of the room acoustic simulation and the reproduction techniques.

Only normal-hearing listeners were tested in this study in an effort to focus on a comparison between the reproduction techniques, as speech intelligibility results from hearing-impaired listeners typically show a markedly higher variance than those measured with normal-hearing listeners. As a next step, hearing-aids or other communication devices should be considered as well, as these devices might behave differently than human listeners in the generated sound fields and the processing algorithms, such as beamformers, might interact in unexpected ways with the applied reproduction methods. However, in the most important frequency range for speech up to about 6 kHz (ANSI, 2017), in which these devices typically operate, the sound field is relatively well controlled by the applied reproduction techniques, and previous work showed only a slight reduction in the efficacy of, e.g., beamforming algorithms (Cubick and Dau, 2016; Oreinos and Buchholz, 2016). Nonetheless, since one of the main application areas of VSEs is the evaluation of such communication de-

vices and their benefit to the user, the interaction between advanced processing algorithms, hearing impairment, and virtual sound environments needs to be explored further. Outcome measures other than speech intelligibility, such as listening effort, scene awareness and head-movements, as for example considered in Hendrikse et al. (2018), might also be explored, as they can be relevant for hearing-aid applications.

## 2.5 Conclusions

This study examined the accuracy of speech intelligibility measurements in a virtual sound environment (VSE) in comparison to a reference room in several conditions and with computational auditory modeling as an analysis tool. Three reproduction methods and specific factors that influence speech perception were considered: room reverberation, interferer type and spatial location of the interferers.

The reproduction based on impulse responses measured with a microphone array provided the closest match to the reverberant reference room in terms of speech reception thresholds (SRTs). The two methods based on room acoustic simulations showed significantly lower SRTs compared to the reference room, but only when target and interferers were separated, while no differences were found when target and interferer were colocated. Lower SRTs in the simulation-based reproductions could be explained by errors in the simulated early reflections, despite a correctly reproduced total reverberation time. The measured SRTs in the real and virtual rooms could be predicted using the auditory model.

Overall, it was demonstrated that room acoustic models, which are successful in capturing average properties of a room, may be limited in their ability to match the exact details of the response at a specific location, which in turn can lead to differences in measured speech intelligibility. This may only be a relevant shortcoming if capturing the response of a specific room at a specific location is crucial. If this is the case, measurement-based methods provide a clear advantage.

## 2.6 Supplementary data

Supplementary data to this article can be found online. The data from the room model can be found at zenodo.org/record/1232317 and the results from the

speech intelligibility model can be found here dx.doi.org/10.17632/2gc4bmn35p.1.

# 3

# The effect of sound source width on speech intelligibility in anechoic and reverberant environments[a]


## Abstract

Previous studies investigated the perception of the spatial size of sounds and reported an insensitivity to source size in hearing-impaired listeners, as well as enlarged sources with hearing aids and hearing aid signal processing. However, the relation between the source size and speech intelligibility remained unclear. Here, virtual sources were generated using ambisonics coding to generate sound sources with a varying source size, where high ambisonic orders lead to narrow sources and low orders to wide sources. In the first experiment, listeners were asked to estimate the location and perceived source size of a speech stimulus. In the second experiment, the spatial release from masking was measured with two interfering talkers and in the third experiment, speech intelligibility was measured in the presence of spatially varying interfering talkers while the target-to-masker ratio was kept constant. Results showed that the perceived source size did not vary with increasing ambisonics order but the spatial release from masking increased with decreasing energy spread. In accordance with these results, a wider separation between target and interfering speech sources was found to achieve equal speech perception for wide sources than for narrow sources. The speech intelligibility results were accounted for by a better-ear listening model. The modeling revealed that the spatial spread of energy limits the available signal-to-interferer ratio at the ears, due to a spread of energy from the contra-lateral interferers.


[a] This chapter is based on Ahrens A, Marschall M, Dau T (submitted).

## 3.1 Introduction

Whereas the perceived size of a visual object is directly related to the size of its retinal image (Hering, 1861; Holway and Boring, 1941), the perception of the size of an auditory object appears to be less obvious. The concept of the perceived size of acoustic sources was first discussed in the context of concert hall acoustics (see Griesinger (1997) for a review) but has since been adopted in other areas within acoustics. The perceived size of an object, often referred to as the apparent source width or the sound image size, has been shown to be affected by early reflections in a given environment and is thus related to the amount of reverberation in the environment (Blauert and Lindemann, 1986a). An increased amount of reverberation results in a decrease of the correlation between the signals at the left and right ear of a listener, i.e. a reduced interaural coherence (IC), which has been linked to larger perceived sources (e.g. Blauert and Lindemann, 1986b). In listeners with a hearing impairment, it was found that the sensitivity to changes in the physical source width is generally reduced compared to that observed in normal-hearing listeners (Whitmer et al., 2014; Whitmer et al., 2012). Other studies demonstrated that dynamic range compression in hearing aids, reflecting a level-dependent amplification scheme commonly used to compensate for loudness recruitment in hearing-impaired listeners, leads to enlarged source width percepts (Hassager et al., 2017; Wiggins and Seeber, 2011, 2012).

While there is evidence that the acoustic environment, the transmission through a device like a hearing aid, as well as effects of hearing impairment can affect human listeners' sound source width perception, only a few studies investigated how such altered spatial perception affects speech intelligibility. It has been shown that spatial differences between target speech and interferers in the horizontal plane (Duquesnoy, 1983) or in the vertical plane (Martin et al., 2012) as well as in terms of distance (Westermann and Buchholz, 2015) are advantageous for speech intelligibility relative to conditions with colocated sources. Cubick et al. (2018) investigated the effect of hearing-aid amplification on spatial release from masking in various spatial configurations of the target and the interfering speakers. In addition, they also estimated the size of the sound images of the target and the interferers in the different conditions and found larger sound images as well as a reduced spatial release from masking in the conditions with hearing aids compared to the conditions without hearing

aids. However, it has not been studied systematically the extent to which point-like sound images might be easier to perceptually segregate from spatially more diffuse sound images, and in which way they affect speech intelligibility in conditions with one or more interferers.

In the present study, the physical source size was varied using ambisonics processing, a method based on spherical harmonic decomposition (Gerzon, 1973). The higher the ambisonics order, the larger the number of spherical harmonic components, and thus, the smaller the spatial energy spread of the reproduced sources (Bertet et al., 2007; Daniel, 2001; Gerzon, 1992; Zotter and Frank, 2012). While explicit source-widening algorithms exist, such as the one proposed by Zotter et al. (2014), here the choice was made to consider the effects of ambisonic reproduction order directly, due to the potential implications for speech tests presented in virtual sound environments (Ahrens et al., 2019).

Three experiments were conducted to investigate the effects of the energy spread on speech intelligibility. Experiment 1 explored to what extent the energy spread affects the corresponding (perceived) sound image, by measuring the location and size of sound images of speech sounds as a function of their physical source size. Experiment 2 investigated if speech intelligibility is affected by the energy spread in conditions with colocated and spatially separated target-interferer configurations. In experiment 3, the minimum separation angle between the target and the interferers at a fixed level of speech intelligibility was measured for the different source sizes to test if larger separation angles are required for broader sound sources than for spatially more compact sources.

To analyze and interpret the potential perceptual cues that may contribute to the obtained speech perception results, a computational speech intelligibility model (Lavandier and Culling, 2010) was employed that includes effects of both "better-ear" listening, driven by the information represented in the left-ear vs. the right-ear signals (Zurek, 1993), as well as effects of "across-ear" processing (Wan et al., 2010), reflecting the benefit of binaural unmasking.

## 3.2 General methods

### 3.2.1 Listeners

The spatial perception and speech intelligibility experiments were performed by young (20 to 27 years) normal-hearing listeners. All listeners were native

Danish speakers and were paid on an hourly basis. Audiograms were measured for all listeners at the octave band frequencies between 250 Hz and 8 kHz. All thresholds were below or equal to 20 dB HL.

The participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). Six listeners participated in the spatial perception experiment (experiment 1), 13 in the speech intelligibility experiment with spatially distributed interfering talkers fixed in space (experiment 2) and nine listeners participated in the speech intelligibility experiment with an adaptive spatial configuration of the interfering talkers (experiment 3). The order of the experiments was randomized for each listener. Single sessions were limited to a duration of 2.5 h and the listeners were encouraged to take breaks during the sessions.

### 3.2.2 Virtual sound environment

All experiments were conducted in an anechoic chamber. The anechoic chamber was equipped with 64 KEF LS50 loudspeakers (KEF Audio, Maidstone, UK), arranged in a spherical array. In the current study, only the 24-loudspeaker horizontal ring at ear height was used. The height of the chair was individually adjusted for each listener. The 24 loudspeakers were equidistantly spaced on a 2.4 m radius (separation of 15°). The loudspeakers were driven by a sonible d:24 amplifier (sonible GmbH, Graz, Austria). The audio signals were generated in MATLAB (The Mathworks Inc., Natick, MA, USA) and fed to the amplifier via a digital audio network through Ethernet (DANTE) and two TESIRA biamp DSP units including TESIRA SOC-4 digital-to-analog converters (biamp Systems Inc., Beaverton, USA). Level, time, and frequency response corrections were applied, based on impulse response measurements at the midpoint of the loudspeaker array.

### 3.2.3 Stimuli and spatialization of sounds

The speech stimuli that were used throughout this study were taken from the multi-talker version of the Dantale II, a Danish matrix sentence test (Behrens et al., 2007; Wagener et al., 2003). The stimuli were spatialized using ambisonics reproduction on the horizontal 24-loudspeaker array. A 24-transducer setup allows for a maximum ambisonics order, M, of 11 (Gerzon, 1973). In addition to

the 11th order reproduction, 1st, 3rd and 5th order ambisonics were investigated using all 24 loudspeakers on the horizontal ring.

To examine possible spectral impairments introduced by ambisonics reproduction at off-center positions (Solvang, 2008), an optimal sub-set of $N = 2*M+2$ loudspeakers (Daniel, 2001) was investigated for $1^{st}$, $3^{rd}$ and $5^{th}$ order ambisonics. However, since no significant differences in results between using the full- and the sub-set of loudspeakers was found, only the full-set results are presented here.

The loudspeaker signals were generated using a dual-band decoder with a cross-over frequency at $M*700$ Hz (Favrot and Buchholz, 2010). Below the cross-over frequency, basic ambisonics decoding was used and above "max-$r_E$" decoding (Daniel, 2001). The loudspeaker signals were presented to the listeners anechoically (direct sound only) and including simulated reverberation from a small, living room type room (IEC listening room, IEC 268-13 (1985)) with a volume of 100 $m^3$ and a reverberation time of about 0.4 s. The room was modeled using the room acoustics simulation software Odeon (Odeon A/S, Lyngby, Denmark) and is available online (Ahrens, 2018). The loudspeaker signals were generated using the LoRA toolbox (Favrot and Buchholz, 2010). Since only loudspeakers in the horizontal plane were employed, the elevated reflections were mapped to the horizontal plane. The simulated sources were placed at a distance of 2.4 m and therefore coincided with the distance of the loudspeaker array.

Ambisonics decoding at different orders can lead to variations in the frequency response, due to the different decoder crossover frequencies as well as due to spectral colorations when more than $2*M+1$ loudspeakers are used (Solvang, 2008). To reduce the influence of spectral colorations on the experimental outcomes, equalization filters were designed to achieve equal frequency responses as measured at the center of the loudspeaker array. The filters were designed to match the direct sound (anechoic) frequency response of the $11^{th}$ order ambisonics reproduction. The reverberant impulse responses were equalized with the same filters as the anechoic impulse responses. Subsequently, the impulse responses for both anechoic and reverberant conditions were set to unity gain of the direct sound. Thus, the reverberant condition was perceived as somewhat louder than the anechoic condition, while the source levels remained equal.

The physical source size of the virtual sources that were reproduced using

ambisonics can be described using the ambisonics energy vector, $r_E$ (Daniel, 2001; Gerzon, 1992). The angular energy spread is defined as the inverse cosine of the length of the energy vector (Bertet et al., 2013; Daniel, 2001; Zotter and Frank, 2012). For an infinite ambisonics order, the energy vector is equal to one, i.e. the energy spread is zero. For lower orders, the length of the energy vector is reduced from one and the energy spread increases. Figure 3.1 shows the ambisonics panning function of the ambisonics orders considered in the current study. The arrow indicates the length of the energy vector which can be related to the physical energy spread in degrees, indicated by the cross (Zotter and Frank, 2012). The length of the energy vector has been shown to correlate with the perceived source width (Frank, 2013). The ambisonics panning function was calculated and plotted using the spherical array processing toolbox (Politis, 2016).

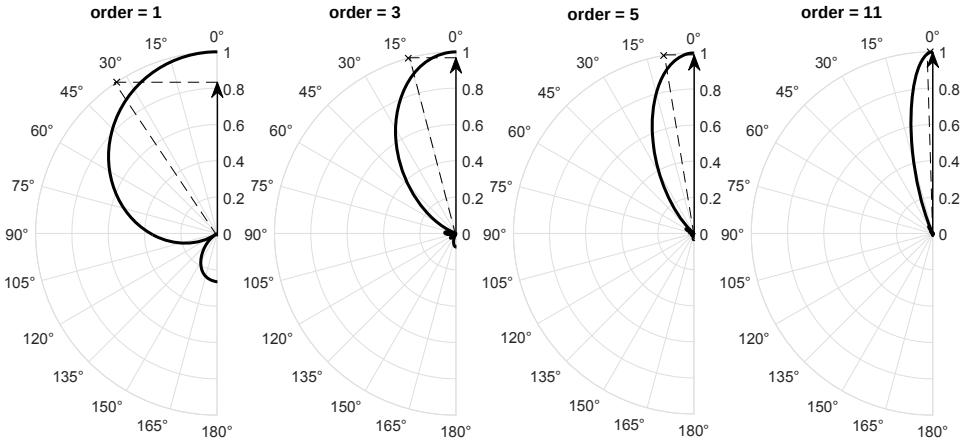



Figure 3.1: Ambisonics panning function of different orders. The arrow indicates the length of the energy vector ($r_E$) and the cross the corresponding energy spread in degrees (calculated as the inverse cosine of the length of the ambisonics energy vector).

### 3.2.4 Statistical Analysis

The results obtained in the three experiments were analyzed employing linear mixed-effects models using the statistics software R and the step function included in the lmerTest package (Kuznetsova et al., 2014). If post-hoc analyses of within-factor comparisons were performed, the “emmean” package was used to estimate marginal means from the mixed-effects linear models (Lenth, 2016). The p-values are reported including Bonferroni significance corrections.

## 3.3 Experiment 1: Measures of sound image location and size as a function of the energy spread

### 3.3.1 Methods

The listeners were asked to localize a single sound source and to judge the size of the perceived sound source. This was done by indicating the location and size of the perceived sound image on the touchscreen of an Apple iPad Air 2 (Apple Inc., Cupertino, CA, USA). Figure 3.2 shows the user interface as shown to the listeners. To indicate the location of the sound image, the listeners were asked to place a cross at the desired location with a finger on the touchscreen. To indicate the size of the sound image, the listeners could vary the size of a circle around the cross by moving a finger closer to the origin or further away from it, as in Hassager et al. (2017). The initial radius of the source size was semi-random to reduce a potential bias. If multiple sound images ("split images") were perceived by the listeners, two or more circles could be placed on the user interface. The listeners were instructed that sound images could be placed at any location and distance from the origin, i.e. also at positions closer to the listener than the loudspeaker ring or further away from it.

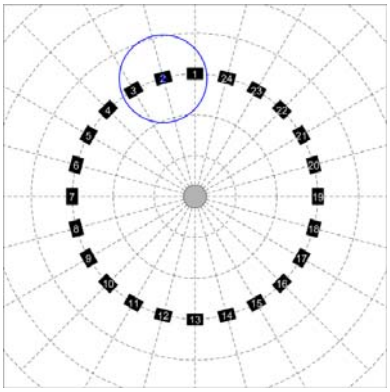


Figure 3.2: Screenshot of the user interface (UI) from the spatial perception experiment. The grey circle in the center depicts the listeners' position, and the black boxes the loudspeaker locations. The numbers in the UI correspond to numbers displayed on the loudspeakers.

The sound sources were generated using different ambisonics orders and in conditions with and without simulated reverberation. A single sentence from the Dantale II database was used as a speech stimulus presented in a speech-modulated noise (SMN) background. The SMN had the same long term spectrum and broadband envelope as the speech sentence but with random

phase (Ahrens et al., 2019; Best et al., 2013; Westermann and Buchholz, 2015). The stimuli were presented from the front and from 15° azimuth to the right. Each condition was repeated three times, leading to 96 trials for each listener. The listeners were allowed to listen to each sound repeatedly before indicating the position and size of the sound image. Additionally, a reference sound was presented to the listeners, providing an anchor with the minimum energy spread. The reference stimulus was generated using the same stimulus (speech sentence or speech modulated noise sentence) as the target but was presented anechoically from a single loudspeaker in the front of the listeners.

To calculate the source image size, the two tangents between the listeners' position (center in Figure 3.2) and the sides of the drawn source circles (blue circle in Figure 3.2) were calculated. The angle between the tangents, i.e. the angular width of the source circle as viewed from the listeners' position, was defined as the source image size. Applying this analysis method results in a larger source image for a source perceived close to the listener than for a source perceived further away. This measure was considered instead of the radius or area of a perceived source because the listeners were asked to judge the source size relative to the loudspeakers.

### 3.3.2 Results and discussion

Figure 3.3 shows the responses of the listeners obtained for the four ambisonics orders M = 1 (red, upper left panel), 3 (green, upper right panel), 5 (blue, lower left panel) and 11 (cyan, lower right panel), with both stimulus types presented from the front and the lateral direction. Each semitransparent circle represents a single response. The size of the sound images did not seem to vary much across the conditions with different ambisonics order. However, the position of the sound images was generally considered to be closer to the listener for low ambisonic orders than for the higher orders.

Figure 3.4 shows the perceived distance as a function of the ambisonics order in the anechoic (light blue) and the reverberant (dark blue) conditions. The statistical analysis showed significance for all main effects [order: $F(3,561)=9.2$, $p<0.0001$; stimulus type: $F(1,561)=4.1$, $p=0.0442$; direction: $F(1,561)=4.1$, $p=0.0428$; reverberation: $F(1,561)=210.9$, $p<0.0001$] as well as the interaction between the ambisonics order and the reverberation condition [$F(3,561)=10.9$, $p<0.0001$]. Thus, for the low orders, the anechoic sources were perceived to be closer to the listener than for the high orders. In fact, only the 11th order condition was

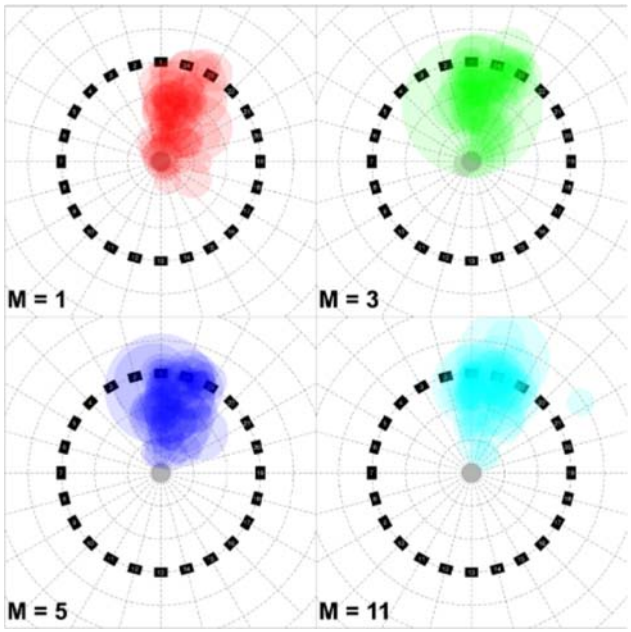



Figure 3.3: Representation of the responses in the spatial perception experiment with the speech and noise signals in the anechoic condition for sources from both 0° and 15° at the right. The signals were reproduced using the ambisonics orders (M) as shown in the subfigures.

not perceived to be significantly further/closer to the actual loudspeaker distance at 2.4 m [t(6.81)=-0.1, p=0.09], while all other orders differed significantly from the actual distance [p<0.0167]. In the reverberant condition, none of the ambisonics orders led to a perceived distance that was significantly different from the actual loudspeaker distance [p>0.69].

Figure 3.5 shows the size (angular width) of the sound images as a function of the ambisonics order. Sound images that were perceived to be very close to the listener or inside the head were not considered in the analysis. A linear mixed model was fitted to the sound image size, where the ambisonics order, the stimulus type, the source location and the reverberation condition were treated as fixed effects and the listeners, as well as the interaction between the listeners and the fixed effects, were treated as random effects. The analysis of the model revealed that only the interaction between the ambisonics order and the reverberation condition contributed significantly to the model [F(3,525.07)=3.5, p=0.0149], while the main effects as well as all other interactions were not significant.

No differences between the sound image sizes obtained for the different ambisonics orders were found, even though larger sound images were expected

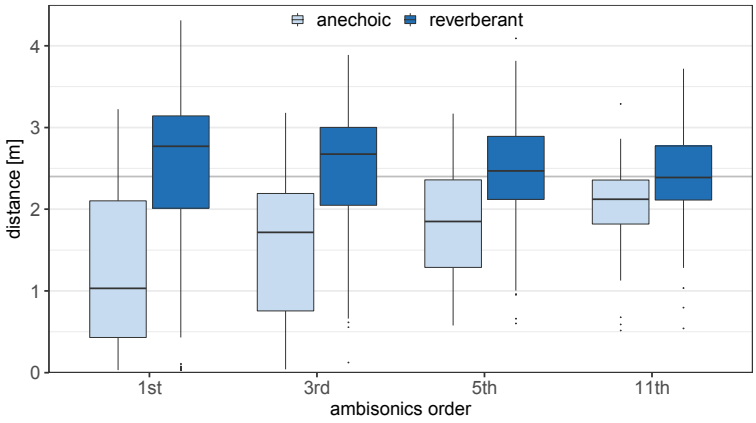



Figure 3.4: Perceived distance of the speech and noise sources in the spatial perception experiment. The distance is defined as the distance between the listener position and the center of the circle placed by the listener. The boxplots indicate the median and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

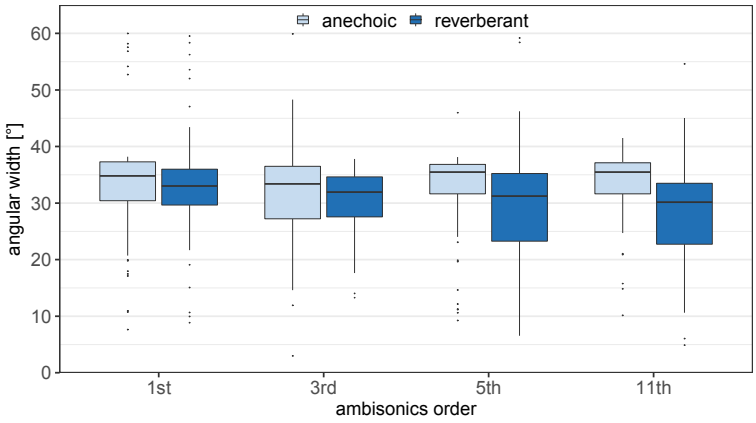



Figure 3.5: Sound image size (angular width) in the spatial perception experiment, calculated as the angular width of the reported source circle from the listener's position. The boxplots indicate the median and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

for low orders as the physical energy spread is larger with low orders, as shown in Figure 3.1. However, the results of the present experiment did show an effect of ambisonics order on the perceived distance. In the anechoic condition, listeners perceived the low ambisonics order stimuli to be closer than the higher order stimuli. In contrast to the current study, Frank (2013) found that the energy spread was highly correlated with the perceived source width when sound was presented over pairs or triplets of loudspeakers at various opening angles. Similarly, Bertet et al. (2013) observed a larger localization blur with

low ambisonics orders, which has also been argued to be a measure of the source width percept (Blauert, 1997). However, these studies were conducted in listening rooms that were not anechoic, and subjects were not asked to consider source distance. Without reverberation, the direct-to-reverberant ratio, which is a major cue for distance perception, is not available (Zahorik et al., 2005). Thus, in the absence of alternate distance cues, listeners might have interpreted the wider spread of energy as a cue for the source distance instead of source size. In the simulated reverberant conditions, listeners appear to have perceived the sources at the correct distance on average, but no effect of the energy spread on sound image size emerged. As a measure, the energy spread only considers the distribution of the direct sound across loudspeakers, even though early reflections have been shown to contribute to the source width percept (Barron and Marshall, 1981; Griesinger, 1997). It is possible that the change of energy spread of the direct sound was insufficient to elicit a perceived change in the sound image due to the presence of the early reflections. Additionally, the incongruence between the audio (small reverberant room) and visual stimulus (large anechoic chamber) may have made the judgements more difficult (Gil-Carvajal et al., 2016).

## 3.4 Experiment 2: Speech intelligibility with two interfering talkers fixed in space

### 3.4.1 Methods

Experiment 2 investigated the influence of the energy spread on speech intelligibility. The speech material of the target and two interfering talkers was taken from the multi-talker version of the Danish matrix sentence test Dantale II (Behrens et al., 2007). Dantale II sentences have a name-verb-numeral-adjective-noun structure. The name was presented as a call-sign and the listeners were asked to identify the remaining four words on a user interface displayed on an iPad Air 2 (Apple Inc., Cupertino, CA, USA) touch screen. The call-sign was continuously shown on the user interface. For each word category, ten words exist in the speech test and are shown as possible response alternatives. The responses were scored on a word basis and speech reception thresholds (SRTs) were measured with an adaptive procedure converging at 70% correct intelligibility (Brand et al., 2002). The sound pressure level (SPL) of the maskers

was kept constant at 60 dB, while the level of the target speech was adjusted adaptively, starting at 70 dB. The speech material contained five female talkers with a similar voice pitch. However, only three talkers (talkers 1, 4, 5) were chosen because the average level of the two other talkers differed strongly.

SRTs were measured in two spatial configurations: a colocated condition with the target and two interfering talkers presented from the front, and a separated condition with the target from the front and the two interferers presented from ±15° azimuth. For each SRT measurement, a call-sign (name) was chosen randomly and kept for all sentences while the three talkers representing the target and interfering sources were chosen randomly for each sentence.

Each listener was introduced to and familiarized with the task by presenting five to ten sentences in quiet. SRTs were then measured in the conditions with the different ambisonics orders, with and without reverberation, and with colocated and separated interferers, leading to 28 (7x2x2) SRT measurements overall. The conditions were presented in random order to the listener.

The model of Lavandier and Culling (2010) was used to predict the SRTs in the corresponding conditions. The model uses the binaural impulse responses (BIRs) measured between the listening position and the target and the interferer locations, convolved with speech-shaped noise. The BIRs were measured in the center of the loudspeaker array using a B&K Head and Torso Simulator (Type 4128-C; Brüel & Kjær A/S, Nærum, Denmark). From the binaural input, the model calculates a long-term better-ear SNR and a binaural unmasking component, based on the binaural masking level difference (BMLD) model from Culling et al. (2005). In the current study, since the BIRs were convolved with the speech shaped noise, no additional frequency weighting was applied in the model.

### 3.4.2 Results and discussion

Figure 3.6 shows results from the speech intelligibility experiment for the anechoic condition (top panel) and the reverberant condition (bottom panel) with spatially colocated (white boxes) and separated (blue boxes) interferers. The statistical analysis of the SRTs revealed significant main effects [order: $F(3,186)=10.8$, $p<0.0001$; interferer configuration: $F(1,186)=321.8$, $p<0.0001$; reverberation: $F(1,186)=51.4$, $p<0.0001$] as well as significant interactions between ambisonics order and interferer configuration [$F(3,186)=10.1$, $p<0.0001$] and between reverberation and interferer configuration [$F(1,186)=22.1$, $p<0.0001$].

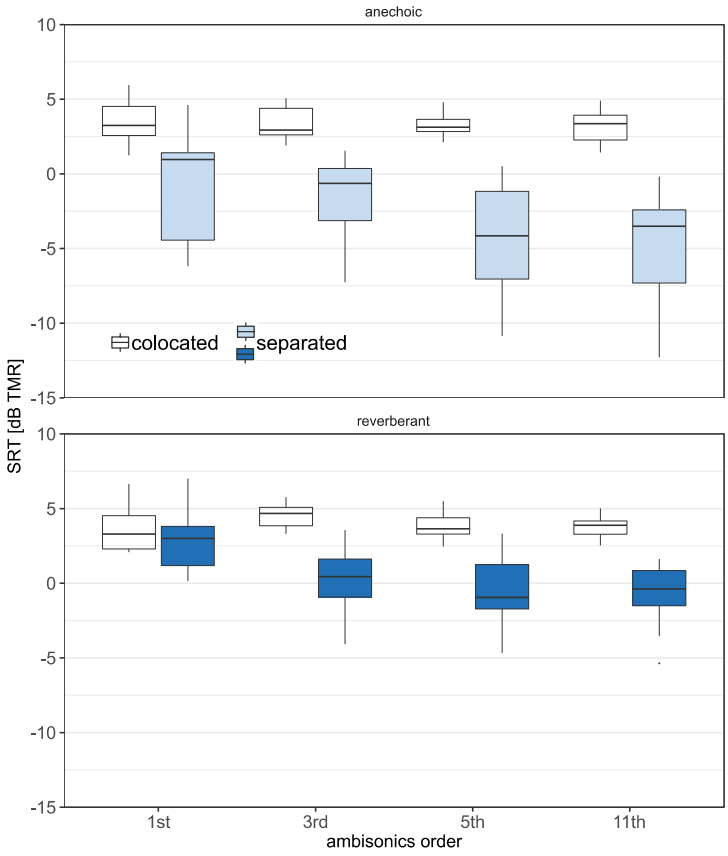



Figure 3.6: Speech reception thresholds (SRT) at 70% correct as target-to-masker ratio in dB with two colocated (white boxplots) or two symmetrically separated interferers (blue boxplots). The top panel represents the anechoic condition and the bottom panel the reverberant condition. The boxplots indicate the median and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

In the colocated interferer configuration, no differences were found between the ambisonics orders [p=1]. Similarly, no effect of reverberation was found when the target and the interfering talkers were colocated [t(186)=-1.7, p=0.17]. These findings are consistent with previous work with this speech material (Ahrens et al., 2019). It has been argued that a positive target-to-masker ratio (TMR), i.e. a target speech with a higher level than each interferer, is needed to segregate the sources in situations with similar target and interfering speech material and no spatial separation (Best et al., 2012; Brungart et al., 2001).

Further analysis was performed on the difference between the colocated and the separated interferer configurations, i.e. the spatial release from masking

(SRM). Figure 3.7 shows the SRM obtained in the anechoic (light blue boxes) and the reverberant (dark blue boxes) condition as a function of the ambisonics order. The analysis of the linear mixed model with the ambisonics order and the reverberation condition as fixed effects and the listeners as random effect revealed significant contributions of both main effects [order: $F(3,87)=12.4$, $p<0.0001$; reverberation: $F(1,87)=27.1$, $p<0.0001$] but no interaction [$F(3,84)=1.7$, $p=0.17$]. The post-hoc analysis between the orders is shown in Table 3.1 and revealed that the SRM is largest for the 11$^{th}$ order and decreases with decreasing order. The ambisonic presentation order, and thus the energy spread, clearly affects the SRM.

Table 3.1: Results from the post-hoc analysis of the spatial release from masking (SRM). Non-significant results with p-value larger than 0.05 are indicated in grey.

| | 1$^{st}$ order | 3$^{rd}$ order | 5$^{th}$ order | 11$^{th}$ order |
|---|---|---|---|---|
| 1$^{st}$ order | - | t(87)=-2.9, p=0.0268 | t(87)=-4.9, p<0.0001 | t(87)=-5.5, p<0.0001 |
| 3$^{rd}$ order | | - | t(87)=-2.0, p=0.3159 | t(87)=-2.6, p=0.0617 |
| 5$^{th}$ order | | | - | t(87)=-0.7, p=1.0 |
| 11$^{th}$ order | | | | - |

Figure 3.7 also shows the predictions of the SRM data obtained with the auditory model. The squared markers indicate the better-ear listening component, which was fitted to the median SRM values in the 11$^{th}$ order conditions. The binaural unmasking component (BMLD) is depicted as circles and is presented relative to the BMLD obtained for 11$^{th}$ order. Both fittings were done separately for the conditions with and without reverberation.

The better-ear component of the model captures the observed trend of an increasing SRM with increasing ambisonics order, while the binaural unmasking component is approximately constant with respect to the ambisonics order. Thus, the SRM results can be predicted by the better-ear SNR component of the model alone. This suggests that the reduced SRM observed for sources presented at a lower ambisonics order, i.e. with a wider spread of energy, is due to a reduced better-ear SNR advantage. However, it is not clear whether this reduced SNR advantage is related to the spatial position of the source, i.e. whether speech intelligibility can be restored by increasing the source-target separation. This was considered in the following experiment, where the target-masker separation angle was investigated.

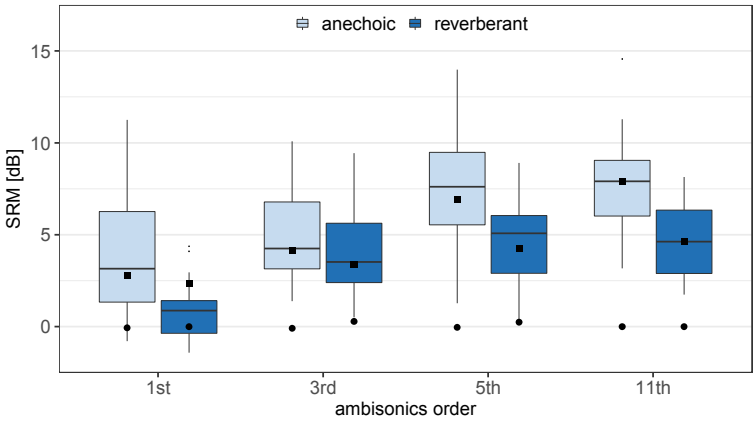



Figure 3.7: The measured spatial release from masking (SRM, boxplots) and modeled better-ear listening component (squares) and binaural masking level difference component (circles) in the anechoic (light blue) and the reverberant (dark blue) condition. The boxplots indicate the median and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

## 3.5 Experiment 3: Speech intelligibility with two interfering talkers varying in space

### 3.5.1 Methods

Experiment 3 investigated speech intelligibility of target sentences from the front direction in the presence of spatially varying interfering talkers. This was done for a fixed TMR of -6 dB, and for the same ambisonics orders ($1^{st}$, $3^{rd}$, $5^{th}$ and $11^{th}$ order) corresponding to different degrees of energy spread as described above. With the fixed TMR, the separation angle of two symmetrically separated interferers was measured to obtain 70% speech intelligibility (speech reception angle; SRA). The particular TMR was chosen based on pilot testing, and set to obtain a reasonable range of angles, avoiding ceiling and floor effects. The speech material was the same as in experiment 2 where the interferers had fixed spatial locations. The SRA was measured using an adaptive procedure as described in Brand et al. (2002). The separation angle of a specific trial was calculated using the same procedure as is used to obtain the level in traditional SRT measurements (Brand et al., 2002). The change in separation angle ($\Delta\theta$) of the subsequent trial was defined as:

$$\Delta\theta = -\frac{f(i)\times(prev-tar)}{slope},$$

where i is the reversal number, prev refers to the discrimination value of the previous sentence, and tar to the discrimination value to which the procedure converges.

The parameters $f(i)$ and $slope$ were adapted from the recommendations provided by Brand et al. (2002) to account for the fact that the separation angle was used here as a tracking variable instead of the SNR. This was needed because speech intelligibility is less sensitive to a change in separation angle than a change in SNR (Rønne et al., 2017). A slope parameter of 0.029 degrees$^{-1}$ and an $f(i) = 1.5 \times 1.15^{-i}$ were used to obtain the different step sizes.

The range of separation angles was limited to where speech intelligibility was expected to be a monotonic function of separation angle. Since the lowest SRM has commonly been found at 0° separation (i.e., no separation between target and interferers) and the largest at a separation angle of 110-120° (Bronkhorst, 2000), the range of angles was set to 0°±105°.The initial separation angle between the target and the interferers was 75°. Each listener repeated each condition twice.

### 3.5.2 Results and discussion

Figure 3.8 shows the angle between the target and the two symmetric interferers that is needed to identify 70% of the words correct (SRA). The statistical analysis of the SRA revealed a significant effect of the ambisonics order [F(3,67)=11.6, p<0.0001] but not of the repetitions [F(1,66)=1.3, p=0.26] and their interaction [F(3,63)=1.2, p=0.32]. The result of the post-hoc analysis is shown in Table 3.2. Generally, smaller SRAs were found for the higher the ambisonics orders. However, when comparing 1$^{st}$ vs. 3$^{rd}$ order, as well as 5$^{th}$ vs. 11$^{th}$ order, no significant differences were found.

Table 3.2: Results from the post-hoc analysis of the speech reception angle (SRA). Non-significant results with p-value larger than 0.05 are indicated in grey.

| | 1$^{st}$ order | 3$^{rd}$ order | 5$^{th}$ order | 11$^{th}$ order |
|---|---|---|---|---|
| 1$^{st}$ order | - | t(67)=-1.2, p=1.0 | t(67)=3.4, p=0.0074 | t(67)=3.6, p=0.0043 |
| 3$^{rd}$ order | | - | t(67)=4.6, p=0.0001 | t(67)=4.8, p=0.0001 |
| 5$^{th}$ order | | | - | t(67)=0.2, p=1.0 |
| 11$^{th}$ order | | | | - |

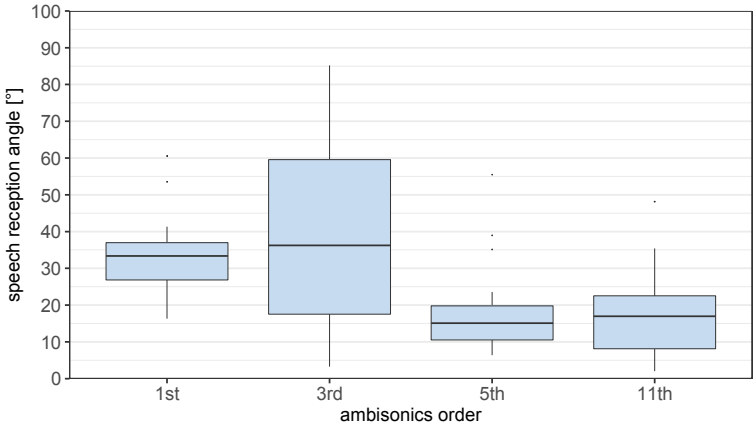



Figure 3.8: Speech reception angle (SRA), i.e. separation angle between the target and two symmetrically spaced interferers that leads to 70% intelligibility, at −6 dB target-to-masker ratio in the anechoic condition. The boxplots indicate the median and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

The results show that the changes in speech intelligibility due to the varying energy spread do relate to the spatial position of the sources: sources with a larger energy spread require a larger angular separation for equal intelligibility. The outcomes are consistent with results from (Lőcsei et al., 2017), who measured the interaural time difference needed to understand 50% of the words presented 3 dB below individually measured SRTs with two colocated interferers. Their results varied between 140 and 370 µs which corresponds to about 15 to 45° azimuth location as measured on an artificial head (e.g. Oreinos and Buchholz, 2013). These angles correspond to the SRA found in the current study for sources reproduced with higher ambisonics orders (narrow source width).

## 3.6 Overall discussion and summary

In the present study, three experiments were conducted to investigate the effect of source size on speech perception. In experiment 1, it was shown that a wider energy spread elicited by ambisonics processing did not lead to perceptually wider sources. Instead, sources were perceived as being closer in distance when presented anechoically. In experiment 2, a lower SRM was found for sources with a wider energy spread than for narrow sources. Simulations using a binaural auditory model suggest that the underlying mechanism is a reduced opportunity for better-ear listening when the energy spread is wide. In the third experiment,

the minimum separation angle between a target speech and interfering speech sources in terms of speech intelligibility was found to be correlated with the energy spread. For equal speech intelligibility, a wide separation was needed for sources with a large energy spread, and a smaller separation was needed for sources with a low energy spread.

In the current study, ambisonics processing was employed to generate sources with varying energy spread as a way to simulate sound sources of varying physical source size. However, varying the ambisonics order does not only control the energy spread, but also introduces varying magnitude and phase errors at higher frequencies due to different frequency range limitations for different orders (Daniel, 2001). While equalization and dual-band decoding were used to reduce these errors, the sound field at the ear positions of the listeners may have differed in other aspects than purely the energy spread of the sources. This, in turn, may have resulted in speech intelligibility degradations that were not related to spatial attributes. While such contributions cannot be excluded in the present framework, the modeling results suggesting an underlying better-ear cue, as well as the results from experiment 3, which demonstrated a significant effect of spatial separation on speech intelligibility, together imply a key role of the spatial properties of the sources.

The percept of the source width has previously been described with binaural features such as the interaural coherence or fluctuations of interaural time differences (Griesinger, 1997; Mason et al., 2001; Whitmer et al., 2012). In the current study, these measures were not considered directly; instead, a binaural speech intelligibility model was used that took binaural unmasking into account, i.e. long-term interaural time differences and interaural coherence. However, the binaural unmasking component was not found to vary with the physical energy spread, and, consistently, no perceptual differences in sound image size were found. In contrast, the varying energy spread did affect speech intelligibility, which the modeling revealed was through a reduction of the better-ear SNR advantage. Thus, the perception of the size of a source and its intelligibility in spatial settings seem to be driven by different cues. While a larger spread of energy may not necessarily lead to a wider perceived sound image, it can still decrease speech intelligibility with spatially separated interferers. This implies that any processing which influences the spatial spread of energy, for example through the signal processing in hearing aids or a low-order reproduction in ambisonics-based virtual sound environments, can lead to degraded speech

intelligibility.

# 4

# Sound source localization with varying amount of visual information in virtual reality[a]


## Abstract

To achieve accurate spatial auditory perception, subjects typically require personal head-related transfer functions (HRTFs) and the freedom for head movements. Loudspeaker-based virtual sound environments allow for realism without individualized measurements. To study audio-visual perception in realistic environments, the combination of spatially tracked head mounted displays (HMDs), also known as virtual reality glasses, and virtual sound environments may be valuable. However, HMDs were recently shown to affect the subjects' HRTFs and thus might influence sound localization performance. Furthermore, due to limitations of the reproduction of visual information on the HMD, audio-visual perception might be influenced. Here, a sound localization experiment was conducted both with and without an HMD and with a varying amount of visual information provided to the subjects. Furthermore, interaural time and level difference errors (ITDs and ILDs) as well as spectral perturbations induced by the HMD were analyzed and compared to the perceptual localization data. The results showed a reduction of the localization accuracy when the subjects were wearing an HMD and when they were blindfolded. The HMD-induced error in azimuth localization was found to be larger in the left than in the right hemisphere. Presenting visual information of hand-location and


[a] This chapter is based on Ahrens A, Lund KD, Marschall M, Dau T (2019); Sound source localization with varying amount of visual information in virtual reality. PLoS ONE 14(3): e0214603.

room dimensions showed better sound localization performance compared to the condition with no visual information. When visual information of the limited set of source locations was provided, the localization error induced by the HMD was found to be negligible. Also adding pointing feedback in form of a virtual laser pointer improved the accuracy of elevation perception but not of azimuth perception.

## 4.1 Introduction

Virtual environments (VE) and virtual reality (VR) systems enable the study of audio-visual perception in the laboratory with a higher degree of immersion than obtained with typical laboratory-based setups. Head-mounted displays (HMDs) may allow the realistic simulation of visual environments, and loudspeaker-based virtual sound environments can reproduce realistic acoustic environments while maintaining the subjects' own head-related transfer functions (HRTFs). Combining HMDs and loudspeaker-based virtual sound environments could, therefore, be valuable for investigating perception in realistic scenarios.

To localize a sound source in the horizontal plane (azimuth) as well as in the vertical plane (elevation; see Blauert (1997) for a review), three major cues are crucial: interaural time differences (ITDs), interaural level differences (ILDs), and monaural spectral cues generated by reflections from the body and the pinnae. An alteration of these cues can lead to a reduced localization accuracy (e.g. Hofman et al., 1998; Shinn-Cunningham et al., 1998a,b). Wearing an HMD alters the sound localization cues (Genovese et al., 2018; Gupta et al., 2018), as well as the perceived spatial quality (Gupta et al., 2018) and might also reduce the sound source localization accuracy.

Another factor that can affect people's sound source localization ability is visual information. The ventriloquism effect describes the capture of an acoustic stimulus by a visual stimulus (Howard and Templeton, 1966; Recanzone, 2009), altering the perceived location of the acoustic stimulus.

Maddox et al. (2014) showed that the eye gaze modulates the localization accuracy of acoustic stimuli. They found an enhanced ITD and ILD discrimination performance when the eye gaze was directed towards the source. Since HMDs have a reduced field-of-view relative to the human visual system, sound source

localization abilities might also be affected due to reduced visual information. Furthermore, when having the room and the hand-location visible, the localization error has been shown to be lower than when subjects are blind-folded (Tabry et al., 2013).

Modern proprietary VR systems have been shown to reproduce immersive visual environments and to provide reliable spatial tracking accuracy, both for the reproduction of virtual visual scenes as well as for headset- and controller-tracking (Borrego et al., 2018; Niehorster et al., 2017). However, Niehorster et al. (2017) showed that when the tracking system of the HMD is lost, for example due to the user blocking the path between the tracking system and the HMD with their hands, the VR system fails to maintain the correct spatial location. Consequently, the calibrated location of the HMD within the room can be offset, i.e. a certain direction in VR does not correspond to the corresponding direction in the real-world. Such offsets can be difficult to track in proprietary VR systems. To overcome this issue, a real-world to virtual-world calibration is proposed.

The aim of the present study was to clarify how sound source localization ability is affected by a VR system, and in particular, how HMD-induced changes of the binaural cues and virtual visual information alter sound localization accuracy. In order to address this, a sound localization experiment was conducted in a loudspeaker environment with and without an HMD. For the localization task, a hand-pointing method was employed utilizing commercially available hand-held VR controllers. To evaluate the accuracy of the pointing method, a visual localization experiment was conducted, where the subjects' task was to point to specific locations. Since previous studies showed that sound localization accuracy can be influenced by visual information (Dufour et al., 2002; Tabry et al., 2013), the amount of visual information provided to the subjects was varied in the VR environment. One condition included no visual information (blind-folded); another provided visual cues regarding the room dimensions and the hand-location; in a third condition, the subjects saw the loudspeaker locations and were additionally provided with a laser pointer for pointing on the perceived sound location. It was hypothesized that effects as observed in previous audio-visual localization experiments in real environments, may also be reflected in VR.

## 4.2 Methods

### 4.2.1 Subjects

Ten subjects (three female, seven male) with an average age of 24 years participated in the study. None of the subjects had seen the experimental room before. The subjects were blind-folded when they were guided into the room. They had normal audiometric thresholds equal to or below 20 dB hearing loss at the octave-band frequencies between 250 Hz and 8 kHz and self-reported normal or corrected vision. Nine of the ten subjects were right handed and were instructed to use their main hand to hold the controller. The data of the left-handed subject were mirrored on the median plane. For each subject, the interpupillary distance was measured and the HMD was adjusted accordingly to ensure a clear binocular image and to minimize eye strain. All subjects provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

### 4.2.2 Acoustic reproduction method

The acoustic reproduction system consisted of 64 loudspeakers (KEF LS50, KEF, Maidstone, United Kingdom) housed in an anechoic chamber. The loudspeakers were arranged in a full sphere, at a distance of 2.4 m to the listening position, and driven by sonible d:24 (sonible GmbH, Graz, Austria) amplifiers. Twenty-seven of the loudspeakers in the frontal hemisphere were used in the present study. The loudspeakers were placed at three heights, at 0° (ear level) and ±28° elevation. Thirteen of the loudspeakers were at ear-level, distributed between -90° and +90° azimuth, with 15° separation. Seven loudspeakers were elevated by +28°, and seven by -28°, distributed between ±90° azimuth with 30° separation. All loudspeakers were equalized in level, delay and magnitude response as measured at the listening position. The loudspeakers were labelled using color coding (yellow, red and blue) to indicate elevation and numbers for azimuth locations (see Figure 4.1). The numbers ranged from one to thirteen, starting at -90° (left). The elevated loudspeakers at ±28° used only odd numbers, such that equal numerals were used for loudspeakers with the same azimuth angle.

### 4.2.3 Visual reproduction method

The real environment (RE), shown in Figure 4.1 (left panel), was replicated to ensure comparable visual information with and without the HMD (Figure 4.1, right panel). To present the VE, the HTC Vive system (HTC Corporation, New Taipei City, Taiwan) was used. This system consists of an HMD and two hand-held controllers to interact with the VE. Three additional Vive Trackers (HTC Corporation, New Taipei City, Taiwan) were used for VE-to-RE calibration (see details below). The spatial position and rotation of all devices were tracked with the infrared ray-tracking system. Blender (Blender Foundation, Amsterdam, The Netherlands) and Unity3D (Unity Technologies, San Francisco, CA) with the SteamVR plugin (Valve Corporation, Bellevue, WA) were used to replicate and present the VE, respectively.

Figure 4.1: Photography (left) and screenshot (right) of the acoustic reproduction system in the real and in the virtual environment (RE and VE). The loudspeakers are numbered in azimuth and color coded in elevation.

When the aim is to replicate a real scenario in VR, while maintaining the interaction with real objects, it is crucial to ensure spatial alignment between the real and the virtual world. To calibrate the virtual world to the real world, the three Vive Trackers were positioned on top of the ear-level loudspeakers at 0° and ±45° azimuth. Discrepancies in the positions of the trackers in the RE and the VE were accounted for as follows:

1. The normal of the plane spanned by the three points corresponding to the positions of the trackers was calculated for the RE and for the VE and the difference in rotation between them was applied to the VE. This ensured the correct orientation of the VE.

2. To correctly position the VE, the difference between one tracker and the respective reference point in the VE was calculated and the VE was shifted

accordingly. This resulted in an alignment of the RE and the VE at the chosen reference point.

3. The final rotation offset around the normal vector was corrected by calculating the angle difference of the vectors from the aligned reference point to an unaligned reference point in the VE and RE, i.e. to the known location of one of the other trackers and its virtual position.

After this procedure, the VE was aligned in both position and rotation relative to the RE. This method continuously accommodated for potential spatial discrepancies that might have occurred from tracking losses, as described by Niehorster et al. (2017). The system was recalibrated when either the tracker position error relative to the true position exceeded 2 cm or when the HMD lost tracking. The maximum allowed positional offset of the reference points resulted in a worst-case rotation error of 3.2°.

### 4.2.4 Acoustic stimuli

The acoustic stimulus consisted of a pink noise burst with a duration of 240 ms and 20 ms tapered cosine ramps at the onset and offset. The noise burst was created in MATLAB (The Mathworks, Natick, MA) using a sampling frequency of 48 kHz. For each stimulus presentation, a new noise burst was created. The stimulus was presented at a sound pressure level (SPL) of 65 dB, and roved by values between ±3 dB, drawn from a uniform distribution. The short duration limits effect of head movements during the stimulus presentation (Perrett and Noble, 1997) and the roving minimizes a spatial cue provided by directional loudness differences (Makous and Middlebrooks, 1990; Musicant and Butler, 1984). The subjects were asked to re-center their viewing direction before each stimulus representation, i.e., to face the 0° on-axis loudspeaker at ear level. The HMD rotation was logged in a subset of the conditions and for a subset of the subjects to evaluate if, on average, the viewing direction was centered at the time of the acoustic stimulus exposure. An initial azimuth rotation of the HMD of -1° ± 3° standard deviation was found.

### 4.2.5 Experimental conditions

Table 4.1 shows an overview of the eight experimental conditions considered in this study. The column ‘visual information’ shows the visual environment

that was presented to the subjects. The stimulus refers to the localization task, which was either visual localization (visual search) or sound localization. The last column indicates whether the HMD was worn or not. Each condition and each of the 27 source locations (see section 4.2.2) was presented five times to each of the subjects. Thus, each condition consisted of 135 stimuli which were presented in fully random order.

The four blocks shown in Table 4.1 were presented in a fixed-order from Block I to Block IV to control for the exposure to the loudspeaker locations. The conditions were randomized within the blocks. No pre-experimental training was conducted to avoid any bias of the subjects with respect to a specific condition or regarding visual information.

Table 4.1: Overview over the conditions considered and grouped into blocks. Conditions were randomized within blocks. The conditions varied in available visual information, target stimulus and if the head-mounted display was worn.

| Block | Visual information | Stimulus | HMD |
|---|---|---|---|
| I | Blind-folded | Acoustic | No |
| | Blind-folded | Acoustic | Yes |
| II | Virtual env., no loudspeaker (LS) | Acoustic | Yes |
| III | Virtual env., LS | Visual | Yes |
| | Real env. | Visual | No |
| | Real env. | Acoustic | No |
| | Virtual env., LS | Acoustic | Yes |
| IV | Virtual env., LS, laser pointer | Acoustic | Yes |

The two blind-folded conditions in Block I were used to examine whether, or to what extent, simply wearing the HMD has an influence on sound source localization. The reference localization accuracy was measured using the acoustic stimuli described above, while subjects were blind-folded with a sleeping mask. To assess the effect of the HMD frame on sound source localization, subjects wore the HMD on top of the sleeping mask.

The visual localization conditions in Block III were employed to investigate the baseline accuracy of the pointing method using the hand-held VR controller in the RE and the VE (Figure 4.1). The subjects were instructed to point at a loudspeaker location shown either on an iPad Air 2 (Apple Inc., Cupertino, CA) in the real environment or on a simulated screen within the VE.

To investigate the influence of the HMD on sound localization when visual information regarding the loudspeaker positions is available, the sound localization accuracy was evaluated in the real and in the virtual loudspeaker environments (Block III). The subjects were informed that sounds could also

come from positions in-between loudspeakers. While the visual information regarding the loudspeaker positions was available in both conditions, the VE provided reduced visual information. The field-of-view of the HTC Vive is about 110° while the visual field of the human visual system is larger. Furthermore, in the VE only the hand-held controller was simulated but not the arm, which was inherently visible in the RE.

In addition to evaluating the pointing accuracy and the HMD-induced localization error, the effect of varying the amount of visual information on sound localization in the VE was investigated. In Block II, the experimental room (anechoic chamber) without the loudspeakers was rendered on the HMD and the subjects were asked to localize the acoustic stimuli as described for the previous conditions. The experiment thus included conditions with various degrees of visual information available to the subjects in the VR: no visual information, a depiction of the empty room including hand-location, and a depiction of the complete room including the locations of the sound sources.

To assess the role of visual feedback of the pointer location, in Block IV, a simulated laser pointer emerging from the hand-held controller was shown while the subjects completed the localization task in the VE with the room and the loudspeaker setup visible.

### 4.2.6 Pointing method

The controller of the VR system was used as a pointing device. The subjects were instructed to indicate the perceived stimulus location by stretching their arm with the wrist straight while holding the controller, in an attempt to minimize intra- and inter-subject variability in pointing. The pointing direction was defined by the intersection point of an invisible virtual ray originating at the tip of the controller extending the base of the device and an invisible virtual sphere with the origin at the listeners head position and the same radius as the loudspeakers (r=2.4 m). The perceived position of the source was indicated with a button press using the index finger.

On each trigger button press, the PC running the VR system transmitted an Open Sound Control (OSC) message via User Datagram Protocol (UDP) over an Ethernet network to the audio processing PC. The audio processing PC subsequently presented the next acoustic or visual stimulus, with a delay of three seconds to allow the subject to re-center the viewing direction. With a responding OSC message, the audio processing PC permitted the reporting of

the perceived location after the playback completed.

A virtual model of the controller was rendered in all conditions containing visual information in the HMD. Thus, the visual feedback of the controller position in Blocks II and III was similar, independent of whether the HMD was worn or not. To standardize the pointing method for all audio-visual conditions, a direction marker, functioning as a visual pointing aid, was not provided in this study, except in the last condition (Block IV, Table 4.1), since a sufficiently comparable method was infeasible in the real environment. Thus, the pointing method in Blocks II and III was similar to free-hand pointing.

### 4.2.7 Physical analysis

The effect of the HMD on the acoustic ear signals was analyzed from B&K Head and Torso Simulator (HATS; Type 4128-C; Brüel & Kjær A/S, Nærum, Denmark) measurements with and without the HTC Vive HMD. Binaural impulse responses were recorded from all 64 loudspeakers with a 22 s long exponential sweep (Müller and Massarani, 2001) in a frequency range from 60 Hz to 24 kHz and truncated to 128 samples (2.7 ms) to remove reflections from other loudspeakers and objects in the room. The dataset of the measurements can be found in a repository (zenodo.org/record/1185335). Acoustic perturbations of the HMD on the frequency spectrum were analyzed for the same set of loudspeakers as employed in the perceptual experiment by calculating the power in auditory filters between 200 Hz and 16 kHz with equivalent rectangular bandwidths (Glasberg and Moore, 1990) using the Auditory Modeling Toolbox (Soendergaard and Majdak, 2013). The power in the auditory filters was averaged in three frequency regions from 200 Hz to 1 kHz, 1 to 5 kHz and 5 to 16 kHz.

Spectral differences (SD) were calculated as the mean absolute power differences of the three frequency regions, measured with and without the HMD. Interaural level differences (ILD) were determined in the same frequency region as the SD using the power differences at the output of the auditory filters between the left- and the right-ear signals. Interaural time differences (ITD) were calculated as the delay between the left- and right-ear signals. The delay of each impulse response was defined as the lag of the peak of the cross-correlation between the impulse response and its minimum-phase version (Nam et al., 2008).

### 4.2.8 Pointing bias

It was hypothesized that subjects might have a bias in pointing direction due to the shape of the hand-held controller, and because they had no knowledge on where the 'invisible ray' was emerging from the controller. A bias for each subject was therefore calculated in azimuth and elevation as the mean of all source locations in the two visual localization conditions (real and virtual). Individual responses were then corrected by the calculated azimuth and elevation biases for all conditions except the condition with visual feedback of the pointer location (laser pointer condition).

### 4.2.9 Analysis of behavioral responses

Statistical analyses on the subject response errors were performed by fitting mixed-effects linear models to the azimuth and elevation errors. The subject responses were corrected by a bias estimation due to the pointing method, as described above. Responses that were localized farther than 45° from the target location in either azimuth or elevation were treated as outliers. Of the 10800 subject responses, 0.29 % were treated as outliers and discarded from the analysis.

Only the sources in the horizontal plane were considered in the statistical analyses of the azimuth localization errors. The azimuth stimulus location and the experimental condition (see Table 4.1) as well as their interaction were treated as fixed effects. Regarding the elevation error, the stimulus location in both azimuth (only azimuth directions that occurred in all elevation directions) and elevation, the experimental condition, as well as their interactions were treated as fixed effects. The influence of the subjects, the repetitions and their interaction were considered as random effects. The p-values were calculated based on likelihood-ratio tests for the random effects and on F-tests for the fixed effects (Kuznetsova et al., 2014). Post-hoc analyses of within factor comparisons were performed using estimated marginal means calculated from the mixed-effects linear models and using Bonferroni p-value adjustments (Lenth, 2016).

## 4.3 Results

### 4.3.1 Pointing Bias

Figure 4.2 shows the pointing bias in azimuth (squares) and elevation (circles) for each subject calculated from the visual localization experiments. Regarding the azimuth bias, the subjects tended to point slightly too far to the left (-3.5° to -0.1°), except for subject S07, who had a slight positive (right) azimuthal bias of 1.3°. Overall, the average bias across subjects in azimuth was 1.6° (left). The only left-handed subject (S08) showed a similar bias as the other subjects. The bias in elevation angle (circle) was found to be higher than the azimuth bias for all subjects, with an average value of 19.0°. The subjects generally tended to point too high. The variance across subjects was between 12.8° and 28.6° and is likely to be related to the shape of the hand-held controller and the internal reference of each subject of where the "invisible ray" emerges from the controller. The responses of the subjects were corrected by the pointing bias for all conditions except the laser pointer condition.

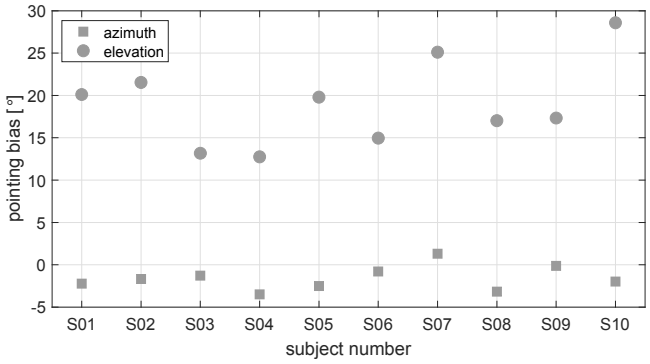



Figure 4.2: Pointing bias in azimuth (squares) and elevation (circles) for each subject. The bias was calculated as the mean error over all source locations in the two visual localization conditions. Negative angles indicate biases to the left and downwards for azimuth and elevation, respectively.

### 4.3.2 Spectral differences and interaural errors

Figure 4.3 shows the spectral difference (SD) of the left-ear signals obtained with and without the HMD on the B&K HATS for the azimuth source locations between -90° and +90° and for the elevation angles of -28° (downward triangles), 0° (diamonds) and 28° (upward triangles). For ipsilateral sources (negative azimuth angles), the SD was low for all frequency regions and all elevation

angles. In the low-frequency region between 200 Hz and 1 kHz (dashed lines), the SD was also below 1 dB for contralateral sources (positive azimuth angles), independent of the elevation angle. The SD in the mid-frequency (dashed-dotted lines) and high-frequency (solid lines) regions was found to be up to 6.3 dB for elevation angles at and above the horizontal plane (0°, 28°). For these elevations, the error was higher in the high-frequency region than in the mid frequency region, except at 60°. For mid- and high-frequency sources below the horizontal plane (i.e. an elevation angle of -28°), the SD was lower than for the other two elevation angles.

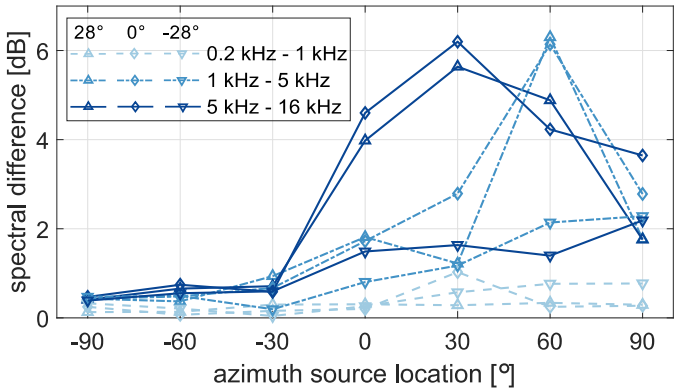



Figure 4.3: Spectral difference (SD) measured at the left ear of the B&K HATS with and without the HTC Vive. The angles in the legend represent the elevation angles considered in the current study. The SD was calculated in auditory bands and averaged over three frequency regions at low-, mid- and high frequencies as shown in the legend.

Figure 4.4 shows the signed errors in ILD (left panel) and ITD (right panel) induced by the HMD measured in the horizontal plane as a function of source azimuth angle. The ILD error in the low-frequency (dashed line) region was below 2 dB. In the mid- and high- frequency regions (dashed-dotted line and solid line), the error was lowest at the frontal source location (0° azimuth) and at 90. The largest error was about 6 dB at source angles of 60° and 30°, for the mid and high frequency regions, respectively. The ITD error was below 1 sample (21 µs) for source angles between 0° and 30° and increased to 62.5 µs for the source angle of 75°. The ITD error was 0 µs at 90° azimuth angle.

### 4.3.3 Pointing accuracy with VR controllers

Figure 4.5 shows the responses of the subjects in the two visual localization experiments in the RE and the VE. In both conditions, the subjects' task was to

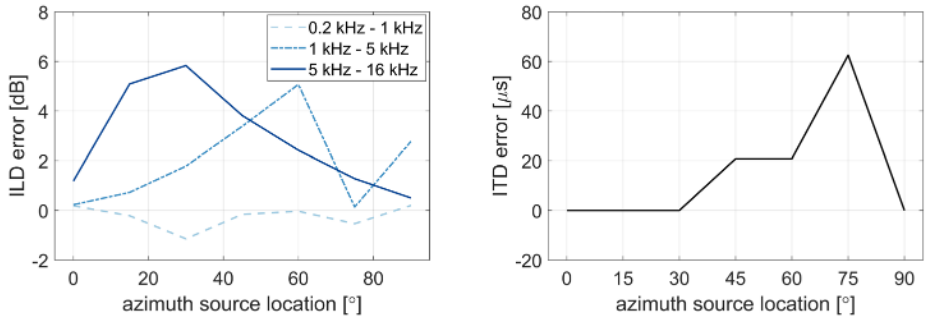



Figure 4.4: Signed errors of interaural differences in level and time (ILD and ITD) with respect to azimuth angles on the horizontal plane (0° elevation). Positive errors indicate larger ILDs and ITDs with than without the HMD. The ILDs were calculated in auditory bands and averaged over three frequency regions at low-, mid- and high frequencies as shown in the legend. The ITDs were calculated from the delays between the broadband binaural impulse responses (see Methods for details).

point to the center of the loudspeaker indicated on a screen. The filled black squares represent the 27 source locations. The small colored symbols represent the individual responses of the subjects and the large open black symbols indicate the mean responses, averaged across subjects and repetitions. The connecting lines between the target location and the mean responses indicate the localization error. The subjects generally pointed close to the correct loudspeaker, whereby the precision of the responses was generally higher for azimuth than for elevation localization.

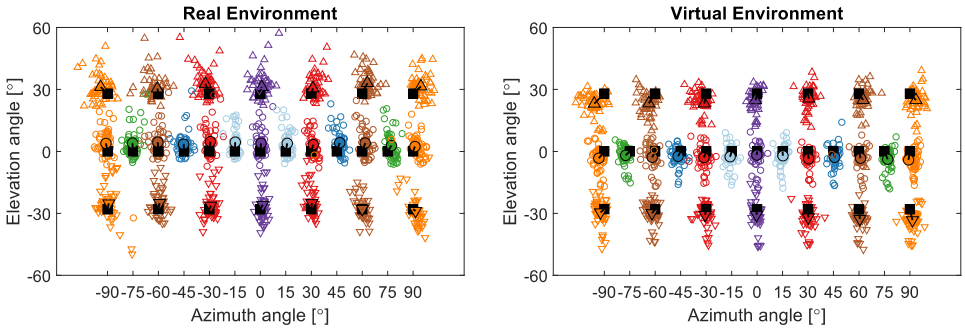



Figure 4.5: Response plot of the visual localization experiment for the real and virtual environment. The black squares represent the source locations. The small markers show the responses of the subjects and the large markers the mean response over subjects and repetitions. Negative angles represent sources to the left and downwards for azimuth and elevation, respectively.

Figure 4.6 shows the azimuth error obtained in the visual pointing experiment. The circles represent the mean absolute error and the boxplots indicate the signed error. The signed azimuth error was significantly affected

by the conditions [$F(1,1263)=60.63$, $p<0.0001$], the azimuth source location [$F(12,1263)=28.02$, $p<0.0001$] and their interaction [$F(12,1263)=1.9$, $p=0.0305$]. The difference between the VE and the RE was only significant for sources on the left at -90° [$t(1263)=4.52$, $p=0.0001$], -75° [$t(1263)=4.26$, $p=0.0003$] and -30° [$t(1263)=3.71$, $p=0.0028$]. At -45° and -60° the difference was smaller and did not lead to significant effects after correction for multiple comparisons [-45°: $t(1263)=2.64$, $p=0.1078$; -60°: $t(1263)= 2.73$, $p=0.0833$]. Generally, the responses showed a small overshoot, i.e. the sources on the right side showed a shift to the right, while the responses for sources on the left tended to show a shift to the left. The overshoot was larger for sources on the left in the virtual environment with the HMD than in the real environment.

The condition (RE vs. VE) was not found to have an effect on the absolute azimuth error [$F(1,1263)=0.76$, $p=0.38$]. Azimuth location [$F(12,1263)=7.27$, $p<0.0001$], as well as the interaction between the condition and azimuth location [$F(12,1263)=1.83$, $p=0.0392$], were found to be significant. However, after correction for multiple comparisons, no statistically significant difference between the RE and the VE was found for any azimuth location [$p>0.11$]. The hemispheric difference, i.e. the difference between the left and the right side, of the absolute azimuth error difference between the RE and the VE was not found to be significant. However, the hemispheric difference was larger at the lateral source locations [±90°: $t(1263)=-2.8$, $p=0.0587$; ±75°: $t(1263)=-2.25$, $p=0.22$] than at the source locations closer to the median plane [$p>0.43$].

The analysis of the absolute elevation error showed that the conditions [$F(1,2042)=9.77$, $p=0.0018$] and the three-way interaction of condition, azimuth location and elevation location [$F(12,1263)=2.02$, $p=0.0196$] were significant. The two-way interaction of the conditions with the source locations in azimuth [$F(6,2042)=2.21$, $p=0.0398$] was significant, but not with the source locations in elevation [$F(2,2042)=2.29$, $p=0.1$]. The interaction between the source locations in azimuth and elevation was not significant [$F(12,2042)=1.42$, $p=0.15$] and also the source locations [Azimuth: $F(6,2042)=0.99$, $p=0.43$; Elevation: $F(2,2042)=1.02$, $p=0.36$] did not reveal significant effects.

The analysis of the signed elevation error showed significant contributions of the conditions [$F(1,2069)=609.11$, $p<0.0001$], the source azimuth [$F(6,2069)=2.42$, $p=0.0245$], the source elevation [$F(2,2069)=7.85$, $p=0.0004$] as well as the interactions between the source elevation and the conditions [$F(2,2069)=9.51$, $p<0.0001$] and both source locations [$F(12,2069)=2.7$, $p=0.0013$]. The interac-

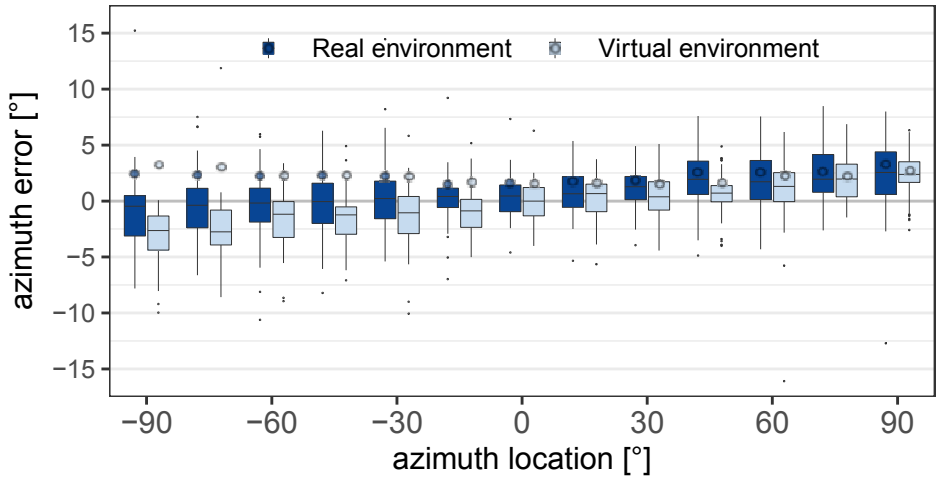



Figure 4.6: Mean absolute (circles) and signed (boxplots) azimuth error for visual localization in the virtual (light blue) and the real (dark blue) environment. The error is shown over the thirteen azimuth angles in the horizontal plane (0° elevation). The boxplots indicate the median (line) and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

tion between the source azimuth and the conditions [$F(6,2069)=0.75$, $p=0.61$] and the three-way interaction [$F(12,2069)=0.51$, $p=0.91$] were not found to be significant. The difference between the RE and the VE was significant at all three source elevations, at 0° elevation [$t(2069)=15.48$, $p<0.0001$], as well as below [$t(2069)=10.76$, $p<0.0001$] and above [$t(2069)=16.52$, $p<0.0001$] the horizontal plane. The signed elevation error was positive (upwards) in the RE and negative (downwards) in the VE, as indicated by the lines between the target markers (black squares) and the response markers (colored squares) in Figure 4.5. On average, the subjects pointed 6.5° higher in the RE than in the VE.

### 4.3.4 Influence of HMD on azimuth error

Figure 4.7 shows the absolute (circles) and the signed (boxplots) azimuth error as a function of the azimuth source locations. Negative angles represent sources on the left and positive angles indicate sources on the right of the subjects. The dark grey boxes and circles represent results for the condition where the subjects were blind-folded and the light grey boxes and circles show the results where the subjects were blind-folded and wore the HMD. The mean absolute azimuth error was always found to be larger with the HMD except for 0°. This difference was larger on the left than on the right side. The analysis of the

signed error, employing a linear mixed effects model, showed that the effect of the conditions was not significant [F(1,1265)=0.6, p=0.44], while the source locations [F(12,1265)=62.04, p<0.0001] and the interaction [F(12,1265)=3.04, p=0.0003] were significant factors.

The median of the signed error showed that the sources were perceived further lateral with the HMD than without. The post-hoc analysis for the sources on the left side showed that the difference between the conditions reached significance only at a source angle of -60° [t(1265)=3.3, p=0.0059]. At -45° the p-value exceeded the 5% significance level after correction for multiple comparisons [t(1265)=2.4, p=0.0944]. On the right side, the difference was significant only at 45° [t(1265)=-3.1, p=0.0119]. The error induced by the HMD on the signed azimuth error was larger on the left than on the right side at 45° [t(1265)=3.9, p=0.0014] and 60° [t(1265)=2.96, p=0.0374].

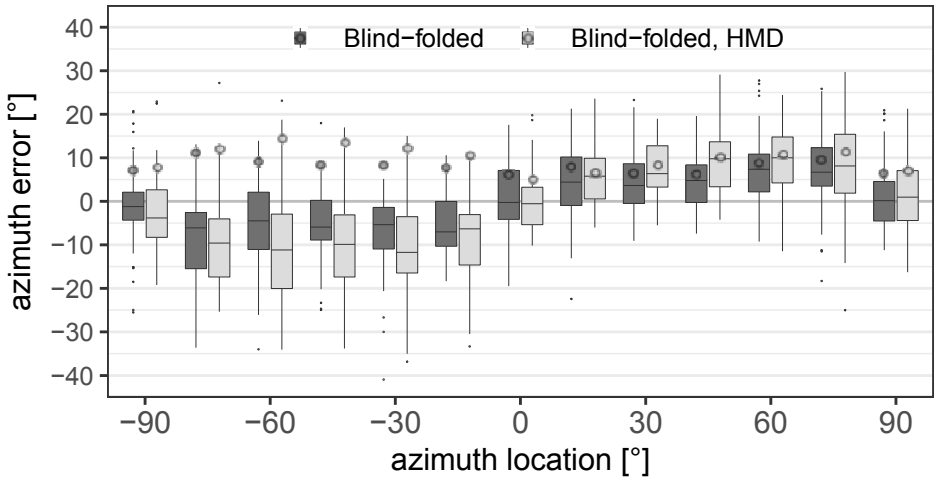



Figure 4.7: Mean absolute (circles) and signed (boxplots) azimuth error for acoustic localization of blind-folded subjects with (light grey) and without (dark grey) the head mounted display (HMD). The error is shown over the thirteen azimuth angles in the horizontal plane (0° elevation). The boxplots indicate the median (line) and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

### 4.3.5 Influence of visual information on azimuth error

Figure 4.8 shows the absolute (circles) and signed (boxplots) azimuth error as a function of the azimuth source locations for five conditions with varying visual information. The subjects were either blind-folded with the HMD (light grey), were provided with information regarding the room dimensions and the hand

position information (white), could see the real room (dark blue), were provided with the virtual version of the real room (light blue), or were provided with a laser pointer in the virtual room (blue).

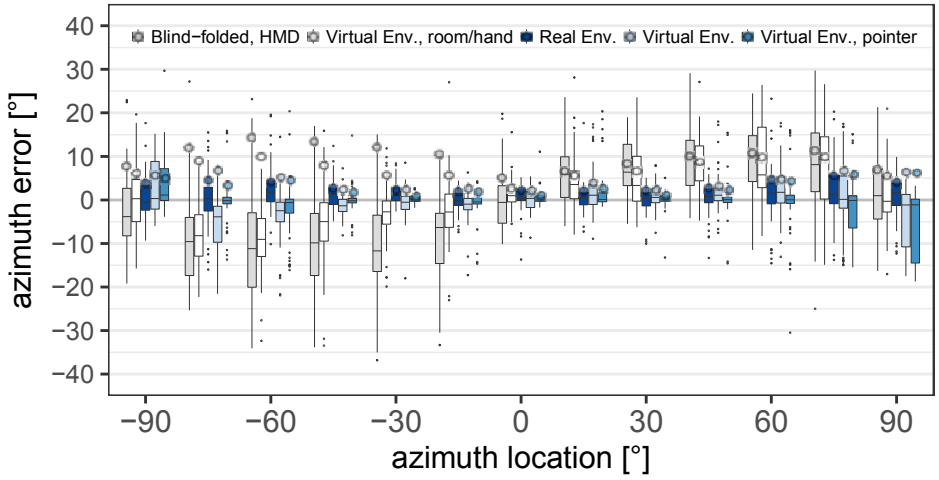



Figure 4.8: Mean absolute (circles) and signed (boxplots) azimuth error for acoustic localization with varying visual information in the virtual environment and the real environment. In all conditions, except in the real environment, subjects wore the head-mounted display (HMD). The conditions depicted with shades of blue color include visual information of possible source locations. The error is shown over the thirteen azimuth angles in the horizontal plane (0° elevation). The boxplots indicate the median (line) and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

The analysis of the linear mixed-effects model of the absolute azimuth error showed that both main effects of azimuth source location [$F(12,3176)=23.17$, $p<0.0001$] and conditions [$F(4,3176)=223.88$, $p<0.0001$] as well as their interaction [$F(48,3176)=4.57$, $p<0.0001$] had a significant effect on the azimuth error. A post-hoc analysis was performed to investigate the effect of the room and hand position information, the loudspeaker locations, and the laser pointer for aided pointing on the perceived sound source.

A significant decrease in error was found when comparing the blind-folded HMD condition (light grey) with the condition where the subjects had visual information of room and hand position (white) at azimuth location between -15° and -75° [$p<0.039$]. No significant change in error was found at -90° [$t(3176)=1.61$, $p=1$], at the right side of the subjects [$p=1$] nor for the frontal source [$t(3176)=2.34$, $p=0.25$].

When also visual information of the loudspeaker locations was provided (light blue), the subjects generally pointed towards the correct loudspeaker.

When the loudspeaker locations were provided, the azimuth localization error decreased in comparison to the condition where only room and hand-location information were given. The reduction in error was significant for azimuth locations in the left hemifield between -15° and -60° [$p<0.03$] and in the right hemifield between 30° and 75° [$p<0.02$]. The lateral sources on the left [$p>0.34$] and on the right [$p=1$], as well as sources close to the midline at 0° and 15° [$p=1$] were not found significantly different in the two conditions with and without visual representations of the loudspeakers.

When the laser pointer was shown in the VE (blue), the absolute azimuth error was not found to be different from the condition without the laser pointer (light blue) [$p=1$], except for the azimuth angle -75° [$t(3176)=3.42$, $p=0.0084$]. Comparing the VE without the laser pointer (light blue) and the RE (dark blue) showed no significant difference of the azimuth error at any of the source locations on the horizontal plane [$p>0.28$].

### 4.3.6 Influence of HMD on elevation error

Figure 4.9 shows the error in elevation as a function of the elevation target location. The results for the blind-folded conditions are shown with the HMD (light grey) and without the HMD (dark grey). The analysis of the linear mixed-effects model including the two conditions, azimuth and elevation locations, revealed significant main effects of conditions [$F(1,2049)=35.29$, $p<0.0001$] and source elevations [$F(2,2049)=8.28$, $p=0.0003$] but no effect of the azimuth locations [$F(6,2049)=1.98$, $p=0.0653$]. The interactions of the source locations with the conditions were not found to be significant [azimuth: $F(6,2043)=0.98$, $p=0.44$; elevation: $F(2,2041)=0.7$, $p=0.5$]. The elevation error was found to increase with the HMD in comparison to the condition without the HMD by 1.8°.

### 4.3.7 Influence of visual information on elevation error

Figure 4.10 shows the absolute elevation error at the three source elevations for the five conditions with varying visual information. The statistical analysis showed significant main effects of the conditions [$F(4,5123)=289.44$, $p<0.0001$] and the azimuth locations [$F(6,5123)=9.47$, $p<0.0001$], but no effect of the elevation locations [$F(2,5123)=2.04$, $p=0.13$]. The interaction between the conditions and the locations was found to be significant for the elevation [$F(8,5123)=5.73$,

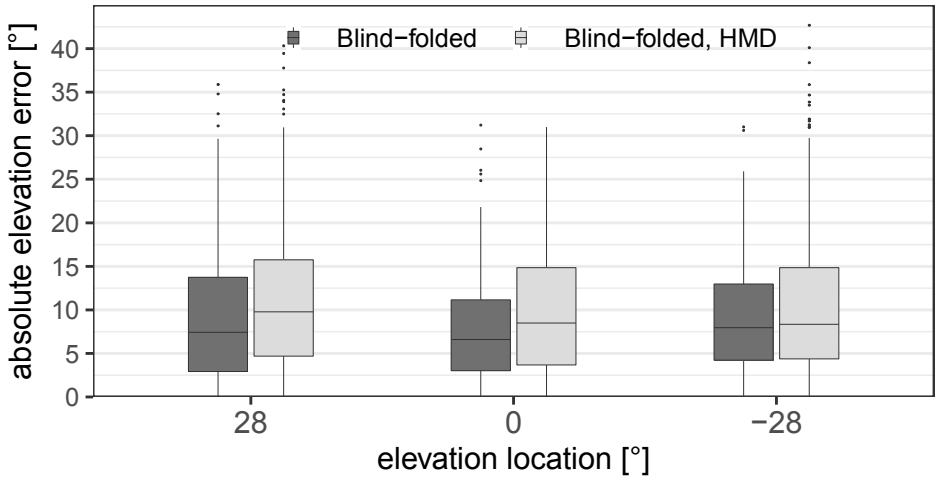



Figure 4.9: Absolute elevation error in degrees for acoustic localization for blind-folded subjects with (light grey) and without (dark grey) the head mounted display (HMD). The error is shown over the three elevation angles and includes the sources from all azimuth locations. The boxplots indicate the median (line) and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

p<0.0001] and non-significant for the azimuth [F(24,5123)=1.43, p=0.0779].

The post-hoc analysis showed a significant drop in error between the blind-folded condition (light grey) and the condition with visual room and hand-position information (white) for all source elevations [p<0.0001]. The mean decrease of the error was 2.1° (from 10.2° to 8.1°). When the loudspeakers were visualized (light blue) in addition to the room and hand position (white), the elevation error was found to further decrease significantly for the elevated sources [p<0.0018], but not for the sources in the horizontal plane [t(5123)=-0.18, p=1]. When the laser pointer was employed (blue), the absolute elevation error was lowest (2.4°) and significantly different from the error in the VE condition without the laser pointer (light blue) [p<0.0001]. The comparison of the elevation error in the RE (dark blue) and the VE (light blue) revealed a significantly larger error for the sources at an elevation angle of 0° [t(5123)=-5.64, p<0.0001] but not for the sources above [t(5123)=-2.04, p=0.16] and below [t(5123)=-0.42, p=1] the horizontal plane.

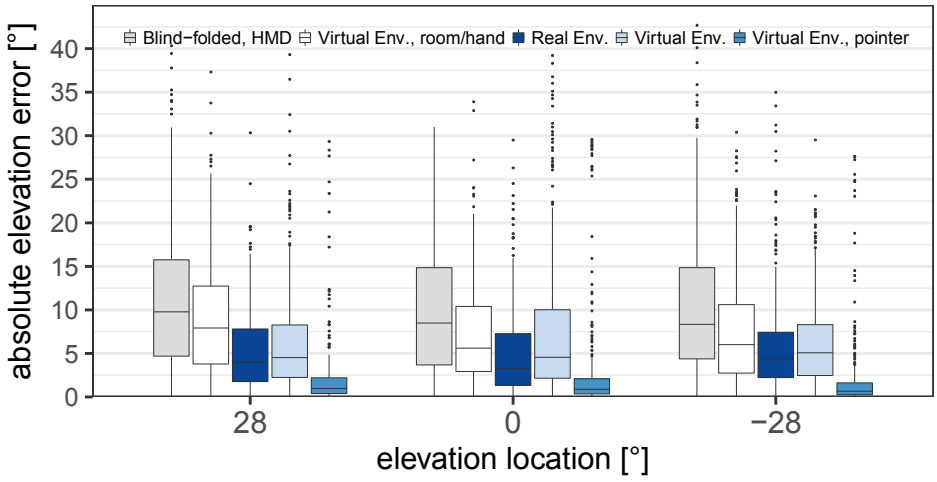



Figure 4.10: Absolute elevation error for acoustic localization with varying visual information in the virtual environment and the real environment. In all conditions, except in the real environment, subjects wore the head-mounted display (HMD). The conditions depicted with shades of blue color include visual information of possible source locations. The error is shown over the three elevation angles and includes the sources from all azimuth locations. The boxplots indicate the median (line) and the first and third quartile. The whiskers extend to 1.5 times the interquartile range.

## 4.4 Discussion

### 4.4.1 Degraded sound localization with HMD

The interaural disparities in the present study were found to be larger when measured with the HMD than without the HMD, consistent with results from the study of Gupta et al. (2018). Since Hafter and De Maio (1975) found ITD just-noticeable differences (JNDs) between 20 and 40 µs for lateral sources, the differences in ITD of up to 60 µs µs induced by the HMD (see Figure 4.4) might be perceptible. Likewise, the ILD changes of up to 6 dB induced by the HMD (see Figure 4.4) were found to be above the JND of 2 dB (Hafter et al., 1977; Mills, 1960).

The effect of the HMD on the perceived location of acoustic stimuli in an anechoic environment was investigated in the blind-folded conditions. The error without the HMD was comparable in azimuth and elevation to values reported in earlier studies (e.g. Oldfield and Parker, 1984). No difference in azimuth localization error with and without the HMD was found for sources at or around the median plane, which is consistent with the small errors of ITDs and ILDs induced by the HMD. However, for lateral sources, the azimuth error

was larger with the HMD than without the HMD, which is a consequence of the larger binaural disparities caused by the HMD. However, the increase in localization error was larger on the left side than on the right side. A comparable difference was also observed in the visual pointing experiment with a larger overestimation of the source location on the left than on the right side when wearing the HMD (see Figure 4.6). However, no difference between the left and the right hemisphere was found in terms of the absolute azimuth pointing error. Thus, there seem to be additional factors contributing to the localization error beyond the acoustical error induced by the HMD. Possibly, the HMD represents an obstacle when pointing with the right hand to the left hemisphere, resulting in a larger pointing error.

To localize sources in elevation, the auditory system mainly relies on direction dependent spectral cues provided by acoustical filtering of the pinnae, head and body (Hebrank and Wright, 1974). Recent studies investigated the effect of HMDs on HRTFs in a similar way as in the current study and showed spectral perturbations up to about 6 dB at high frequencies (Genovese et al., 2018; Gupta et al., 2018). In the current study, comparable spectral differences were found as well as an increase in elevation error by about 2°. However, the spectral differences as shown in Figure 4.3 are not in line with the localization error differences in Figure 4.9. While the elevation localization error was the same at all tested elevations, the spectral differences varied with source elevation. Specifically, smaller spectral differences were found for sources below the horizontal plane than at or above the horizontal plane. Thus, the spectral difference is not a good predictor of the localization accuracy. In fact, it has been shown that spectral differences do not correlate well with elevation localization errors (Middlebrooks, 1999) but that elevation perception is based on multi-feature, template-based matching (Baumgartner et al., 2014; Macpherson and Sabin, 2013; Van Opstal et al., 2017).

### 4.4.2 Visual information influences sound localization in VR

In the virtual condition representing the simulated anechoic room without loudspeakers, a reference frame was provided by the room and the subjects could see the pointer. The sound localization error was found to be smaller with this visual information than in the blind-folded condition with the HMD. The contributions of the visual information about the hand position and about the room dimensions cannot be separated, since the current study was not

designed to distinguish between these two visual aspects. Tabry et al. (2013) also observed lower errors of both azimuth and elevation sound localization in conditions similar to those in the current study whereby real visual information of the subjects' body and of the room were presented instead of virtual visual information. Thus, the amount of visual information provided in the current study may be considered to resemble those provided in a real environment. However, Tabry et al. (2013) found substantially larger elevation errors than in the present study both in the condition with and without visual information. The smaller elevation errors found in the present study might be due to the limited set of elevation source locations (-28°, 0°, 28°) as compared to the study of Tabry et al. (2013), where the range of elevations was between -37.5° and 67.5°. Subjects might be able to learn possible source locations which can improve the localization accuracy (Ege et al., 2018).

When the source locations were visible, the azimuth and elevation errors decreased by 3° and 1.5°, respectively, consistent with results obtained in real environments (Shelton and Searle, 1980). No improvement in localization accuracy was found in azimuth for frontal sources and in elevation for sources in the horizontal plane, because the auditory localization accuracy, even without visible source locations, was already high compared to those that are away from the midline and horizontal plane. However, there was likely a high visual bias towards the loudspeakers in this condition. Thus, mainly pointing accuracy and not sound localization accuracy was measured. The location priors have an even higher impact on the elevation than on the azimuth accuracy since only three elevation locations were used.

The additional information provided by the visual feedback from the laser pointer had only a negligible effect on the localization accuracy in azimuth, but a clear effect on the elevation accuracy. The elevation error was larger when no laser pointer was visible which might be partly due to the shape of the VR controller. The shape of the controller led to a biased pointing direction. The bias correction as described above (see Figure 4.2) was intended to reduce the influence of the controller. However, even though the subjects were asked to always hold the controller in the same way, the controller positioning might have varied leading to a larger pointing error when no visual feedback of the laser pointer was provided. Even though the effect of the laser pointer on the mean azimuth error was negligible, the variance of the subjects' responses decreased when the visual feedback of the pointing direction was available.

Thus, the responses with the laser pointer become more reliable.

Comparing the localization error in the real and the virtual environments showed no differences in terms of the azimuth error. The elevation error was significantly increased at 0° elevation and was found to be slightly, but not significantly, larger above the horizontal plane even though not significantly. Below the horizontal plane no difference between RE and the VE was found. Even though the provided visual information was supposed to be the same in the two environments, some differences were unavoidable. In the real loudspeaker environment the subjects could see their arms, but not in the VE, where only the controller was visible. However, Van Beers et al. (1999) showed that the visual feedback of the arm does not seem to increase visual object localization accuracy compared to the situation when only the finger is visible. Nevertheless, for pointing to sources on and above the horizontal plane the arm might have been helpful visual information for more accurate pointing, however, no such evidence was found in the visual pointing experiment.

### 4.4.3 Potential training effects

The subjects did not receive training before starting the experiment, but were introduced to the task and the controller. Previous studies indicated that training can improve sound localization accuracy (Carlile et al., 1997; Majdak et al., 2010; Makous and Middlebrooks, 1990). Since the subjects could not be introduced to the visual loudspeaker environment beforehand, no training was provided to avoid potential benefits in certain conditions but not in others. It is possible that the localization performance of the subjects improved throughout the course of the experiment. To minimize this effect, the conditions were presented in a random order within the experimental blocks. The random-order presentation avoided a bias within the blocks but could not exclude an inter-block bias. Such inter-block bias was unavoidable because visual information needed to be revealed to the subjects reflecting an increase of information content.

### 4.4.4 Implications for VR in hearing research

VR glasses in combination with audio reproduction techniques may allow for novel ways of conducting research in simulated realistic environments. While previous research typically involved simple audio-visual information, with VR, research can be conducted in ecologically more valid environments while main-

taining controllability and reproducibility of the experiment. Even though the localization error increased with the HMD in the blind-folded condition, these errors may not be noticeable in realistic environments including visual information. This might also be the case for hearing-impaired subjects for whom sound localization accuracy is commonly reduced compared to normal-hearing subjects (Dobreva et al., 2011).

Even though only a single device, the HTC Vive, was investigated in the current study the findings may generalize with regards to other commercial virtual reality glasses. It has been shown that other HMDs, including the Oculus Rift (Oculus VR LLC, Menlo Park, CA) or the Microsoft HoloLens (Microsoft, Redmond, WA), lead to comparable or even smaller acoustic perturbations (Genovese et al., 2018; Gupta et al., 2018). Thus, the sound localization error due to the HMD is likely to be comparable or lower than that of the HTC Vive. Visual reproduction and tracking specifications seem comparable between current commercial products.

## 4.5 Conclusions

VR systems and loudspeaker-based audio reproduction allow for full immersion into an audio-visual scene. In the present study, sound source localization accuracy with an HMD providing a varying amount of visual information was investigated. A calibration system to align the real world and the virtual world was developed. Hand-pointing accuracy to visual targets was evaluated using commercially available controllers. The accuracy of the hand pointing to visual targets was found to be high in the azimuth direction, whereas a large bias was found in terms of elevation accuracy due to the shape of the controller. The sound localization experiment demonstrated a small detrimental effect of the HMD on the sound localization accuracy. While the azimuth error induced by wearing the HMD was negligible in the frontal area, it was significant at more lateral sound source locations which is in line with changes in binaural disparities. However, the error induced by the HMD was found to be larger on the left than on the right side in the acoustic localization experiment. Similarly, in the visual pointing experiment a larger overshoot was found with the HMD than without in the left but not in the right hemisphere. Thus, the error might not be purely of acoustic nature but also due to the HMD influencing the motion behavior. The elevation error was about 2° larger with the HMD for all azimuth

and elevation directions.

Generally, the sound localization accuracy was found to be higher when visual information was available than in the conditions without visual information. The lowest accuracy was found when the subjects were blind-folded and a significant improvement was found for both azimuth and elevation when room and hand position information were provided. An additional laser pointer for pointing guidance did not lead to an improvement of azimuth localization but an improved elevation localization.

## 4.6 Supplementary data

All data are available from the zenodo database. The perceptual data can be found here: 10.5281/zenodo.1293059 and the impulse response data can be found here: 10.5281/zenodo.1185335.

# 5 General discussion

In this thesis, human perception in auditory and audio-visual virtual environments was investigated. Three main research questions were considered, as outlined in the introduction:

- How well can an acoustic virtual room, created with state-of-the-art techniques, match a real room in terms of speech intelligibility?
- What is the relationship between the source size and speech intelligibility in spatial conditions?
- What is the role of visual information, and the impact of virtual reality glasses, on sound localization?

## 5.1 Summary of main findings

Regarding the first research question, it was shown that an acoustic reproduction based on impulse responses measured with a microphone array provided the closest match to a reverberant reference room in terms of speech reception thresholds, while the reproduction based on room acoustic simulations showed significantly lower SRTs compared to the reference room. The differences in speech intelligibility could be accounted for by using a binaural speech intelligibility model that considers better-ear signal-to-noise ratio differences and binaural unmasking effects. Both components were found to contribute to the speech intelligibility differences.

To study the relation between the spatial width of virtual sound sources and speech intelligibility, the second research question, sources with different physical source size were created using ambisonics. Ambisonics processing leads to varying degrees of energy spread, depending on its order. It was found that the energy spread of the virtual sources did not show an effect on the perceptual size; however, speech intelligibility was found to be worse (higher SRTs) with wider sources than with narrow, point-like sources. The relationship

between the energy spread and speech intelligibility could be accounted for by using the better-ear listening component in a binaural speech intelligibility model, while the binaural unmasking component did not contribute.

Finally, visual information was shown to improve sound source localization accuracy in virtual reality. Virtual reality glasses were shown to affect the acoustic field around the head and alter the HRTFs, which decreased localization accuracy. When visual information was presented in VR, sound localization accuracy improved similarly as observed in real environments. The lowest accuracy was found when the subjects were blind-folded and a significant improvement was found when room and hand position information were provided. An additional laser pointer for pointing guidance did not lead to an improvement of azimuth localization but an improved elevation localization.

## 5.2 Virtual environments for hearing research

Virtual sound environments have previously been shown to be a valuable tool for hearing research as they allow the creation of highly realistic and reproducible sound scenes with a large number of sound sources (Cubick and Dau, 2016; Favrot and Buchholz, 2010; Oreinos and Buchholz, 2016). In the following, capabilities, opportunities and limitations of virtual environments for hearing research are discussed.

### 5.2.1 Headphone playback

The acoustic scenes in this thesis were reproduced using a loudspeaker array. The advantage of loudspeaker-based reproduction over headphone-based reproduction is that head movements are inherently possible and ear-worn devices, such as hearing aids, can be used. To use heaphones for realistic reproductions, individual HRTFs need to be measured and applied. However, individual measurements of HRTFs are still cumbersome and expensive. On the other hand, loudspeaker-based virtual sound environments are also not feasible or practical in all research environments. For example, for clinical studies or entertainment purposes, headphones or smaller loudspeaker setups, e.g. using cross-talk cancellation systems (e.g. Akeroyd et al., 2007; Pausch et al., 2018), are preferable. To make headphone-based reproductions feasible, a solution to personalize HRTFs is needed. Together with virtual reality glasses,

such headphone-based approaches would allow a reproduction of a low-cost audio-visual cocktail party for everybody.

### 5.2.2 Room acoustic simulations

When investigating speech perception in virtual sound environments, errors can be introduced by multiple sources in the processing chain. The scene capture employing room acoustic simulations introduces two main error sources. First, when employing room acoustic simulations, the acoustic properties of the surfaces need to be measured or estimated. Since measurements or exact values of the acoustic properties of the materials are often not available, absorption coefficients as well as scattering coefficients need to be estimated. These estimations often need to be based on the experience of the researcher or acoustician. To improve the estimations, algorithms have been developed to adjust the surface properties using measurements of room acoustic features (Christensen et al., 2014). These algorithms generally take only global measures, such as the reverberation time or clarity, into consideration and not single reflections. Thus, the optimization might not necessarily converge to the correct materials as multiple optimization solutions can lead to the same result of reverberation time or clarity. Furthermore, it is possible that materials that occur in the room are found, but get assigned to different surfaces as spatial information regarding single reflections is not considered in the optimization algorithm.

The second major error source that originates from room acoustic simulations is the simulation process itself and the simplifications considered therein. The method applied in this thesis is based on a hybrid approach of the image-source method and stochastic approaches (Naylor, 1993). Even though this method does not lead to physically exact representations of reverberant spaces, the predictions of classical room acoustic parameters have been shown to be close to just-noticeable differences (Bork, 2000, 2005a,b; Vorländer, 1995). However, other methods that simulate room acoustics in a physically more accurate way might be favourable. These methods include finite- and boundary element methods or methods including phase information in addition to energy information (Marbjerg et al., 2015). Nevertheless, it is unclear if more advanced room acoustic simulations also lead to perceptually different results.

### 5.2.3 Ambisonics playback

When using ambisonics reproduction, a finite number of loudspeakers leads to errors in the reproduction at high frequencies (Ward and Abhayapala, 2001). The spectral error can be perceivable, particularly by trained listeners, but it may not affect speech intelligibility.

In chapter 2, no differences in speech intelligibility were found between the NLM and the HOA methods using the room acoustic simulation. Even though, the binaural speech intelligibility model predicted a slightly larger better-ear SNR advantage for the NLM method than for the HOA method when the interferers were asymmetrically distributed. When using NLM, the direct sound as well as the early reflections are reproduced using single loudspeakers, while a larger number of loudspeakers is used when using HOA. The wider spread of energy of HOA in comparison to NLM might be the reason for the lower better-ear SNR advantage when HOA is applied.

A comparable effect was obtained in chapter 3, where speech intelligibility was measured using a range of ambisonics orders. It was shown that the speech intelligibility increased with increasing ambisonics order. Since the spread of energy is lower for higher ambisonics orders, the physical difference between NLM and HOA becomes smaller as the order is increased.

### 5.2.4 Ambisonics capture

When using microphone arrays to capture an acoustic scene, the errors due to the finite ambisonics order occur twice, once at the capture side and once at the playback side. Oreinos (2015) showed that the increase in pressure error is mainly seen at high frequencies and at the contralateral ear (opposite ear relative to the sound source) when applying both capture and reproduction as opposed to reproducing simulated sound fields. Oreinos (2015) simulated both the capture and the playback stages and other error sources, such as positioning errors and phase/magnitude difference of the transducers, were not taken into account. Errors introduced by reflections in the playback room were also not considered. However, it is difficult to estimate what effect the errors have on outcome measures such as speech intelligibility. Some errors might lead to better and some errors to worse speech intelligibility. Thus, these errors might eventually not be detectable in speech intelligibility experiments. In chapter 2, the microphone array-recording method led to the most accurate

results as compared to the reference room. While the error due to the ambisonics processing is larger when using both ambisonics capture and playback, the microphone array-based reproduction leads to the physically most accurate reproduction regarding the methods considered in this thesis as it is the only method that attempts to directly reproduce the sound field.

On the capture side there is an additional problem which might affect speech intelligibility. The array size and the regularization method introduce a frequency-dependent directivity, i.e. effectively lowering orders at low frequencies (Favrot and Marschall, 2012). As described in chapter 3, a lower ambisonics orders result in a wider energy spread and thus lead to a reduced speech intelligibility due to a lower better-ear SNR advantage. On the other hand, better-ear listening mainly occurs at high frequencies and might therefore not be influenced by the effective ambisonics order at low frequencies.

Oreinos (2015) showed that the ambisonics pressure error is reduced in reverberant environments compared to an anechoic environment due to lower spatial variations caused by the diffuse sound. A larger impact of errors in anechoic than in reverberant environments was also shown in chapter 3, where the difference in speech intelligibility between the ambisonics orders was smaller when reverberation was present.

### 5.2.5 Speech as an outcome measure

To investigate and detect errors in the acoustic reproduction system, speech intelligibility might not be the most accurate measure as speech is robust and its perception dependents on many factors (Bronkhorst, 2000). Instead, traditional psychoacoustic tasks that are more sensitive to a certain feature in the sound might be beneficial. For example, to test differences in perceived reverberation, modulation depth discrimination experiments might reveal differences that are not reflected in speech intelligibility experiments.

### 5.2.6 Applicability of virtual reality glasses for hearing research

In this thesis, it has been shown that localization accuracy is worse when subjects wear virtual reality glasses. The reduced accuracy occurs partly due to altered HRTFs and partly due to a different movement behavior with the virtual reality glasses. However, the increased error was small in comparison to common distances between talkers in realistic environments. Thus, while one

needs to be cautious when conducting sound localization experiments with loudspeakers in virtual reality, the perception in more realistic situations might not be affected by the presentation in virtual reality.

## 5.3 Towards a realistic audio-visual cocktail party

In this thesis, virtual acoustic and visual information were mainly considered independently. However, to investigate a realistic cocktail party-like scenario in the laboratory, the two virtual worlds need to be combined. One of the challenges in the reproduction of virtual audio-visual talkers is the realistic animation of head and body motion behavior as well as lip movements. While systems for the reproduction of lip movements for hearing research have been proposed recently (Devesse et al., 2018; Hendrikse et al., 2018), they still remain unnatural as humans are highly sensitive to inaccurate visualizations of facial features. In the movie industry, the visualization of motion and facial expressions is further developed but requires a large amount of manual work (Lewis et al., 2014). For computer graphics, recent work showed promising progress of automatic lip movement simulations that might become available for a larger audience and for research (Joo et al., 2019; Lombardi et al., 2018; Wu et al., 2018) which would finally allow for realistic virtual audio-visual speech.

Virtual reality glasses and virtual sound environments allow for a higher realism of the stimuli compared to classical laboratory experiments. However, the task of the listeners often remains unnatural. For example, in speech experiments, listeners are commonly asked to repeat what they heard, either orally or via a user-interface. This task differs from typical real-world behavior. Thus, more realistic speech tests and tasks are also needed. Recently, a speech perception experiment was proposed, where the task was to answer questions based on a speech stimulus instead of simply repeating the stimulus (Best, 2016). Moreover, Weller et al. (2016) suggested a method to locate and identify a number of talkers in a multi-talker environment. Similarly, Stecker et al. (2018) used an audio-visual virtual environment in which listeners were asked to identify a certain talker out of a multi-talker mixture. While these tasks are better for testing real-world perception than traditional speech tests in some ways, no current test for perception in highly realistic audio-visual cocktail party-like scenes exists. An ideal measure would incorporate the same tasks and would elicit the same behavior that a listener has in a real cocktail party-like scenario.

To create a realistic audio-visual cocktail party, three components are needed. Foremost, the audio needs to be reproduced accurately in terms of temporal, spectral and spatial details. This can be achieved to a certain degree as discussed throughout this thesis. Second, matching visual information needs to be correctly presented. This includes the periphery, such as a room and objects within the room, as well as other people, including their lip movements and motion. At an actual cocktail party, the protagonists interact with each other. Interactions can be acoustic, through conversations, as well as visual, through head, hand and eye movements. More research is needed to determine which interactive and visual features affect perception and behavior. Additionally, more development is needed to create the tools for such an interactive audio-visual cocktail party in the laboratory.

# Bibliography


ANSI (2017). *Methods for Calculation of the Speech Intelligibility Index (S3.5-1997)*.

Agus, T. R., M. A. Akeroyd, S. Gatehouse, and D. Warden (2009). "Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise". In: *The Journal of the Acoustical Society of America* 126.4, p. 1926.

Ahrens, A. (2018). *Room acoustics model of a listening room [Dataset]*.

Ahrens, A., M. Marschall, and T. Dau (2019). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments". In: *Hearing Research*.

Akeroyd, M. A. et al. (2007). "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics". In: *The Journal of the Acoustical Society of America* 121.2, pp. 1056–1069.

Ando, Y. (2007). "Concert Hall Acoustics Based on Subjective Preference Theory". In: *Springer Handbook of Acoustics*. Ed. by T. Rossing. New York, NY: Springer New York, pp. 351–386.

Arweiler, I. and J. M. Buchholz (2011). "The influence of spectral characteristics of early reflections on speech intelligibility". In: *The Journal of the Acoustical Society of America* 130.2, pp. 996–1005.

Barron, M. and A. H. Marshall (1981). "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure". In: *Journal of Sound and Vibration* 77.2, pp. 211–232.

Batteau, D. W. (1967). "The Role of the Pinna in Human Localization". In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 168.1011, pp. 158–180.

Batteau, D. W. (1968). "Role of the Pinna in Localization: Theoretical and Physiological Consequences". In: *Hearing Mechanisms in Vertebrates*, pp. 234–243.

Baumgartner, R., P. Majdak, and B. Laback (2014). “Modeling sound-source localization in sagittal planes for human listeners”. In: *The Journal of the Acoustical Society of America* 136.2, pp. 791–802.

Behrens, T., T. Neher, and B. Johannesson (2007). “Evaluation of a Danish speech corpus for assessment of spatial unmasking”. In: *Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007)*. Isaar, pp. 449–457.

Bertet, S., J. Daniel, L. Gros, and E. Parizet (2007). “Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3D microphones”. In: *Audio Engineering Society Convention*, pp. 1–9.

Bertet, S., J. Daniel, E. Parizet, and O. Warusfel (2013). “Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources”. In: *Acta Acustica united with Acustica* 99.4, pp. 642–657.

Berzborn, M., R. Bomhardt, J. Klein, J. G. Richter, and M. Vorländer (2017). “The ITA-Toolbox : An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing”. In: *Fortschritte der Akustik*, pp. 222–225.

Best, V. (2016). “A Flexible Question-and-Answer Task for Measuring Speech Understanding”. In: *Trends in Hearing* 20, pp. 1–8.

Best, V., N. Marrone, C. R. Mason, and G. Kidd Jr (2012). “The influence of non-spatial factors on measures of spatial release from masking”. In: *The Journal of the Acoustical Society of America* 131.4, pp. 3103–3110.

Best, V., E. R. Thompson, C. R. Mason, and G. Kidd (2013). “An energetic limit on spatial release from masking”. In: *JARO - Journal of the Association for Research in Otolaryngology* 14.4, pp. 603–610.

Beutelmann, R. and T. Brand (2006). “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners”. In: *The Journal of the Acoustical Society of America* 120.1, pp. 331–342.

Beutelmann, R., T. Brand, and B. Kollmeier (2010). “Revision, extension, and evaluation of a binaural speech intelligibility model”. In: *The Journal of the Acoustical Society of America* 127.4, pp. 2479–2497.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT Press.

Blauert, J. and W. Lindemann (1986a). “Auditory spaciousness : Some further psychoacoustic analyses”. In: *J. Acoust. Soc. Am.* 80.2, pp. 533–542.

Blauert, J. and W. Lindemann (1986b). “Spatial mapping of intracranial auditory events for various degrees of interaural coherence”. In: *J. Acoust. Soc. Am.* 79.3, pp. 806–813.

Bork, I. (2000). “A comparison of room simulation software-The 2nd round robin on room acoustical computer simulation”. In: *Acta Acustica united with Acustica* 86.6, pp. 943–956.

Bork, I. (2005a). “Report on the 3rd Round Robin on Room Acoustical Computer Simulation - Part I: Measurement”. In: *Acta Acustica united with Acustica* 91.4, pp. 740–752.

Bork, I. (2005b). “Report on the 3rd Round Robin on Room Acoustical Computer Simulation - Part II: Calculations”. In: *Acta Acustica united with Acustica* 91.4, pp. 753–763.

Borrego, A., J. Latorre, M. Alcañiz, and R. Llorens (2018). “Comparison of Oculus Rift and HTC Vive: Feasibility for Virtual Reality-Based Exploration, Navigation, Exergaming, and Rehabilitation”. In: *Games for Health Journal* 7.2, g4h.2017.0114.

Bradley, J. S. (2011). “Review of objective room acoustics measures and future needs”. In: *Applied Acoustics* 72.10, pp. 713–720.

Bradley, J. S., H. Sato, and M. Picard (2003). “On the importance of early reflections for speech in rooms”. In: *The Journal of the Acoustical Society of America* 113.6, p. 3233.

Bradley, J., R Reich, and S. Norcross (1999). “A just noticeable difference in C50 for speech”. In: *Applied Acoustics* 58.2, pp. 99–108.

Brand, A., O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe (2002). “Precise inhibition is essential for microsecond interaural time difference coding.” In: *Nature* 417.6888, pp. 543–7.

Bronkhorst, A. W. (2000). “The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions”. In: *Acta Acustica united with Acustica* 86.1, pp. 117–128.

Brungart, D. S. and N. Iyer (2012). “Better-ear glimpsing efficiency with symmetrically-placed interfering talkers”. In: *The Journal of the Acoustical Society of America* 132.4, pp. 2545–2556.

Brungart, D. S., B. D. Simpson, M. A. Ericson, and K. R. Scott (2001). “Informational and energetic masking effects in the perception of multiple simultaneous talkers”. In: *The Journal of the Acoustical Society of America* 110.5, pp. 2527–2538.

Carlile, S., P Leong, S Hyams, and D Pralong (1997). *The nature and distribution of errors in the localization of sounds in humans*.

Chabot-Leclerc, A., E. N. MacDonald, and T. Dau (2016). "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain". In: *The Journal of the Acoustical Society of America* 140.1, pp. 192–205.

Cherry, E. C. (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *The Journal of the Acoustical Society of America* 25.5, pp. 975–979.

Christensen, C. L., G. Koutsouris, and J. H. Rindel (2014). "Estimating absorption of materials to match room model against existing room using a genetic algorithm". In: *Forum Acusticum*. Krakow.

Coleman, P. D. (1962). "Failure to Localize the Source Distance of an Unfamiliar Sound". In: *The Journal of the Acoustical Society of America* 34.3, pp. 345–346.

Cubick, J. and T. Dau (2016). "Validation of a Virtual Sound Environment System for Testing Hearing Aids". In: *Acta Acustica united with Acustica* 102.3, pp. 547–557.

Cubick, J., J. M. Buchholz, V. Best, M. Lavandier, and T. Dau (2018). "Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners". In: *The Journal of the Acoustical Society of America* 144.5, pp. 2896–2905.

Culling, J. F. and E. R. Mansell (2013). "Speech intelligibility among modulated and spatially distributed noise sources". In: *The Journal of the Acoustical Society of America* 133.4, pp. 2254–2261.

Culling, J. F., M. L. Hawley, and R. Y. Litovsky (2004). "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources". In: *The Journal of the Acoustical Society of America* 116.2, pp. 1057–1065.

Culling, J. F., M. L. Hawley, and R. Y. Litovsky (2005). "Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]". In: *The Journal of the Acoustical Society of America* 118.1, pp. 552–552.

Daniel, J. (2001). "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia". PhD thesis. Université Paris 6.

Devesse, A., A. Dudek, A. van Wieringen, and J. Wouters (2018). “Speech intelligibility of virtual humans”. In: *International Journal of Audiology* 57.12, pp. 908–916.

Dobreva, M. S., W. E. O’Neill, and G. D. Paige (2011). “Influence of aging on human sound localization”. In: *Journal of Neurophysiology* 105.5, pp. 2471–2486.

Dufour, A., O. Després, and T. Pebayle (2002). “Visual and auditory facilitation in auditory spatial localization”. In: *Visual Cognition* 9.6, pp. 741–753.

Duquesnoy, A. J. (1983). “Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons”. In: *The Journal of the Acoustical Society of America* 74.3, pp. 739–743.

Duquesnoy, A. J. and R. Plomp (1980). “Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis”. In: *The Journal of the Acoustical Society of America* 68.2, pp. 537–544.

Durlach, N. I. (1963). “Equalization and Cancellation Theory of Binaural Masking-Level Differences”. In: *The Journal of the Acoustical Society of America* 35.8, pp. 1206–1218.

Durlach, N. I. (1972). “Binaural signal detection Equalization and cancellation theory”. In: *Binaural signal detection Equalization and cancellation theory*, pp. 371–460.

Ege, R., A. J. V. Opstal, and M. M. Van Wanrooij (2018). “Accuracy-Precision Trade-off in Human Sound Localisation”. In: *Scientific Reports* 8.1, pp. 1–12.

Epain, N. et al. (2010). “Objective evaluation of a three-dimensional sound field reproduction system”. In: *Proceedings of 20th International Congress on Acoustics*. Sydney.

Ewert, S. D., W. Schubotz, T. Brand, and B. Kollmeier (2017). “Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers”. In: *The Journal of the Acoustical Society of America* 142.1, pp. 12–28.

Favrot, S. and J. M. Buchholz (2010). “LoRA: A loudspeaker-based room auralization system”. In: *Acta Acustica united with Acustica* 96.2, pp. 364–375.

Favrot, S. and J. M. Buchholz (2009). “Validation of a loudspeaker-based room auralization system using speech intelligibility measures”. In: *126th AES Convention*. Munich.

Favrot, S. and M. Marschall (2012). “Metrics for performance assessment of mixed-order Ambisonics spherical microphone arrays”. In: *25th Conference: Spatial Audio in Today’s 3D World*.

Fisher, H. G. and S. J. Freedman (1968). “The role of the pinna in auditory localization.” In: *Journal of Auditory Research* 8.1, pp. 15–26.

Frank, M. (2013). “Source Width of Frontal Phantom Sources: Perception, Measurement, and Modeling”. In: *Archives of Acoustics* 38.3, pp. 311–319.

Freyman, R. L., K. S. Helfer, D. D. McCall, and R. K. Clifton (1999). “The role of perceived spatial separation in the unmasking of speech”. In: *Journal of the Acoustical Society of America* 106.6, pp. 3578–3588.

Genovese, A., G. Zalles, G. Reardon, and A. Roginska (2018). “Acoustic perturbations in HRTFs measured on Mixed Reality Headsets”. In: *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond.

Gerzon, M (1992). “General Metatheory of Auditory Localisation”. In: *92nd Convention of Audio Engineering Society*. Vienna.

Gerzon, M. (1973). “Periphony: With-Height Sound Reproduction”. In: *Journal of the Audio Engineering Society* 21.1, pp. 2–10.

Gil-Carvajal, J. C., J. Cubick, S. Santurette, and T. Dau (2016). “Spatial Hearing with Incongruent Visual or Auditory Room Cues”. In: *Scientific Reports* 6.

Glasberg, B. R. and B. C. Moore (1990). “Derivation of auditory filter shapes from notched-noise data.” In: *Hearing research* 47.1-2, pp. 103–38.

Glyde, H., J. M. Buchholz, and H. Dillon (2013). “The importance of interaural time differences and level differences in spatial release from masking”. In: *The Journal of the Acoustical Society of America* 134.EL147.

Griesinger, D. (1997). “The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces”. In: *Acta Acustica* 83.4, pp. 721–731.

Grimm, G., B. Kollmeier, and V. Hohmann (2016). “Spatial Acoustic Scenarios in Multichannel Loudspeaker Systems for Hearing Aid Evaluation”. In: *Journal of the American Academy of Audiology* 27.7, pp. 557–566.

Gupta, R., R. Ranjan, J. He, and W.-S. Gan (2018). “Investigation of effect of VR/AR headgear on Head related transfer functions for natural listening”. In: *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond.

Hafter, E. R. and J. De Maio (1975). “Difference thresholds for interaural delay”. In: *The Journal of the Acoustical Society of America* 57.1, pp. 181–187.

Hafter, E. R., R. H. Dye, J. M. Neutzel, and H. Aronow (1977). “Difference thresholds for interaural intensity”. In: *The Journal of the Acoustical Society of America* 61.3, pp. 829–834.

Hassager, H. G., A. Wiinberg, and T. Dau (2017). “Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment”. In: *The Journal of the Acoustical Society of America* 141.4, pp. 2556–2568.

Hebrank, J. and D. Wright (1974). “Spectral cues used in the localization of sound sources on the median plane”. In: *The Journal of the Acoustical Society of America* 56.6, pp. 1829–1834.

Hendrikse, M. M., G. Llorach, G. Grimm, and V. Hohmann (2018). “Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters”. In: *Speech Communication* 101.June 2017, pp. 70–84.

Hering, E. (1861). *Beitraege zur Physiologie*. Leipzig: Verlag von Wilhelm Engelmann.

Hirsh, I. J. (1948). “The Influence of Interaural Phase on Interaural Summation and Inhibition”. In: *The Journal of the Acoustical Society of America* 20.4, pp. 536–544.

Hofman, P. M., J. G. Van Riswick, and A. J. Van Opstal (1998). “Relearning Sound Localization with New Ears”. In: *Nature Neuroscience* 1.5, pp. 417–421.

Holway, A. H. and E. G. Boring (1941). “Determinants of Apparent Visual Size with Distance Variant”. In: *The American Journal of Psychology* 54.1, pp. 21–37.

Houtgast, T., H. J. M. Steeneken, and R. Plomp (1980). “Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics”. In: *Acta Acustica united with Acustica* 46.1, pp. 60–72.

Howard, I. P. and W. B. Templeton (1966). *Human spatial orientation*. Oxford, England: John Wiley & Sons, p. 533.

IEC 268-13 (1985). *Sound System Equipment Part 13: Listening Tests on Loudspeaker*. Geneva.

ISO26101 (2012). *Acoustics - Test methods for the qualification of free-field environments*.

Jelfs, S., J. F. Culling, and M. Lavandier (2011). “Revision and validation of a binaural model for speech intelligibility in noise”. In: *Hearing Research* 275.1-2, pp. 96–104.

Joo, H. et al. (2019). “Panoptic Studio: A Massively Multiview System for Social Interaction Capture”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1, pp. 190–204.

Koski, T., V. Sivonen, and V. Pulkki (2013). “Measuring Speech Intelligibility in Noisy Environments Reproduced with Parametric Spatial Audio”. In: *Audio Engineering Society 135th Convention*. New York.

Kuznetsova, A., R. H. B. Christensen, C. Bavay, and P. B. Brockhoff (2014). “Automated mixed ANOVA modeling of sensory and consumer data”. In: *Food Quality and Preference* 40.PA, pp. 31–38.

Lavandier, M. and J. F. Culling (2007). “Speech segregation in rooms: Effects of reverberation on both target and interferer”. In: *The Journal of the Acoustical Society of America* 122.3, pp. 1713–1723.

Lavandier, M. and J. F. Culling (2010). “Prediction of binaural speech intelligibility against noise in rooms”. In: *The Journal of the Acoustical Society of America* 127.1, pp. 387–399.

Lavandier, M., S. Jelfs, J. F. Culling, A. Watkins, A. Raimond, and S. Makin (2012). “Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources”. In: *Journal of the acoustical society of america* 131.1, pp. 218–230.

Lőcsei, G., S. Santurette, T. Dau, and E. Macdonald (2017). “Lateralized speech perception with small ITDs in normal-hearing and hearing-impaired listeners”. In: *Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR): Adaptive Processes in Hearing*. Vol. 6.

Lenth, R. V. (2016). “Using lsmeans”. In: pp. 1–43.

Lewis, J., K. Anjyo, R. Taehyun, M. Zhang, F. Pighin, and Z. Deng (2014). “Practice and Theory of Blendshape Facial Models”. In: *Eurographics (State of the Art Reports)* 1.8.

Licklider, J. C. R. (1948). “The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise”. In: *The Journal of the Acoustical Society of America* 20.2, pp. 150–159.

Lochner, J. and J. Burger (1964). “The influence of reflections on auditorium acoustics”. In: *Journal of Sound and Vibration* 1.4, pp. 426–454.

Lombardi, S., J. Saragih, T. Simon, and Y. Sheikh (2018). “Deep appearance models for face rendering”. In: *ACM Transactions on Graphics* 37.4, pp. 1–13.

Macpherson, E. A. and J. C. Middlebrooks (2002). “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited”. In: *The Journal of the Acoustical Society of America* 111.5, p. 2219.

Macpherson, E. A. and A. T. Sabin (2013). “Vertical-plane sound localization with distorted spectral cues”. In: *Hearing Research* 306, pp. 76–92.

Maddox, R. K., D. A. Pospisil, G. C. Stecker, and A. K. Lee (2014). “Directing Eye Gaze Enhances Auditory Spatial Cue Discrimination”. In: *Current Biology* 24.7, pp. 748–752.

Majdak, P., M. J. Goupell, and B. Laback (2010). “3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training”. In: *Attention, Perception, & Psychophysics* 72.2, pp. 454–469.

Makous, J. C. and J. C. Middlebrooks (1990). “Two-dimensional sound localization by human listeners”. In: *The Journal of the Acoustical Society of America* 87.5, pp. 2188–2200.

Marbjerg, G., J. Brunskog, C.-H. Jeong, and E. Nilsson (2015). “Development and validation of a combined phased acoustical radiosity and image source model for predicting sound fields in rooms”. In: *The Journal of the Acoustical Society of America* 138.3, pp. 1457–1468.

Marschall, M. (2014). “Capturing and reproducing realistic acoustic scenes for hearing research”. PhD thesis. Technical University of Denmark.

Marschall, M., S. Favrot, and J. M. Buchholz (2012). “Robustness of a mixed-order Ambisonics microphone array for sound field reproduction”. In: *Audio Engineering Society 132nd Convention, Budapest, Hungary, 2012 April 26–29*, pp. 1–11.

Martin, R. L., K. I. McAnally, R. S. Bolia, G. Eberle, and D. S. Brungart (2012). “Spatial release from speech-on-speech masking in the median sagittal plane”. In: *The Journal of the Acoustical Society of America* 131.1, pp. 378–385.

Mason, R., F. Rumsey, and B. De Bruyn (2001). “An investigation of interaural time difference fluctuations, part 1: the subjective spatial effect of fluctuations delivered over headphones”. In: *110th AES Convention*. Amsterdam.

McGregor, P., A. G. Horn, and M. A. Todd (1985). “Are Familiar Sounds Ranged More Accurately?” In: *Perceptual and Motor Skills* 61.3_suppl, pp. 1082–1082.

McGurk, H. and J. MacDonald (1976). “Hearing lips and seeing voices”. In: *Nature* 264.5588, pp. 746–748.

Meyer, J. and G. W. Elko (2004). “Spherical Microphone Arrays for 3D Sound Recording”. In: *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Ed. by Y. Huang and J. Benesty. Boston: Springer, pp. 67–89.

Middlebrooks, J. C. (1992). “Narrow-band sound localization related to external ear acoustics”. In: *The Journal of the Acoustical Society of America* 92.5, pp. 2607–2624.

Middlebrooks, J. C. (1999). “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency”. In: *The Journal of the Acoustical Society of America* 106.3, pp. 1493–1510.

Middlebrooks, J. C., J. Z. Simon, A. N. Popper, and R. R. Fay, eds. (2017). *The Auditory System at the Cocktail Party*. Vol. 60. Springer Handbook of Auditory Research. Cham: Springer International Publishing.

Mills, A. W. (1960). “Lateralization of High-Frequency Tones”. In: *The Journal of the Acoustical Society of America* 32.1, pp. 132–134.

Minnaar, P., S. Favrot, and J. M. Buchholz (2010). “Improving hearing aids through listening tests in a virtual sound environment”. In: *Hearing Journal* 63.10, pp. 40–44.

Monaghan, J. J. M., K. Krumbholz, and B. U. Seeber (2013). “Factors affecting the use of envelope interaural time differences in reverberation”. In: *The Journal of the Acoustical Society of America* 133.4, pp. 2288–2300.

Müller, S. and P. Massarani (2001). “Transfer-Function Measurement with Sweeps”. In: *Journal of the Audio Engineering Society* 49.6, pp. 443–471.

Musicant, A. D. and R. A. Butler (1984). “The influence of pinnae-based spectral cues on sound localization”. In: *The Journal of the Acoustical Society of America* 75.4, pp. 1195–1200.

Nam, J, J. Abel, and J. S. Iii (2008). “A method for estimating interaural time difference for binaural synthesis”. In: *Audio Engineering Society Convention 125*, pp. 0–6.

Naylor, G. M. (1993). “ODEON-Another hybrid room acoustical model”. In: *Applied Acoustics* 38.2-4, pp. 131–143.

Niehorster, D. C., L. Li, and M. Lappe (2017). “The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research”. In: *i-Perception* 8.3, p. 204166951770820.

Oldfield, S. R. and S. P. a. Parker (1984). “Acuity of Sound Localization: A Topography of Auditory Space. I. Normal Hearing Conditions”. In: *Perception* 13.5, pp. 581–600.

Oreinos, C. and J. M. Buchholz (2013). “Measurement of a full 3D set of HRTFs for in-ear and hearing aid microphones on a head and torso simulator (HATS)”. In: *Acta Acustica united with Acustica* 99.5, pp. 836–844.

Oreinos, C. and J. M. Buchholz (2016). “Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids”. In: *Journal of the American Academy of Audiology* 27.7, pp. 541–556.

Oreinos, C. (2015). “Virtual Acoustic Environments for the Evaluation of Hearing Devices Christos Oreinos”. PhD thesis.

Pausch, F., L. Aspöck, M. Vorländer, and J. Fels (2018). “An Extended Binaural Real-Time Auralization System With an Interface to Research Hearing Aids for Experiments on Subjects With Hearing Loss”. In: *Trends in Hearing* 22, p. 233121651880087.

Perrett, S. and W. Noble (1997). “The effect of head rotations on vertical plane sound localization”. In: *The Journal of the Acoustical Society of America* 102.4, pp. 2325–2332.

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)”. In: *Acta Acustica united with Acustica* 34.4, pp. 200–211.

Plomp, R. and A. M. Mimpen (1979). “Improving the reliability of testing the speech reception threshold for sentences”. In: *International Journal of Audiology* 18.1, pp. 43–52.

Poletti, M. A. (2005). “Three-dimensional surround sound systems based on spherical harmonics”. In: *AES: Journal of the Audio Engineering Society* 53.11, pp. 1004–1024.

Politis, A. (2016). “Microphone array processing for parametric spatial audio techniques”. PhD thesis. Aalto University.

Pulkki, V. (2007). “Spatial Sound Reproduction with Directional Audio Coding”. In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516.

Rayleigh, L. (1907). “On our Perception of Sound Direction”. In: *Philosophical Magazine* 13, pp. 214–232.

Recanzone, G. H. (2009). “Interactions of auditory and visual stimuli in space and time”. In: *Hearing Research* 258.1-2, pp. 89–99.

Rennies, J., T. Brand, and B. Kollmeier (2011). “Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet”. In: *The Journal of the Acoustical Society of America* 130.5, pp. 2999–3012.

Rønne, F. M., S. Laugesen, and N. S. Jensen (2017). “Selection of test-setup parameters to target specific signal-to-noise regions in speech-on-speech intelligibility testing”. In: *International Journal of Audiology* 56.8, pp. 559–567.

Schroeder, M. R., D. Gottlob, and K. F. Siebrasse (1974). “Comparative study of European concert halls: correlation of subjective preference with geometric and acoustic parameters”. In: *The Journal of the Acoustical Society of America* 56.4, pp. 1195–1201.

Seeber, B. U., S. Kerber, and E. R. Hafter (2010). “A system to simulate and reproduce audio–visual environments for spatial hearing research”. In: *Hearing Research* 260.1-2, pp. 1–10.

Shelton, B. R. and C. L. Searle (1980). “The influence of vision on the absolute identification of sound-source position.” In: *Perception & psychophysics* 28.6, pp. 589–596.

Shinn-Cunningham, B. G., N. I. Durlach, and R. M. Held (1998a). “Adapting to supernormal auditory localization cues. I. Bias and resolution”. In: *The Journal of the Acoustical Society of America* 103.6, pp. 3656–3666.

Shinn-Cunningham, B. G., N. I. Durlach, and R. M. Held (1998b). “Adapting to supernormal auditory localization cues. II. Constraints on adaptation of mean response”. In: *The Journal of the Acoustical Society of America* 103.6, pp. 3667–3676.

Soendergaard, P. and P. Majdak (2013). “The Auditory Modeling Toolbox”. In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer, pp. 33–56.

Solvang, A. (2008). “Spectral impairment for two-dimensional higher order ambisonics”. In: *AES: Journal of the Audio Engineering Society* 56.4, pp. 267–279.

Soulodre, G., N. Popplewell, and J. Bradley (1989). “Combined effects of early reflections and background noise on speech intelligibility”. In: *Journal of Sound and Vibration* 135.1, pp. 123–133.

Stecker, G. C., T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami (2018). “Toward objective measures of auditory co-immersion in virtual and aug-

mented reality". In: *Conference on Audio for Virtual and Augmented Reality*. Redmond.

Stitt, P., S. Bertet, and M. van Walstijn (2014). "Off-Centre localisation performance of ambisonics and HOA for large and small loudspeaker array radii". In: *Acta Acustica united with Acustica* 100, pp. 937–944.

Stitt, P., S. Bertet, and M. Van Walstijn (2016). "Extended energy vector prediction of ambisonically reproduced image direction at off-center listening positions". In: *AES: Journal of the Audio Engineering Society* 64.5, pp. 299–310.

Sumby, W. H. and I. Pollack (1954). "Visual Contribution to Speech Intelligibility in Noise". In: *The Journal of the Acoustical Society of America* 26.2, pp. 212–215.

Tabry, V., R. J. Zatorre, and P. Voss (2013). "The influence of vision on sound localization abilities in both the horizontal and vertical planes". In: *Frontiers in Psychology* 4.DEC, pp. 1–7.

Van Beers, R. J., A. C. Sittig, and J. J. Denier Van Der Gon (1999). "Localization of a seen finger is based exclusively on proprioception and on vision of the finger". In: *Experimental Brain Research* 125.1, pp. 43–49.

Van Opstal, A. J., J. Vliegen, and T. Van Esch (2017). "Reconstructing spectral cues for sound localization from responses to rippled noise stimuli". In: *PLoS ONE* 12.3, pp. 1–29.

Van Wanrooij, M. M. (2005). "Relearning Sound Localization with a New Ear". In: *Journal of Neuroscience* 25.22, pp. 5413–5424.

Vorländer, M. (1995). "International Round Robin on Room Acoustical Computer Simulations". In: *15th International Congress on Acoustics*. Trondheim, pp. 689–692.

Wagener, K., J. L. Josvassen, and R. Ardenkjær (2003). "Design, optimization and evaluation of a Danish sentence test in noise". In: *International Journal of Audiology* 42.1, pp. 10–17.

Wan, R., N. I. Durlach, and H. S. Colburn (2010). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers". In: *The Journal of the Acoustical Society of America* 128.6, pp. 3678–3690.

Wan, R., N. I. Durlach, and H. S. Colburn (2014). "Application of a short-time version of the Equalization-Cancellation model to speech intelligibility ex-

periments with speech maskers". In: *The Journal of the Acoustical Society of America* 136.2, pp. 768–776.

Ward, D. and T. Abhayapala (2001). "Reproduction of a plane-wave sound field using an array of loudspeakers". In: *IEEE Transactions on Speech and Audio Processing* 9.6, pp. 697–707.

Watson, C. S. (2005). "Some Comments on Informational Masking". In: *Acta Acustica united with Acustica* 91.2005, pp. 502–512.

Weller, T., V. Best, J. M. Buchholz, and T. Young (2016). "A Method for Assessing Auditory Spatial Analysis in Reverberant Multitalker Environments". In: *Journal of the American Academy of Audiology* 27.7, pp. 601–611.

Westermann, A. and J. M. Buchholz (2015). "The influence of informational masking in reverberant, multi-talker environments". In: *The Journal of the Acoustical Society of America* 138.2, pp. 584–593.

Whitmer, W., B. U. Seeber, and M. a. Akeroyd (2014). "The perception of apparent auditory source width in hearing-impaired adults." In: *The Journal of the Acoustical Society of America* 135.6, p. 3548.

Whitmer, W. M., B. U. Seeber, and M. A. Akeroyd (2012). "Apparent auditory source width insensitivity in older hearing-impaired individuals". In: *The Journal of the Acoustical Society of America* 132.1, pp. 369–379.

Wierstorf, H. (2014). "Perceptual Assessment of Sound Field Synthesis". PhD thesis. Fakultät IV – Elektrotechnik und Informatik, Technischen Universität Berlin, pp. 265 –292.

Wiggins, I. M. and B. U. Seeber (2011). "Dynamic-range compression affects the lateral position of sounds". In: *The Journal of the Acoustical Society of America* 130.6, pp. 3939–3953.

Wiggins, I. M. and B. U. Seeber (2012). "Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners". In: *Ear and Hearing* 33.3, pp. 399–410.

Wu, C., T. Shiratori, and Y. Sheikh (2018). "Deep incremental learning for efficient high-fidelity face tracking". In: *ACM Transactions on Graphics* 37.6, pp. 1–12.

Zahorik, P. (2002). "Assessing auditory distance perception using virtual acoustics". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1832–1846.

Zahorik, P., D. S. Brungart, and A. W. Bronkhorst (2005). "Auditory distance perception in humans: A summary of past and present research". In: *Acta Acustica united with Acustica* 91.3, pp. 409–420.

Zotter, F. and M. Frank (2012). “All-round ambisonic panning and decoding”. In: *AES: Journal of the Audio Engineering Society* 60.10, pp. 807–820.

Zotter, F., M. Frank, M. Kronlacher, and J.-W. Choi (2014). “Efficient phantom source widening and diffuseness in ambisonics”. In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics* 2.April, pp. 69–74.

Zurek, P. M. (1993). “Binaural advantages and directional effects in speech intelligibility”. In: *Acoustical factors affecting hearing aid performance* 2, pp. 255–175.

# Contributions to Hearing Research

**Vol. 1:** *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.
External examiners: Mark Lutman, Stefan Stenfeld

**Vol. 2:** *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.
External examiners: Brian Moore, Kathrin Krumbholz

**Vol. 3:** *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.
External examiners: Michael Akeroyd, Armin Kohlrausch

**Vol. 4:** *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.
External examiners: Jesko Verhey, Steven van de Par

**Vol. 5:** *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.
External examiners: Björn Hagerman, Ejnar Laukli

**Vol. 6:** *Helen Connor*, Hearing aid amplification at soft input levels, 2010.
External examiners: Inga Holube, Birgitta Larsby

**Vol. 7:** *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.
External examiners: Birger Kollmeier, Ray Meddis

**Vol. 8:** *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.
External examiners: David Kemp, Stephen Neely

**Vol. 9:** *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.
External examiners: Bernhard Seeber, Michael Vorländer

**Vol. 10:** *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.
External examiners: Christopher Plack, Christian Lorenzi

**Vol. 11:** *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.
External examiners: Joost Festen, Jürgen Tchorz

**Vol. 12:** *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.
External examiners: Bob Burkard, Stephen Neely

**Vol. 13:** *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.
External examiners: Stuart Rosen, Christian Lorenzi

**Vol. 14:** *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.
External examiners: Michael Stone, Oded Ghitza

**Vol. 15:** *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.
External examiners: John Culling, Martin Cooke

**Vol. 16:** *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.
External examiners: Lawrence Rosenblum, Matthias Gondan

**Vol. 17:** *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.
External examiners: Shihab Shamma, Guy Brown

**Vol. 18:** *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.
External examiners: Sascha Spors, Ville Pulkki

**Vol. 19:** *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.
External examiners: Bernhard Seeber, Steven van de Par

**Vol. 20:** *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.
External examiners: Christopher Plack, Enrique Lopez-Poveda

**Vol. 21:** *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.
External examiners: Steven van de Par, John Culling

**Vol. 22:** *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.
External examiners: Ingrid Johnsrude, Christian Lorenzi

**Vol. 23:** *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.
External examiners: Judy Dubno, Martin Cooke

**Vol. 24:** *Johannes Käsbach*, Characterizing apparent source width perception, 2016.
External examiners: William Whitmer, Jürgen Tchorz

**Vol. 25:** *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.
External examiners: Thomas Brand, Armin Kohlrausch

**Vol. 26:** *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.
External examiners: Laurel Carney, Bob Carlyon

**Vol. 27:** *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.
External examiners: Volker Hohmann, Piotr Majdak

**Vol. 28:** *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.
External examiners: Maria Chait, Christian Lorenzi

**Vol. 29:** *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.
External examiners: Ville Pulkki, Pavel Zahorik

**Vol. 30:** *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.
External examiners: Roland Schaette, Ian Bruce

**Vol. 31:** *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.
External examiners: Enrique Lopez-Poveda, Tim Jürgens

**Vol. 32:** *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.
External examiners: Armin Kohlrausch, James Kates

**Vol. 33:** *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.
External examiners: Stefan Bleeck, Jürgen Tchorz

**Vol. 34:** *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.
External examiners: Julie Arenberg, Olivier Macherey

**Vol. 35:** *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.
External examiners: Christophe Micheyl, Monita Chatterjee

**Vol. 36:** *Søren Fuglsang*, Characterizing neural mechanisms of attention-driven speech processing, 2019.
External examiners: Shihab Shamma, Maarten de Vos

**Vol. 37:** *Borys Kowalewski*, Assessing the effects of hearing-aid dynamic-range compression on auditory signal processing and perception, 2019.
External examiners: Brian Moore, Graham Naylor

**Vol. 38:** *Helia Relaño Iborra*, Predicting speech perception of normal-hearing and hearing-impaired listeners, 2019.
External examiners: Ian Bruce, Armin Kohlrausch

**Vol. 39:** *Axel Ahrens*, Characterizing auditory and audio-visual perception in virtual environments, 2019.
External examiners: Pavel Zahorik, Piotr Majdak

*The end.*

*To be continued…*

One of the challenges in hearing research is to explain the human ability to understand speech in complex, noisy environments, commonly referred to as a cocktail party scenario. To gain a better understanding of how the auditory system performs in complex acoustic environments, one approach is to reproduce such listening situations in the laboratory.

Virtual reality glasses and loudspeaker-based virtual sound environments are promising tools to bring the cocktail party into the laboratory. In this thesis, both of the tools were used to investigate auditory and audio-visual perception. It was found that the perception of speech in virtual sound environments is similar to real scenes, however, small differences in terms of speech intelligibility were found in some conditions. Virtual reality glasses were shown to lead to perturbations of the sound field around the head which affects the sound source localization accuracy when no visual information is available. When adding visual information, the accuracy increases.

Concluding, throughout this thesis it has been shown that virtual reality glasses and loudspeaker-based virtual sound environments are promising tools for the reproduction of realistic audio-visual scenarios. These realistic environments can be used for hearing research and to investigate new technologies such as hearing aids and other audio wearables.

**DTU Health Tech**
Department of Health Technology

Ørsteds Plads
Building 352
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 39 50
www.dtu.dk