Human Sound Externalization in Reverberant Environments

PhD thesis by
Jasmina Catic

Technical University of Denmark
2014

Preprint version for the assessment committee
Pagination will differ in the final published version.

This PhD-dissertation is the result of a research project at the Centre for Applied Hearing Research (CAHR), Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark). The work was completed in 2014.
The project was financed 1/2 by the Technical University of Denmark and 1/2 by the Danish hearing aid industry (GN ReSound A/S, Oticon A/S, and Widex A/S)

Supervisors

Prof. Torsten Dau
Assist. Prof. Sébastien Santurette
Centre for Applied Hearing Research (CAHR)
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Sébastien Santurette and Torsten Dau, for all their support, motivation, and guidance throughout the project. I would also like to thank Jörg Buchholz for his support and the discussions we had.

I also express my thanks to all the listeners who participated in my experiments and to Sonion for providing the miniature microphones necessary for the HRTF measurements.

Finally, I would like to thank my family, friends and colleagues for the good times in and outside of the office.

CONTENTS

ABSTRACT

In everyday environments, listeners perceive sound sources as externalized. In listening conditions where the spatial cues that are relevant for externalization are not represented correctly, such as when listening through headphones or hearing aids, a degraded perception of externalization may occur. In this thesis, the spatial cues that arise from a combined effect of filtering due to the head, torso, and pinna and the acoustic environment were analysed and the impact of such cues for the perception of externalization in different frequency regions was investigated. Distant sound sources were simulated via headphones using individualized binaural room impulse responses (BRIRs).

An investigation of the influence of spectral content of a sound source on externalization showed that effective externalization cues are present across the entire frequency range. The fluctuation of interaural level differences (ILDs) that occurs in reverberant environments was altered via modifications of the signal envelope in the left and right ear. It was found that the dynamic ILDs had an effect on externalization for broadband and highpass filtered speech, while no effect was found for lowpass filtered speech. Moreover, the compression of low frequency ITD fluctuations did not influence externalization.

Further, the influence of binaural and monaural cues from reverberation was investigated. It was found that monaural reverberation cues were sufficient for the externalization of a lateral source, whereas a frontal source required an increased amount of binaural cues from reflections in order to attain convincingly externalized sound images. It was concluded that the disparity in the interaural cues of the direct sound and the reverberation was important as the interaction of the two played a role for the perception of externalization. Moreover, similar binaural effects of reverberation were found at low- and high frequencies.

The performance of a multi-microphone noise reduction algorithm designed to preserve binaural cues in hearing aids was investigated in conjunction with a voice activity detector (VAD) for noise estimation. Intelligibility weighted improvements in signal-to-noise ratio (SNR) of 6 dB and 18 dB were found for diffuse multi-talker babble noise and speech shaped directional noise, respectively, at input SNRs close to the speech reception threshold (SRT) of hearing impaired listeners.

Overall, this work contributes to the understanding of the auditory processing of spatial cues that are important for externalization in reverberant environments and may have implications for hearing instrument signal processing.

RESUMÉ

I de naturlige akustiske miljøer bliver lydkilder opfattet som eksternaliserede. Under lytteforhold hvor de lydmæssige signalkarakteristika, som er vigtige for eksternalisering, ikke bliver gengivet korrekt, kan opfattelsen af eksternalisering blive forringet. Dette forekommer eksempelvis ved lytning gennem høretelefoner og ved brug af høreapparater. Formålet med dette projekt var at undersøge, hvordan de rumlige cues som bruges til identifikation af en lydkildes retning og afstand indvirker på eksternalisering. Specielt de cues der opstår grundet den kombinerede effekt af filtrering pga. hoved, torso og øre sammen med det akustiske miljø blev analyseret.

En undersøgelse af effekten af et signals båndbredde på eksternalisering viste at effektive cues til eksternalisering findes i hele frekvensområdet. Fluktuationerne i interaurale niveauforskelle, som de forekommer grundet efterklangen i et rum, blev modificeret ved manipulation af signalerne i det højre og venstre øre. De dynamiske interaurale niveauforskelle viste en effekt på lydkilder der indeholder høje frekvenser, mens der ikke blev fundet nogen effekt ved lavpasfiltreret tale.

Yderligere blev effekten af binaurale og monaurale cues fra efterklang undersøgt. De monaurale cues var tilstrækkelige til eksternalisering af en lydkilde som har stærke binaurale cues i den direkte lyd, mens en lydkilde med en lav grad af binaurale cues i den direkte lyd krævede en øget mængde binaurale cues fra refleksioner for at opnå overbevisende eksternaliserede lydbilleder. Det blev konkluderet at samspillet mellem de interaurale cues i den direkte lyd og efterklangen spillede en rolle for opfattelsen af eksternalisering.

En støjreduktionsalgoritme designet til at bevare binaurale cues i høreapparater blev undersøgt i forbindelse med en detektor for taleaktivitet (VAD) som bruges til estimering af støjen. Den taleforståelighedsvægtede forbedring af signal-støj-forholdet ved signal-støj-forhold som ligger tæt på hørehæmmedes talemodtagelsestærskel var henholdsvis 6 dB og 18 dB for diffus fluktuerende og retningsbestemt stationær støj.

Samlet set bidrager dette arbejde til forståelsen af den auditive behandling af rumlige informationer der er vigtige for eksternalisering i akustiske miljøer med efterklang og kan have betydning for udvikling af signalbehandlingsalgoritmer til høreapparater.

RELATED PUBLICATIONS

Journal articles

- Catic, J., Santurette, S., Buchholz, J. M., Gran, F., Dau, T. (2013). The effect of interaural level difference fluctuations on the externalization of Sound. *Journal of the Acoustical Society of America*, 134 (2), p.1232-1241
- Catic, J., Dau, T., Buchholz, J. M., Gran, F. (2010). The Effect of a Voice Activity Detector on the Speech Enhancement Performance of the Binaural Multichannel Wiener Filter. *EURASIP Journal on Audio, Speech, and Music Processing*. Article ID 840294

Conference papers

- Catic, J., Buchholz, J. M., Gran, F., Dau, T. (2011). The role of spectro-temporal fluctuations in interaural level differences for the externalization of sound. *Proceedings of the Forum Acusticum, Aalborg, Denmark*, June 2011.

Published abstracts

- Catic, J., Santurette, S., Buchholz, J. M., Dau, T. (2012). Effects of interaural level differences on the externalization of sound, *Acoustical Society of America (ASA), Hong Kong*, May 2012, p. 3520
- Dau, T., Catic, J., Santurette, S., Buchholz, J. M. (2012). Effects of interaural level and time differences on the externalization of sound. *35th Midwinter Meeting of the Association for Research in Otolaryngology, San Diego, CA, United States*. February 2012, Volume 35, p. 229

LIST OF ABBREVIATIONS

ADM	Adaptive Directional Microphone
AI-DI	AI weighted Directivity Index
ANOVA	Analysis of Variance
BMWF	Binaural Multichannel Wiener Filter
BRIR	Binaural Room Impulse Response
BRTF	Binaural Room Transfer Function
BTE	Behind the Ear
CLUE	Conversational Language Evaluation
D/R	Direct to Reverberant ratio
E-rating	Externalization rating
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
HATS	Head and Torso Simulator
HF	High-Frequency
HI	Hearing Impaired
HPIR	Headphone Impulse Response
HRTF	Head-Related Transfer Function
IC	Interaural Coherence
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
KEMAR	Knowles Electronics Manikin for Acoustic Research
LF	Low-Frequency
MMSE	Minimum Mean Square Error
MO	Condition with monaural reverberation
MVDR	Minimum Variance Distortionless Response
NH	Normal-Hearing
ROC	Receiver Operating Characteristic
SC	Speech Cancellation
SD	Standard Deviations
SI	Speech Intelligibility
SNR	Signal to Noise Ratio
SNR_{INT}	Intelligibility Weighted SNR
SPL	Sound Pressure Level

SRT	Speech Reception Threshold
TR	Truncated BRIR condition
VAD	Voice Activity Detector

1 GENERAL INTRODUCTION

The auditory system has a remarkable ability to analyse sounds in complex acoustic environments. This ability is clearly demonstrated in the well-known cocktail party effect (Cherry, 1953), where many talkers are active at the same time, yet the auditory system is able to enhance certain sound characteristics while suppressing unwanted information in order to facilitate speech understanding of the desired speaker. Another fascinating aspect of the human hearing is the way that acoustic information is utilized to provide the listener with a proper sense of auditory space in complex environments. Even when reverberation and multiple sound sources are present at the same time, the auditory system is able to extract the relevant acoustic information that enables identification of the direction and distance of a chosen sound source in space.

When a sound source arrives at a listener's ears, it is filtered by the head, torso and pinna. Due to this filtering, the sound received at the two ears contains differences in level and time, commonly known as the interaural level differences (ILDs), and the interaural time differences (ITDs). Furthermore, the spectrum of the incoming sound is boosted or attenuated in different frequency bands due to diffraction and scattering of the head and outer ears. The binaural cues (ITDs and ILDs) also differ across frequency. Hence, for each position in space there is a unique combination of different spatial cues, which is then mapped into a single source location by the auditory system. Such a combination of cues is defined by the head related transfer functions (HRTFs) at the two ears. The monaural spectral notches in the head related transfer function facilitate elevation specifically (Macpherson & Sabin, 2007), while the binaural cues generally facilitate horizontal localization (Blauert, 1997).

In our natural environments, both in rooms and outdoors, the sound received at a listener's ears does not only arrive from the direction of the active source, but also from other directions due to reflections from walls and objects. If the sound is continuous, which most of the sounds we normally encounter are, such as speech or music, the sound energy from the original source and reflections overlap, and thereby the spatial cues are modified in such a way that they do not resemble the cues from the original source direction. However, the auditory system has mechanisms that give greater weight to the first arriving wave front (the direct sound), while suppressing the localization information of the reflections, thereby making the localization of the actual sound source possible (Litovsky *et al.*, 1999). Still, this process has

limitations, and reverberation has been shown to reduce localization accuracy (Ihlefeld & Shinn-Cunningham, 2011) compared to sound presentation where direct sound dominates.

Contrary to the adverse effect that reverberation has on localization accuracy, other spatial attributes, such as distance perception, are greatly improved by the presence of reflections (*e.g.* Mershon & King, 1975; Nielsen S. H., 1993). In this case, the auditory system extracts information from the reflections and their interaction with the direct sound in a way that improves perception. The direct sound energy decreases as a function of distance while the reverberation energy remains fairly constant. Thereby, the ratio of the energy in the direct sound and reverberant sound (D/R) decreases as the source-receiver distance is increased, and can therefore be used as a cue to absolute sound source distance. In contrast, the HRTF related cues do not show changes as a function of source-receiver distance except in the special case of small source-receiver distances of less than 1 m (Brungart & Rabinowitz, 1999). Moreover, distance estimation in anechoic rooms is much less accurate and highly underestimated compared to that obtained in reverberant rooms.

When both HRTF related cues and the reverberation related cues are available to the listener in their natural form, the sound sources are perceived at their correct location with a clear perception of distance. Such sound sources are clearly perceived as being outside the head, i.e. they are perceived as externalized. For example, headphone reproduction that captures both the directional cues due to filtering of the head, torso, and pinna, and the cues that arise from the filtering of the environment, can produce very convincingly externalized sound images, that are indistinguishable to the natural sound source they are made to simulate. However, if a headphone presentation does not capture the required spatial cues, most often an internalized sound image is perceived. This would usually occur if frequency independent binaural cues are presented (Hartmann & Wittenberg, 1996) or reverberation is absent (Begault *et al.*, 2001). Although it is known that HRTF related cues and reverberation contribute to externalization, it is not clear which features of the HRTF filtered sound signals are crucial for externalization in reverberant environments, where complex patterns of cues are generated due to the interaction of reverberation with HRTF related localization cues.

Understanding the characteristic features of sound that facilitate externalization in natural environments and with sound source types that we normally encounter in everyday lives (e.g. speech) can be useful for hearing aid applications. In order to improve speech intelligibility, which is a major problem for hearing-impaired listeners in multi-talker environments, signal processing algorithms which either employ dynamic range compression or noise reduction are designed for hearing aids. However, it has been shown that hearing aid processing may distort binaural cues

and thereby the hearing aid user may lose the correct representation of the auditory space (Keidsler *et al.*, 2006). Some listeners then perform better without hearing aids in localization tasks compared to when wearing hearing aids (Van den Bogaert *et al.*, 2006). If it would be known which cues are essential for externalization, hearing aid algorithm design could either be designed to leave these cues undistorted after processing, or could enhance certain sound features in order to improve externalization. Furthermore, such knowledge could be relevant for applications where the synthesis of distant sound sources is desired, such as headphone synthesis of 3D sound for, e.g., computer games. In those types of applications, accurate measurements of the individual HRTFs and specific room environments are unrealistic. Thus, knowledge about which spatial cues need to be enhanced for improving externalization might be helpful and useful in various applications.

This thesis has its focus on the interaction between HRTF related binaural cues and reverberation, and their impact on externalization perception. Psychoacoustic experiments were performed using simulated externalized sound sources via headphones.

Chapter 2 investigates the impact of reverberation on the interaural level differences and their effect on externalization. A method for manipulating dynamic ILDs is developed which is based on envelope modifications where an auditory-based filterbank is used for analysis and synthesis. Psychoacoustic listening tests are performed with speech stimuli with modified ILDs and it is investigated how bandwidth reduction affects the externalization of the speech with natural and modified ILDs.

Chapter 3 focuses on the importance of monaural and binaural reverberation related cues. The detection of changes in how reverberant a sound source is perceived has been shown to depend on monaural cues (Larsen *et al.*, 2008). As reverberation plays an important role in externalization, it could be anticipated that a monaural presentation of the reverberant part of the sound along with a binaural presentation of HRTFs is sufficient for externalization. Here, binaural room impulse responses (BRIRs) are modified by first gradually increasing the amount of reflections and subsequently gradually increasing the amount of binaural reverberation when having identical reverberation in the two ears in the modified BRIR as the starting point. The dynamic binaural cues that arise from the interaction between the direct sound and reverberation are then analysed and compared to the obtained experimental data.

Chapter 4 investigates the performance of a multi-microphone binaural hearing aid algorithm for noise reduction. Often, hearing aid processing distorts binaural cues with adverse consequences for spatial perception (Keidsler *et al.*, 2006). The algorithm considered here is designed to preserve binaural cues and has been shown

to yield better spatial perception than conventional comparable algorithms (Van den Bogaert *et al.*, 2008). However, it requires a voice activity detector (VAD) for the estimation of interfering noise and, so far, its performance has been evaluated with an ideal VAD. It is well known that VADs only work well for certain types of noise and at high signal-to-noise ratios (SNRs) and their performance is significantly degraded in challenging acoustic environments, such as a cocktail party like situation. Here, the noise reduction performance is evaluated with an envelope-based VAD for different noise types and spatial scenarios in order to investigate how the degradation of the VAD performance at low signal-to-noise ratios for certain noise types affects the SNR improvement compared to less challenging scenarios where the VAD has negligible effects.

Chapter 5 summarizes the main findings of this study, discusses the implications of the obtained results for the understanding of the sound characteristics that are important for externalization and for possible applications where externalization is desirable or required, and suggests aspects for future research within the field.

2 THE EFFECT OF INTERAURAL-LEVEL-DIFFERENCE FLUCTUATIONS ON THE EXTERNALIZATION OF SOUND

Abstract

Real-world sound sources are usually perceived as externalized and thus properly localized in both direction and distance. This is largely due to 1) the acoustic filtering by the head, torso, and pinna, resulting in modifications of the signal spectrum and thereby a frequency-dependent shaping of interaural cues and 2) interaural cues provided by the reverberation inside an enclosed space. This study first investigated the effect of room reverberation on the spectro-temporal behaviour of interaural level differences (ILDs) by analysing dummy-head recordings of speech played at different distances in a standard listening room. Next, the effect of ILD fluctuations on the degree of externalization was investigated in a psychoacoustic experiment performed in the same listening room. Individual binaural impulse responses were used to simulate a distant sound source delivered via headphones. The ILDs were altered using a gammatone filterbank for analysis and resynthesis, where the envelopes of the left and right-ear signals were modified such that the naturally occurring fluctuations of the ILDs were restricted. This manipulation reduced the perceived degree of externalization. This was consistent with the analysis of short-term ILDs at different distances showing that a decreased distance to the sound source also reduced the ILD fluctuations.

This chapter is based on Catic *et al.* (2013)

2.1 Introduction

Natural sound sources in our environment are perceived as externalized, i.e., located out in space with a clear perception of distance. In contrast, sound signals played via headphones are typically perceived as internalized, i.e., located inside the listener's head. It is widely accepted that the acoustic filtering produced by the head, torso, and pinna, as described by the head-related transfer function (HRTF), along with the filtering due to the acoustic environment, are responsible for externalization (Blauert, 1997). However, this particular filtering of the signals reaching the two ears gives rise to a variety of spatial auditory cues, and it is not yet understood which of these various cues are crucial for externalization.

Hartmann & Wittenberg (1996) examined the importance of binaural cues, such as interaural level differences (ILDs) and interaural time differences (ITDs), for externalization. They used binaural synthesis to simulate a distant lateral sound source via headphones in an anechoic room, which offered control over the presented stimulus such that interaural cues could be systematically modified from their baseline (original) values. More specifically, this was done by independently adjusting the amplitudes and phases of the harmonics of a synthesized vowel, such that the influence of ILDs was assessed while the ITDs were kept unchanged, or vice versa. Hartmann & Wittenberg (1996) found that ITDs were important only below 1.5 kHz, whereas ILDs were important at all frequencies without weighting of specific frequency regions. Furthermore, it was shown that the frequency dependence of ITDs due to the dispersion around the head did not influence externalization, whereas presenting correct spectral information in each ear was necessary.

Kulkarni & Colburn (1998) found that the details of the magnitude spectrum of the HRTF were not crucial either for externalization, since smoothing the magnitude of the HRTF only had an effect on the elevation of the sound source. Their findings indicated that the sharp spectral peaks and notches produced by the pinna at very high frequencies did not play a major role for externalization.

The above studies were conducted in anechoic rooms. Since the presence of reflections is known to influence spatial perception, particularly in connection with distance perception (Zahorik *et al.*, 2005) and externalization (Begault *et al.*, 2001), the obtained results may not easily be generalized to echoic conditions. Reverberation substantially improves the accuracy of distance perception compared to the performance in anechoic rooms where sound intensity is the only available cue. For instance, listeners either strongly underestimate the distance to a sound source in anechoic rooms, or, in the case of an unfamiliar sound source, the distance judgments converge to a specific distance value irrespective of the actual sound source distance (Coleman, 1962; Mershon & King, 1975). Given this role of

reverberation on accurate distance perception, the ratio of direct to reverberant sound (D/R) has often been considered as a primary source of information for distance perception, because variations in D/R result in multiple cues to which the auditory system may be sensitive, such as changes in the length of reverberant tails, in interaural coherence, or in the spectral envelope of the sound (*e.g.* Zahorik *et al.* 2005).

A further reason to consider echoic settings is that, when listeners are placed in a reverberant environment, they may completely change the weight they assign to different spatial cues from those used in anechoic settings. Brungart & Rabinowitz (1999) showed that distance judgments for nearby sources (less than 1 m.) in anechoic settings are dominated by low frequency ILDs. A similar study investigating distance perception of nearby sources in reverberant settings (Kopčo & Shinn-Cunningham, 2011) suggested that the role of low- frequency ILDs is diminished in echoic surroundings, and listeners instead rely heavily on cues provided by the direct-to-reverberant ratio (D/R).

While there is no doubt that D/R-related cues play a critical role in distance perception, having access to these cues alone does not necessarily result in externalized sound images. Ohl *et al.* (2010) used a stereo recording of a lateral sound source in a reverberant room to create an internalized signal presented via headphones. In that setting, two microphones with interspacing approximately corresponding to the distance between the human ears were used to capture room impulse responses at the position where the listener was seated in the following headphone playback. Such a signal contained all the room information (hence the D/R-related cues) and a frequency-independent ITD. However, the natural ILDs and the frequency dependent ITDs were missing, as would be captured by binaural room impulse responses (BRIRs) acquired on a human head. Such a signal was internalized, strongly suggesting that the head-related ILDs are involved in externalization not only in anechoic rooms, as shown by Hartmann & Wittenberg (1996), but also in reverberant settings. While it might be possible for listeners to judge the distance of internalized signals based on intensity or D/R-related cues, head-related binaural cues seem necessary to provide externalized sound images.

Although spatial cues are well defined in anechoic conditions, this is not the case in reverberant environments, since they are affected by the superposition of the reflected sound with the direct sound. In such acoustic settings, the binaural cues available to the listener are a combination of head-related and room-related cues. Therefore, it is of interest to characterize such complex combined cues in order to better understand their impact on spatial perception. Shinn-Cunningham *et al.* (2005) investigated the effects of reverberation on spatial cues by analysing binaural room transfer functions (BRTFs) acquired on an acoustic manikin at different positions in

a typical classroom. The frequency fluctuations in the magnitude spectra of the BRTFs increased with decreasing D/R while the ILD magnitude decreased at all frequencies. Accordingly, the binaural cues were rendered less reliable compared to anechoic space, and this effect has since been attributed to the reduced accuracy of directional localization in reverberant environments (Ihlefeld & Shinn-Cunningham, 2011). On the other hand, since reverberation improves distance perception and binaural cues are involved in externalization, it might be that the combined binaural cues that arise from the interaction of the head-related binaural cues with reverberant energy play a role in externalization.

The aim of the present study was to characterize the statistical distributions of such dynamic (short-term) binaural cues as a function of source distance and frequency, and to investigate how the shaping of these distributions affects externalization. First, the behaviour of short-term ILDs was analysed by extracting the ILD distributions in critical bands for speech recorded on a head and torso simulator (HATS) at different distances in a standard listening room. Then, the effect of ILDs on the perceived externalization of speech sources in reverberant settings was investigated. For that purpose, the ILD distributions of a distant sound source were modified by altering the short-term ILDs via the signal envelopes in the left and right ears, using a frequency analysis and resynthesis based on a gammatone filterbank. A subjective listening test was performed in which listeners rated the degree of externalization for sound sources with natural and modified ILD distributions. In addition, the effect of source spectral content on externalization was considered, as well as the extent to which ILDs played a role in different frequency regions.

2.2 Methods

2.2.1 Measurement of individual BRIRs

Individual BRIRs were measured in seven listeners in order to simulate a distant sound source delivered *via* headphones. A standard IEC 268-13 listening room located at the Technical University of Denmark was used as the test environment. The room had a reverberation time T_{30} of approximately 500 ms, corresponding to a typical living room environment. The dimensions of the room were 7.52×4.74×2.76 m (L×W×H). The listener was seated on a wooden chair at a distance of 2.1 m from the left side wall and 2.7 m from the front wall. The listener was facing the front wall and was instructed to keep the head still. The sound source, a Dynaudio BM6 loudspeaker, was positioned in the horizontal plane at 30 degrees azimuth relative to the listener. The distance of the listener's head to the loudspeaker was 1.7 m. Miniature microphones (Sonion type 8002) were inserted into the ear canals of the

listener. The ears were blocked with foam ear rings in which rubber holders were inserted such that the microphones were kept in place inside the foam. The microphones had a custom built power supply and their outputs were amplified by an RME QuadMic preamplifier before being fed to an RME DIGI96/8 PAD soundcard. The excitation signal used for capturing the response was a 5-second logarithmic sweep (Muller & Massarani, 2001) with 10 repetitions. The sampling frequency was 44.1 kHz. The captured response was transformed into the frequency domain where the inverse sweep spectrum was applied. The resulting BRIR was then windowed with a $\cos^2$ window at 500 ms, thereby removing soft nonlinear components that appeared at negative times of the BRIR.

In order to compensate for the effect of the headphones, individual headphone impulse responses (HPIRs) were also measured for the same microphone position. The sweep duration was in this case 2 seconds and the response was averaged 10 times to avoid the occurrence of extreme spectral dips. $\cos^2$ windows with transition times of 0.5 ms were applied before the resulting impulse-response onset and at its end, such that the total duration was 5.5 ms. The HPIRs were then transformed into the frequency domain by a 2048-point fast Fourier transform (FFT) and inverted using frequency dependent regularization (Kirkeby *et al.*, 1998) in order to avoid boosting of very low and very high frequencies where the HPIRs had low energy. Thus, the regularization parameter had very small values in the band of interest where full compensation was required, and very high values below 50 Hz and above 18 kHz. Since the headphone system was mixed-phase, the inversion of HPIRs resulted in an acausal response, and a delay was imposed on the inverse filter (a cyclic shift) such that the resulting equalization filter was causal.

Finally, the BRIRs, HPIRs, and speech material from the "conversational language evaluation (CLUE)" test (Nielsen & Dau, 2009) were convolved via the frequency domain using a 65536-point FFT. In order to verify that the obtained virtual sound image played *via* headphones was perceived as fully externalized and coincident in location with the real source, the listeners were first presented with the headphone playback followed by the loudspeaker playback, and were asked to judge whether the sounds came from exactly the same position. All listeners perceived the two sounds as externalized and located exactly at the loudspeaker.

2.2.2 Measurement and analysis of ILDs

In order to analyse the behaviour of short-term ILDs as a function of distance, BRIRs were also measured in the same IEC 268-13 room on a Bruel & Kjær HATS. The measurements were carried out at distances of 0.3, 0.5, 1, 2, and 3 m and an azimuth of 30 degrees relative to the HATS, which was placed 2.1 m from the side walls and 2.2 m from the back wall. The BRIRs were captured using miniature

microphones (Sonion type 8002) inside the HATS ear canal, as described in Section 2.2.1. They were then convolved with speech material from the CLUE test (Nielsen & Dau, 2009). These binaural speech signals were then analysed by a bank of logarithmically-spaced gammatone bandpass filters (Patterson & Moore, 1986; Hohmann, 2002) with a bandwidth of 1 equivalent rectangular bandwidth (ERB); (Glasberg & Moore, 1990). The Hilbert envelopes of each filter output in the left and right ear signals were used to calculate the ILD at each time instant. These short-term ILDs were lowpass filtered at 500 Hz and collected from 1 second of speech to obtain ILD distributions in the form of histograms with 1-dB resolution.

2.2.3 Modification of ILDs

To investigate how ILD statistics affect perception of externalization, a method was required to systematically manipulate ILDs in different frequency regions. The procedure used in the present study, illustrated in Figure 1, is based on frequency analysis and resynthesis using a complex gammatone filterbank (Hohmann, 2002). This filterbank consists of a series of logarithmically-spaced all-pole 4th order filters that are commonly used to model auditory filters in the inner ear. It delivers a complex output $\underline{w_i}(n)$ for each frequency channel i, whose magnitude is an approximation of the Hilbert envelope. The resynthesis is then performed by summing the real part of all filter outputs, after a complex gain $g_{E,i}$ and a delay τ_i are applied to each frequency channel to ensure that the summed output $y(n)$ equals the input signal $w(n)$ when no further processing is applied.

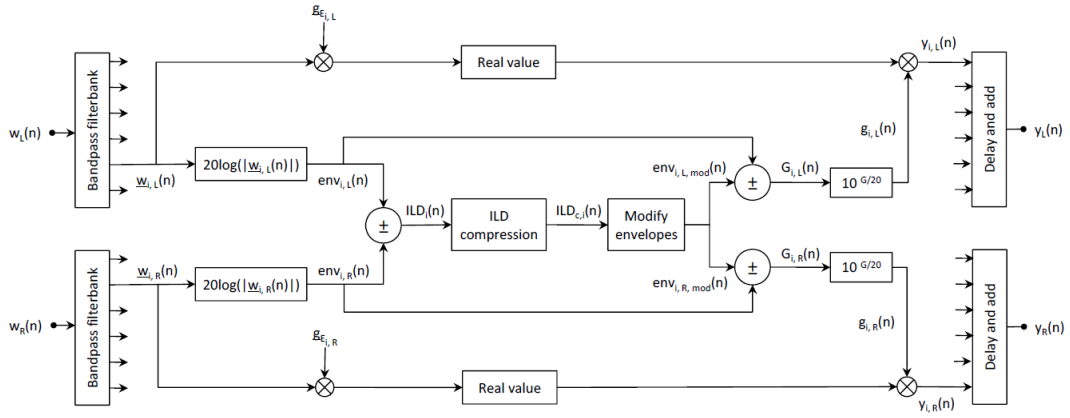

Figure 1 ILD-modification method using a gammatone filterbank with binaural input and output. In each frequency channel, the log-transformed envelopes of the left and right complex filter outputs are subtracted to estimate the instantaneous input ILD. This ILD is compressed and the average input envelope is modified to obtain left and right envelopes with the compressed instantaneous ILD. A channel-specific gain and delay are then applied to the real part of each auditory-filter output. These outputs are finally summed to resynthesize the ILD-compressed signal.

Here, a binaural input was used for the gammatone filterbank, and the ILD modification was integrated into the frequency analysis and resynthesis. In order to

minimize overlap of the frequency channels in the resynthesis, the gammatone filters had a reduced bandwidth of 0.5 ERB and only every third channel was used in the reconstruction. It should be noted that this differs from the filterbank parameters used for the analysis of ILD distributions (Section 2.2.2), where the full spectrum and a 1-ERB bandwidth were used.

In every frequency channel *i*, the left and right envelopes were first converted into the logarithmic domain, and the instantaneous input ILDs, $ILD_i(n)$, were calculated by subtracting the left and right envelopes. The mean value of the ILDs, $\overline{ILD_i(n)}$, was then subtracted, and the input ILDs compressed as follows:

$$ILD_{c,i}(n) = \alpha\left(ILD_i(n) - \overline{ILD_i(n)}\right) \quad \textbf{Eq. (1)}$$

where α is the compression parameter. An average envelope, $env_{i,avg}(n)$, was derived from the left and right envelopes:

$$env_{i,avg}(n) = \frac{env_{i,L}(n) + env_{i,R}(n)}{2} \quad \textbf{Eq. (2)}$$

The modified left and right envelopes $env_{i,L,mod}(n)$ and $env_{i,R,mod}(n)$ were then obtained by adding the compressed ILD, $ILD_{c,i}(n)$, and the mean ILD value to the average envelope:

$$env_{i,L,mod}(n) = env_{i,avg}(n) + \frac{ILD_{c,i}(n) + \overline{ILD_i(n)}}{2}$$
$$env_{i,R,mod}(n) = env_{i,avg}(n) + \frac{ILD_{c,i}(n) + \overline{ILD_i(n)}}{2} \quad \textbf{Eq. (3)}$$

Using these modified envelopes, a gain g_i was calculated and applied to the real part of the filter outputs to obtain the compressed ILD signal in channel *i*, *e.g.*, for the left ear:

$$y_{i,L}(n) = g_{i,L}(n)Re\left(g_{E,i,L} \times \underline{w}_{i,L}(n)\right) \quad \textbf{Eq. (4)}$$

The gain was lowpass-filtered at 500 Hz with a zero-phase filter (with effective order 4) before it was applied. This was done in order to reduce audible artefacts associated with instantaneous compression, and because it is unlikely that the auditory system can detect faster fluctuations in ILDs. Subsequently, all the channels were delayed and summed to form the left and right output signals with compressed ILDs.

This processing resulted in a quasi-linear mapping (on a dB scale) between the input and the output ILDs, with a degree of compression specified by the parameter α in Eq. (1). Due to the low-pass filtering of $g_i(n)$ and the following summation of the channels in the reconstruction of the signals, the mapping of input and output ILDs was not strictly linear and the final compression of the ILDs was smaller than that specified by α. With this method, the compression of the ILD fluctuations was symmetrical around the distribution mean, and thus the mean was not modified by the method. Importantly, this method also allowed an alteration of the ILDs while preserving the original ITDs.

Figure 2 shows an example histogram of ILD distributions in one frequency channel centred at 2.1 kHz for an unprocessed signal ("original", solid line) and the corresponding ILD-compressed signal with α = 0 ("processed", dashed line), for a single sentence spoken at 1.7-m distance and 30 degrees azimuth. The occurrence of each ILD in 1-dB bins is plotted relative to the total number of samples, for distributions obtained from 1-ERB-wide filters after resynthesis. It can be seen that the ILDs fluctuate within a smaller range in the compressed signal compared to the original signal, while the processing has little effect on the mean of the ILD distribution.

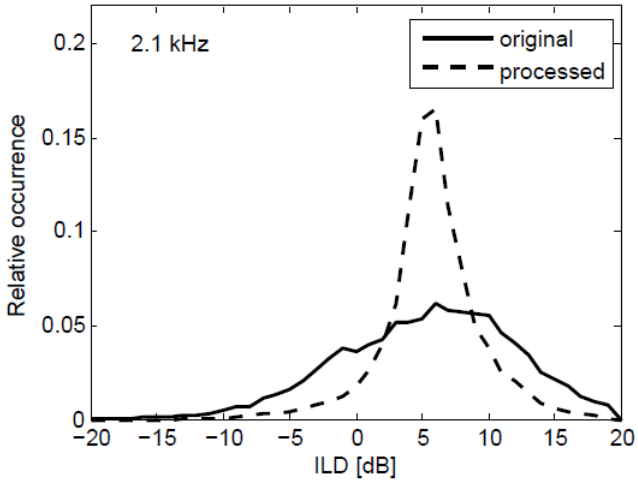

Figure 2 ILD distributions of a BRIR-processed sentence at 1.7 m distance and 30 degrees azimuth ("original", solid line) and of the same sentence after ILD-compression with α = 0 ("processed", dashed line), for a single frequency channel centred at 2.1 kHz. The histograms show the occurrence of each ILD in 1-dB bins relative to the total number of samples. The distributions were obtained from 1-ERB-wide filters after resynthesis.

2.2.4 Listening test procedure

Seven normal-hearing listeners participated in the experiment, which was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-KA-04149-g). Individual BRIRs were first obtained in all listeners as described in Section 2.2.1, and the fully-externalized BRIR-processed speech source,

consisting of a single sentence of 1.34-s duration, was used as a reference signal. This signal was then processed by the method described in Section 2.2.3. Three values of the compression parameter α were used: α = 1, corresponding to natural ILD fluctuations, α = 0.5, corresponding to medium compression, and α = 0, which is the maximum compression that can be achieved with the present method. Nine different bandwidth conditions were considered: one broadband condition, in which the speech had a bandwidth of 14 kHz, 4 lowpass conditions (cutoff frequencies: 0.5, 1, 2, and 4 kHz), and 4 highpass conditions (cutoff frequencies: 4, 2, 1, and 0.5 kHz). The lowpass and highpass filtering were obtained by discarding the frequency channels up to (or down to) the channel whose centre frequency was the closest to the required cutoff frequency. The broadband reference was presented at a level of 63 dB SPL (Sound Pressure Level) via Sennheiser HD580 calibrated headphones. The presentation level was not modified after removal of specific channels in the lowpass and highpass conditions, such that the SPL in a given auditory filter of the listeners remained constant across conditions, while the overall SPL varied.

Since the aim was to measure how externalized a sound was perceived, a subjective rating scale similar to that used by Hartmann & Wittenberg (1996) was introduced. The scale consisted of 4 possible externalization ratings by the listener: (0) the sound is in my head; (1) the sound is closer to me; (2) the sound is closer to the loudspeaker; and (3) the sound is at the position of the loudspeaker. The listeners were instructed to ignore other attributes of the sound such as frequency content or audible level differences due to ILD processing, and to only focus on the perceived amount of externalization. The listening setup was the same as that described in Section 2.2.1. A small touch screen was placed in front of the listeners such that they could easily rate externalization without moving their head. The four possible ratings were indicated on a graphical user interface via buttons with the corresponding number and text description. Although the visual cues provided by the loudspeaker in this setup allowed the listener to identify the fully externalized source, they were not strong visual speech cues, such as a talking face, that could introduce a response bias by pulling the auditory image towards the visual source. However, the fact that the listeners were sitting in the same location and environment where the BRIRs were recorded may have contributed to a convincing percept of externalization.

In each trial, the broadband reference with α = 1 was always presented first, followed after a 400-ms silent gap by one of the 27 processed signals that the listeners were asked to rate. Each listener performed 10 runs, in which each of the 27 combinations (3 α values × 9 bandwidths) was presented once in a random order. In addition, one training run was completed for which the data were discarded. The final externalization rating (E-rating) for each condition was defined as the mean of the listener's 10 ratings.

In order to test whether externalization was possible based on monaural cues only, the speech was also presented to the listeners using the input from one ear only in one additional condition with $\alpha = 1$. A level of 40 dB SPL was used for this monaural presentation, such that audible sound did not cross to the other ear. This additional condition was presented separately from the above 27 binaural conditions.

2.3 Results

2.3.1 Analysis of ILD distributions

Figure 3 shows the measured ILD distributions for the sound source at 30 degrees azimuth in the form of histograms for distances of 50 cm (solid lines), 100 cm (dashed lines), and 200 cm (dotted lines). The distributions are shown for a low-frequency channel at 0.3 kHz (top panel), a mid-frequency channel at 2.4 kHz (middle panel), and a high-frequency channel at 4.6 kHz (bottom panel). The resolution of the histograms was chosen to be 1 dB. For the nearby source at 50 cm, where the direct sound dominates, the distributions were narrow in all three frequency channels, and the ILDs fluctuated mostly around +/- 3 dB around their mean value. As the distance to the sound source increased, the ILDs were less likely to take on values close to their mean and became distributed over a larger ILD range. This is due to the decrease in direct sound compared to the amount of reflected sound energy as the source is moved further in distance. However, this increase in ILD fluctuations was more pronounced at high frequencies (middle and bottom panels). At low frequencies (top panel), the ILD distributions were narrower compared to other frequency regions for the same distance and, although they still did broaden with increasing distance, the effect was smaller than in higher frequency channels. At very low frequencies below 300 Hz (not shown here), the ILD fluctuations did not show any clear dependence on distance.

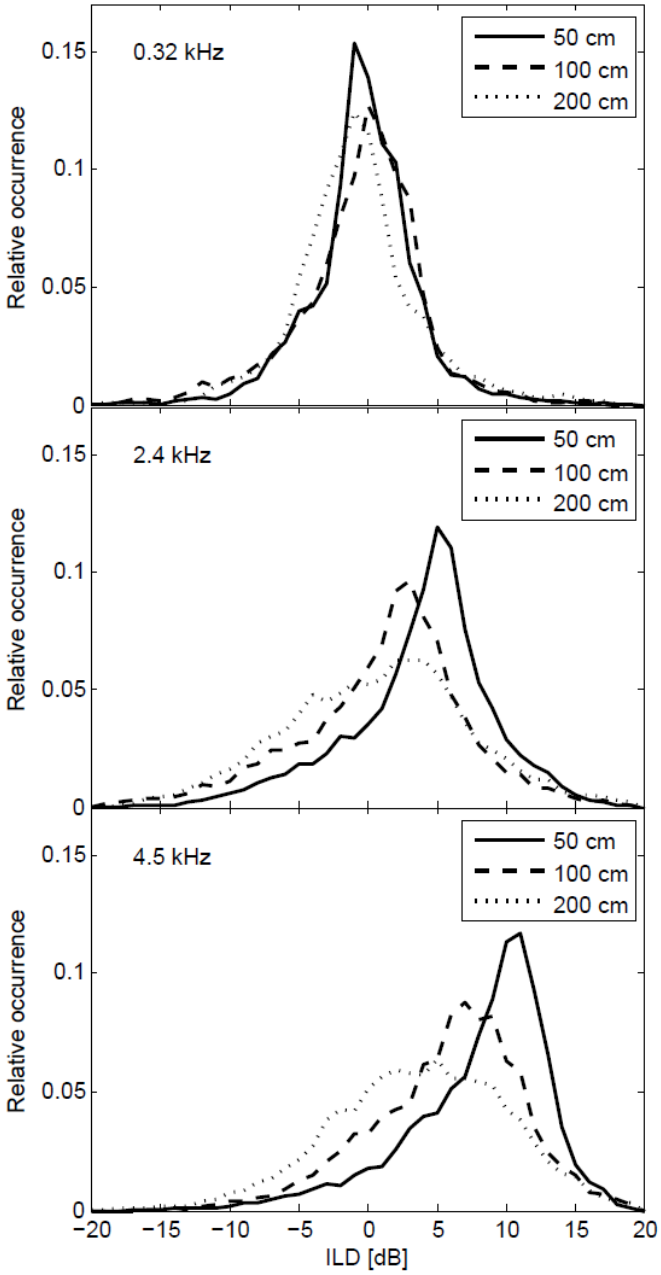

Figure 3 ILD distributions for speech that was processed with BRIRs acquired on a HATS at distances of 50 cm (solid lines), 100 cm (dashed lines) and 200 cm (dotted lines). The distributions are shown for a low-frequency channel at 0.32 kHz (top panel), a mid-frequency channel at 2.4 kHz (middle panel), and a high-frequency channel at 4.5 kHz (bottom panel). All the channels have 1-ERB bandwidth

ILDs are generally strongly frequency dependent for lateral sources due to the head shadow effect, which means that they increase at frequencies above 1 kHz. This increase was observed in the ILD distributions for the nearby source (Figure 3, solid lines), for which the mean ILD increased with frequency. For the farther sources, giving rise to a more diffuse sound field, this frequency dependence was reduced, *i.e.*, the mean of the ILDs remained towards zero, especially for the mid-frequency channel. Hence, the overall magnitude of the ILDs was reduced with distance, whereas their variation around the mean was increased.

2.3.2 Externalization perception

Figure 4 shows the across-listener average externalization ratings, referred to as E-ratings in the following, for lowpass-filtered speech (left panel) and highpass-filtered speech (right panel), as well as the broadband speech with spectral content up to 14 kHz (rightmost data points in left panel, leftmost data points in right panel). The data are shown for the three values of the compression parameter α, with unmodified ILDs (circles, α = 1), medium ILD compression (squares, α = 0.5), and full compression (triangles, α = 0).

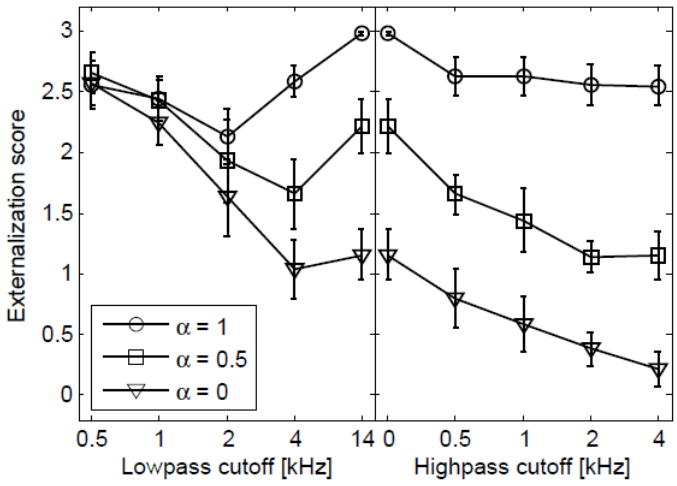

Figure 4 Across-listener mean externalization ratings for lowpass-filtered speech (left panel) and highpass-filtered speech (right panel). The broadband speech is shown in both panels (rightmost data points in left panel, leftmost data points in right panel) for a cutoff frequency of 14 kHz. The ratings are shown for speech with natural ILD fluctuations (α = 1, circles), medium compression (α = 0.5, squares) and full compression (α = 0, triangles). The error bars indicate the standard error of the mean.

For the speech conditions with altered spectral content but natural ILD fluctuations (circles, α = 1), the E-ratings were between 2.5 and 3, indicating that these sounds were always perceived very close to the loudspeaker, with the exception of the lowpass-filtered speech at 2 kHz, where the ratings showed a small dip (E-rating = 2.2) compared to the other bandwidth conditions. For broadband speech (data points common to the two panels), reduced ILD fluctuations led to less externalization, whereby the reduction of α to 0.5 resulted in slightly reduced externalization (E-rating = 2.2), and further reduction of α to 0 moved the sound image close to the listener (E-rating = 1.2).

The externalization of highpass-filtered speech (right panel) was strongly affected by the shape of ILD distributions, whereby the reduction of α to 0.5 resulted in a medium degree of externalization (E-rating ≈ 1.5), and further reduction of α to 0 resulted in the sound being perceived almost inside the head (E-rating ≈ 0.5). Moreover, the effect of reduced externalization due to compression of the ILD distributions was found to increase as the highpass cutoff frequency was increased.

In contrast, the lowpass-filtered speech (left panel) was only affected by the ILD compression if the sound contained mid to high frequencies, whereby full compression ($\alpha = 0$) resulted in E-ratings of 1.7 and 1.1 for speech lowpass-filtered at 2 and 4 kHz, respectively. For speech with lowpass cutoff frequencies of 1 and 0.5 kHz, the sound was always rated as being close to the loudspeaker (E-rating ≈ 2.5), regardless of the value of α.

A two-way ANOVA performed on the listener's E-ratings with the frequency content and the ILD-compression parameter α as factors confirmed that both factors affected externalization, with significant main effects of frequency content [$F(8,162) = 14.19$, $p<0.0001$] and α [$F(2,162) = 104.72$, $p<0.0001$]. The interaction between frequency content and α was also significant [$F(16,162)=5.17$, $p<0.0001$], reflecting the fact that α had a clear effect on E-ratings for broadband and highpass-filtered speech, but not for speech lowpass-filtered below 1 kHz.

The monaurally-presented speech (not shown) was always perceived inside the head, corresponding to an E-rating of 0. The listeners reported that the sound in this monaural condition was located "at the ear" to which it was presented. This indicates that the availability of binaural information is essential for externalization.

2.4 Discussion

2.4.1 Effects of bandwidth and monaural presentation

The current study showed that sounds with relatively narrow bandwidth could be well externalized, as the lowpass and highpass-filtered speech signals with natural ILD fluctuations ($\alpha = 1$) were perceived very close to the loudspeaker. Thus, only small deviations in externalization from the fully-externalized broadband source were observed when high or low-frequency content was removed. This suggests that the cues that contribute to externalization are effective in both high and low-frequency regions when listening to speech in reverberant environments. Moreover, the fact that the single-channel (monaural) headphone presentation did not result in externalized sound images suggests that some of these externalization cues need to be binaural. This monaural control condition might not have been optimal, because monaural signals are likely to be perceived at the ear to which they are presented due to the natural occurrence of very large ILDs only for sound sources close to the ear. However, the use of a diotic presentation mode in further informal listening tests, in which ILDs were removed by sending the left-ear signal to both ears, always led to internalized sound images, thus confirming the importance of binaural cues for externalization.

2.4.2 Role of ILD fluctuations

The hypothesis that the ILDs stemming from a combination of head-related and room-related ILDs are an essential cue to externalization was largely supported by the results, which showed that listeners were sensitive to changes in the shape of the ILD distributions, and that these have a significant impact on externalization. Furthermore, this finding was consistent with the physical ILD distributions measured in a typical room environment at different distances, whereby narrow distributions were found at close distances and broader distributions at farther distances. Externalization ratings had a close relationship with the compression parameter α for broadband and highpass-filtered speech, but also for lowpass-filtered speech as long as the cutoff frequency was above about 1 kHz.

For signals containing only low frequencies (below 1 kHz), there was no effect of α and the speech was well externalized even when full compression was applied, suggesting that ILD fluctuations do not play a dominant role for externalization at low frequencies, and that other low-frequency cues are sufficient for full externalization of lowpass-filtered sound. For signals containing high frequencies, the presence of frequency content below 500 Hz was also found to improve externalization slightly when compression was applied. However, the effect was rather small, as the sound source moved from being almost inside the head to being close to the listener. This indicates that the low-frequency cues that allow full externalization of lowpass-filtered sound do not facilitate externalization when high-frequency content is present. Therefore, ILD cues are predominant for externalization as soon as the sound contains mid to high frequencies.

The finding that the role of ILDs for externalization differs at high and low frequencies contrasts with the results of Hartmann & Wittenberg (1996), who found that ILDs appeared to be equally important in all frequency regions. This could be due to differences in the experimental settings, *i.e.*, the use of a reverberant room instead of an anechoic environment, or to differences in the stimuli and the method used for ILD modification.

2.4.3 Externalization at low frequencies

There are at least four possible reasons for why the lowpass-filtered speech remained well externalized at low frequencies despite a full compression of ILD fluctuations ($\alpha = 0$). This might be explained by a) differences in how effective the method used for ILD-fluctuation compression is in low vs high-frequency channels, b) a role of ITDs at low frequencies, c) a lowpass-filtering effect linked to changes in D/R, which is known to occur in distance perception, and/or d) an increased influence of the visual cue provided by the loudspeaker due to less reliable auditory information. These possibilities are discussed in the following.

- *Insufficient compression of ILD fluctuations at low frequencies*. It might be that either the applied processing was ineffective in low-frequency channels, or that the compression was effective but the listeners were not sensitive to it. Figure 5 shows the original ILD distributions (solid lines) and processed ILD distributions extracted after resynthesis (dashed lines) for a high-frequency (HF) channel at 3 kHz (upper panel) and a low frequency (LF) channel at 0.44 kHz (lower panel). It can be seen that the ILD compression is effective in both frequency channels, suggesting that the auditory system may just not be sensitive to changes in the shape of ILD distributions at low frequencies. However, the amount of compression relative to the original ILD distributions is different for the two channels, as the original distributions have their maxima at a relative occurrence of 0.05 and 0.08 for the HF and LF channel, respectively, while the processed ILD distributions both have their maxima at about 0.15. Moreover, the distributions as such do not show which ILD-fluctuation rates are reduced by the compression, another aspect that differs between the HF and LF channels. An inspection of the FFT spectra of the instantaneous ILD fluctuations indeed revealed that, in LF channels, the applied processing mainly affected instantaneous changes in ILD fluctuations up to a fluctuation rate of about 50 Hz, while the fluctuations above this rate remained nearly identical to those of the unmodified sound source. In contrast, the reduction of ILD fluctuations was effective for fluctuation rates up to at least 150 Hz in HF channels, and even 200-300 Hz in some cases.

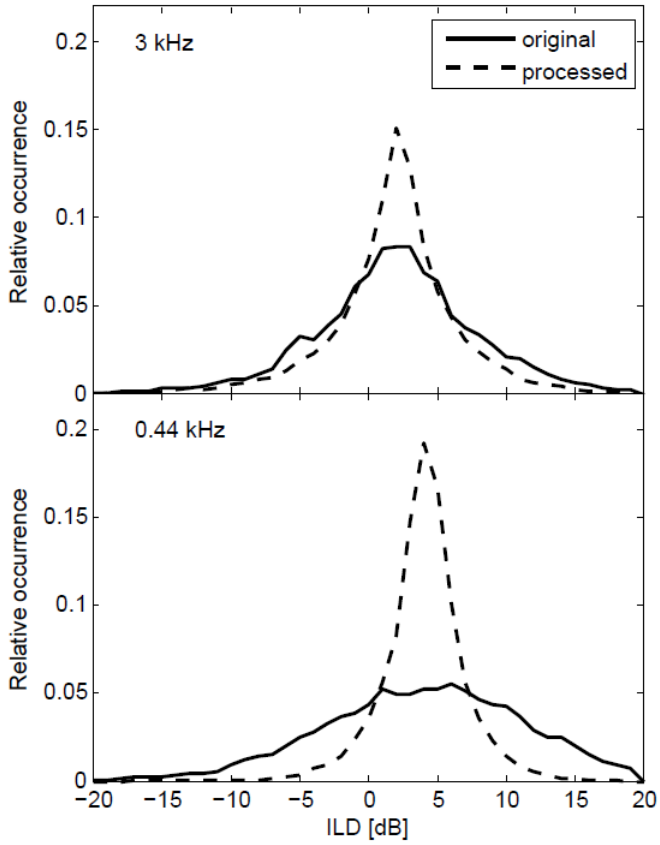

Figure 5 Difference between the achieved ILD-fluctuation compression (α = 0 , dashed lines) in a high-frequency (3 kHz, upper panel) and a low-frequency (0.44 kHz, lower panel) channel, compared to the uncompressed condition (α = 1, solid lines). The histograms show the occurrence of each ILD in 1-dB bins relative to the total number of samples. The distributions are obtained from filters with 1-ERB bandwidth after frequency resynthesis.

- *Role of ITDs.* Another cue to externalization at low frequencies may be the ITDs, as their fluctuations also increase with the presence of reverberant energy. This was confirmed by an analysis of the short-term ITD distributions obtained with the HATS BRIRs (see Section 2.2.2), which showed that the width of these distributions broaden as the level of the direct sound relative to that of the reverberation is decreased. Therefore, a compression of ITD distributions could be expected to reduce the perception of externalization. In order to verify this, an additional informal listening test was performed in four listeners with signals whose ITD distributions were compressed. The width of the ITD distributions was reduced while retaining their peak by modifying the signal phase in the low-frequency channels, with a similar method to that described in Section 2.2.3. Figure 6 shows ITD distributions, in the form of histograms, obtained by collecting short-term ITDs in speech of about 1-second duration. The distributions are shown for the original sound source at 1.7 m distance (solid line) and for the phase-modified source (dashed line). The listening tests showed that externalization was *not* affected

by such a compression of the ITD distributions, suggesting that ITD *fluctuations* do not play a significant role for externalization. However, removing the constant ITDs in addition to compressing the ITD fluctuations (Figure 6, dotted line) did result in a loss of externalization for lowpass-filtered speech, with sounds perceived inside the head. This indicates that, although ITD fluctuations may not be important for externalization, the overall static ITDs must be retained.

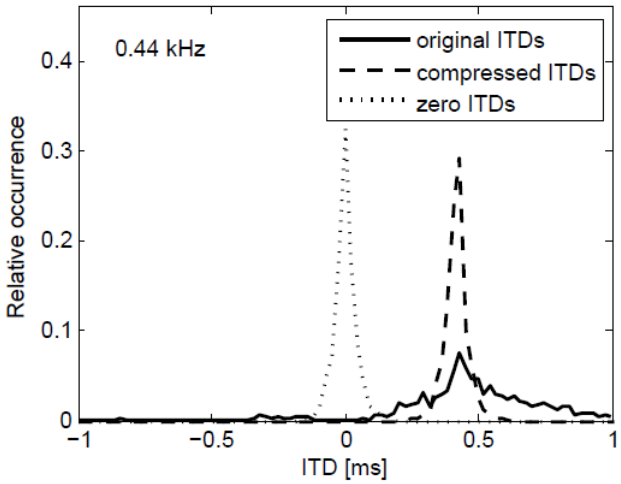

Figure 6 ITD distributions for an unmodified speech signal (solid line) and a phase-modified speech signal (dashed line) for a channel with a centre frequency of 0.44 kHz. The ITD distribution of a speech signal in which the ITDs were removed (dotted line) is also shown. The histograms show the occurrence of each ILD in 1-dB bins relative to the total number of samples.

- *Lowpass filtering effect due to changes in D/R*. Neither the ILD nor the ITD fluctuations were found to have a clear effect on externalization of lowpass-filtered speech in this study. However, another effect related to changes in D/R might have contributed to externalization at low frequencies. As the D/R changes, the spectral envelope varies due to the relatively higher absorption of reflected high-frequency energy compared to low-frequency energy. This is due to the air and materials in a typical room, and leads to an effective lowpass filtering of the sound. Because of this, lowpass-filtered sound sources are sometimes judged as being further away than broadband and highpass-filtered sources, an effect that has been reported in studies on auditory distance perception, (*e.g.* Little *et al.*, 1992; Nielsen S. H., 1993). In the present study, some listeners indeed perceived the lowpass-filtered sounds as being further away than the loudspeaker, suggesting that this distance-related effect might have influenced the results. Despite this, it is questionable to what degree distance-related effects can be transferred to externalization perception. For instance, monaural distance judgments may be

based on D/R-related or intensity cues (Ashmead *et al.*, 1990; Shinn-Cunningham *et al.*, 2000), but the present results showed that monaurally-presented sounds were not externalized. Although reverberation facilitates the fluctuations of binaural cues, the sensitivity to changes in D/R seems to be very similar for monaural and binaural input. Larsen *et al.* (2008) investigated the minimum audible difference in D/R via manipulation of BRIRs by scaling the amount of direct sound. Such a manipulation would also affect the amount of fluctuations in binaural cues, but the listeners' sensitivity to the scaling was found to be the same for the monaurally and binaurally-presented sounds. However, Larsen et al. argued that, although the listeners were able to detect changes in D/R in the monaurally presented sounds, these did not evoke natural distance percepts.

- *Increased visual-capturing effect.* In the listening experiment, the loudspeaker placed at the location of the fully-externalized reference was visible to the listeners at all times, which might have provided an additional visual cue. In cross-modal localization tasks, visual cues have been shown to bias auditory localization more strongly as the auditory information becomes less reliable (Alais & Burr, 2004). Although it is unclear whether such an interaction also applies to externalization (Zahorik, 2001), it is possible that low-pass filtering rendered the relevant auditory information sufficiently unreliable for a visual-capturing effect to occur. Such an effect might have led to full externalization in the lowpass-filtered conditions regardless of how much the ILD fluctuations were compressed. However, the reliability of available auditory cues for externalization would also be expected to decrease with high-pass filtering and with compression of the ILD fluctuations. The fact that these two types of processing clearly led to lower E-ratings indicates that either visual capturing did not occur, or that visual bias only became dominant when high-frequency information was removed.

2.4.4 Limitations and perspectives

The processing method used in the present study proved successful at reproducing externalized sound images *via* headphones in a reverberant setting, and the applied compression of short-term ILDs was sufficient to demonstrate a clear effect on perceived externalization. One of the limitations of this way of processing ILDs lies in the fact that, while a specific amount of ILD compression was desired in each narrow frequency channel and was obtained by modification of the signal envelopes, the reconstructed broadband signal presented to the listener actually contained a smaller amount of compression. Another caveat might come from the difficulty to exclude a role of envelope ITDs on externalization, as the short-term

envelope ITDs and ILDs cannot be modified without affecting one another, even though the envelope was not shifted in time by the processing method used here and only the envelope amplitude was manipulated. Moreover, the compression of ILD fluctuations might have disrupted potential cues provided by time-varying changes in the spectral envelope of the signal. Even though the steady-state HRTF was not affected by the applied processing, because the mean of the applied time-varying gain was 0 dB in all frequency channels and in both ears, it remains possible that the temporal fluctuations in the HRTF induced by such processing had an impact on externalization.

The results demonstrated that the binaural cues used for externalization were effective in different frequency regions, even when large parts of the frequency spectrum were filtered out. However, it is possible that audible spatial cues remained in other frequency regions than those defined by the specified cutoff frequencies, and that the listeners relied on information present at the edges of the filter skirts, despite lower energy. Further work is thus needed to clarify whether externalization can be obtained for sounds with very narrow bandwidths.

Another limitation of the present study is that only speech signals were considered, and it is possible that the large dynamic fluctuations of speech may have played a role in externalization. Moreover, intensity-related cues have been argued to play a larger role for distance perception when speech stimuli rather than noise stimuli are used (Zahorik, 2002), possibly due to the listeners' prior knowledge about the level of natural speech (Brungart & Scott, 2001). However, the results of informal tests indicated that stationary noise that did not contain the sharp onsets and offsets of speech could be just as well externalized as speech with the present processing method. Despite this, the low-pass and high-pass filtering used here might have introduced sufficient changes in loudness and timbre to provide useable cues to vocal effort, and thus source level, which could affect externalization. Even though a contribution of such level cues cannot be excluded, changes in the ILD-compression factor clearly affected externalization without introducing changes in overall level or timbre. Furthermore, while level cues are known to be more important in anechoic settings, (*e.g.* Gardner, 1969), loudness and distance have been shown to interact much less in reverberant settings (Zahorik & Wightman, 2001).

Also, presenting the same broadband reference in each trial might have introduced some bias, such that other qualitative differences between the reference and processed signals could have influenced the listeners' E-ratings. For example, the listeners might have been discouraged from assigning the highest rating to fully-externalized, but qualitatively different, sounds. Although E-ratings remained above 2.5 in the most extreme high-pass and low-pass conditions, it is not clear to what

extent other perceptual attributes, such as timbre and apparent source width, may have interacted with externalization.

Finally, the observed effects of ILD-fluctuation compression on externalization might be relevant for hearing-aid design, as both dynamic-range compression and noise-reduction algorithms alter the natural ILDs of sounds. Wiggins & Seeber (2012) investigated the effect of hearing-aid dynamic-range compression on externalization in normal-hearing listeners. They found that compression reduced the perceived externalization significantly, but the effect was small. In the present study, the effect of ILD-fluctuation compression was very clear for broadband speech. This difference could be due to the fact that Wiggins & Seeber (2012) used non-individualized HRTFs for spatialization, or to the fact that their listening experiments were performed in a sound-insulated booth, whereas the present playback environment was the original room where the BRIRs were recorded. Moreover, the ILD modification employed by the present method is not identical to that used in a hearing-aid compressor. A hearing-aid compressor act on the input level of the sound, which leads to the ILDs either being reduced or kept natural at a certain time instant, depending on that input level. In the present method, the compression modified the ILDs such that the fluctuations around the mean ILD were reduced, depending on the input ILD. Further work is needed to assess to what extent hearing-impaired listeners utilize ILD fluctuations for externalization in different frequency regions, and whether the enhancement of ILD fluctuations in an internalized or slightly externalized sound can improve its perceived externalization degree.

2.5 Summary and Conclusion

This study investigated the effect of ILD-fluctuation compression on the perceived externalization of sound in reverberant settings. The results of a listening test performed by normal-hearing listeners showed that compressing the ILD fluctuations of BRIR-processed speech presented *via* headphones reduced the perceived externalization substantially in all conditions in which the sound contained frequencies above 1 kHz. In contrast, when the sound was lowpass-filtered below 1 kHz, externalization remained unaffected by ILD-fluctuation compression. These subjective results were consistent with an analysis of the variation of short-term-ILD distributions with distance in the experimental listening room, obtained from objective measurements on a HATS. Overall, the present findings suggest that the ILD fluctuations resulting from the combination of head-related and room-related binaural information play an important role in the externalization of sounds containing mid- to high frequencies. A similar role of ILD fluctuations remains possible for low-frequency sounds, and might not have been observed here due to

limitations in the amount of achievable ILD compression with the chosen processing method. Alternatively, cues provided by ITDs or the D/R may contribute to externalization at low frequencies, although ITD compression was not found to affect externalization and monaurally-presented sounds were always internalized in the present study.

3 THE ROLE OF REVERBERATION-RELATED BINAURAL CUES FOR THE EXTERNALIZATION OF SOUND

Abstract

Reverberation affects many types of spatial perception and is known to improve the perception of distance and externalization. Here, externalization perception was investigated in relation to the monaural and binaural cues that occur due to reverberation. Individualized BRIRs were used to simulate externalized sound sources via headphones. The measured BRIRs were subsequently modified such that the proportion of the BRIRs that contained binaural and monaural information, respectively, was varied. It was found that monaural reverberation cues were sufficient for the externalization of a lateral sound source, whereas for a frontal source, an increased amount of binaural cues from reflections was required in order to obtain well externalized sound images. The results suggested that the level of disparity in the interaural cues of the direct sound and the reverberation was important as the interaction of the two influenced the perception of externalization. An analysis of short-term binaural cues showed that for the considered conditions, the amount of fluctuation in binaural cues corresponded well to the externalization ratings obtained in the listening tests. The results further suggested that the precedence effect is involved in the auditory processing of the dynamic binaural cues that are utilized for externalization perception. Moreover, similar effects of reverberation were found at low- and high frequencies.

This chapter represents a journal article manuscript in preparation

3.1 Introduction

In natural listening situations, sounds are typically located correctly according to their position in space, i.e. the listener has a clear perception of direction and distance. Such listening situations provide both undistorted localization cues as well as cues that arise due to reflections of the original sound from, e.g., objects and walls. It is well known that, if a sound is presented via headphones, a convincingly externalized image of the sound source can be obtained if the headphone reproduction includes the spatial cues that occur due to filtering by the head, torso, and pinna as described by the head related transfer functions (HRTFs), and the information about reflections from the environment. In contrast, if the sound being presented via headphones or other listening devices lacks these characteristics, the sound is likely to be perceived inside the head, i.e. internalized. Reflections have been shown to play an important role in distance perception (Mershon & King, 1975; Bronkhorst & Houtgast, 1999; Zahorik, 2002) and externalization (Begault *et al.*, 2001), but it has remained unclear which characteristics of the reverberant sounds are utilized in externalization perception. As reflections are involved in many aspects of sound perception, such as speech intelligibility (Nábelek & Robinette, 1978; Bradley *et al.*, 2003), binaural echo suppression (Litovsky *et al.*, 1999), apparent source width (Griesinger, 1997) and the perception of room size (Sanvad, 1999), it appears that the auditory system has various strategies for dealing with reflections, which involve both binaural and monaural processing.

Since externalization and distance perception are closely related, similar strategies for utilizing reverberation might be used by the listener. It is generally accepted that the ratio of direct to reverberant sound (D/R) is indicative of the perceived sound source distance, but it is not well understood which features of D/R are used by the auditory system. A simple approach to the perception of D/R is that it can be determined by a comparison of the acoustic energy at the time where the direct sound arrives at the ear and the acoustic energy during the reverberant decay. Such an energy ratio can be extracted for sounds that contain sharp onsets and offsets, such as clicks, where the direct sound energy and the reverberant sound energy can be separated. However, this is difficult to extract for sounds that are continuous in nature, yet it has been shown that listeners' ability to perceive changes in D/R is effectively the same for noise with gradual onsets and offsets as for clicks (Zahorik, 2002). Cues that co-vary with D/R and do not rely on the above described acoustic energy comparisons are also available. The reflections that arrive after the direct sound are filtered by walls and objects in the room, and this type of filtering usually attenuates high frequencies more than low frequencies, since most materials absorb more high- than low frequency energy, leading to an effective lowpass

filtering of the sound signal. Therefore, the spectral shape of the sound is changed as more of the lowpass filtered reverberant energy is added compared to the direct sound. Furthermore, adding reverberation also causes an increase in the frequency-to-frequency variations in the signal power spectrum that occurs as a result of interference of reflected sound waves. It has been shown that such cues can be utilized in D/R discrimination (Larsen *et al.*, 2008). In Larsen *et al.* (2008), the listeners' ability to discriminate changes in D/R was investigated using a paradigm where listeners were asked to indicate which sound was more reverberant in a forced choice task. They demonstrated that D/R can be discriminated based on the above mentioned spectral cues, which are purely monaural. Furthermore, Larsen *et al.* (2008) did not find any difference between monaural and binaural listening mode when assessing D/R discrimination. This indicated that such monaural cues are utilized by the auditory system at least in terms of perceiving how reverberant a sound is.

Although externalization perception is different from the perception of the amount of reverberation in a sound, the results may suggest that monaural reverberant cues are efficient and sufficient for externalization as well. However, Catic *et al.* (2013) showed that binaural cues in the form of interaural level differences (ILDs) were essential for externalization. These time-varying ILDs occurred due to the interaction between the HRTF related cues and reverberation. Their impact on externalization was investigated in Catic *et al.* (2013) by considering the statistical distributions of short-term ILDs collected over a long duration signal, i.e. extending an entire speech sentence. Because of the long-term analysis of the speech signal, there was no distinction between how direct sound, early reflections or late reverberation affected the ILD fluctuations and the perception of externalization. Given the considerations above, it might be sufficient to access binaural HRTFs only in the direct sound and use monaural cues for the reverberant part. Furthermore, in Catic *et al.* (2013), modifications of the statistics of the binaural cues affected the externalization of broadband as well as highpass filtered speech, but not that of lowpass filtered speech (with $\leq$1 kHz cutoff), which was well externalized despite the modifications. Thus, it might be that monaural cues are more crucial for externalization at low frequencies.

The aim of the present study was to clarify how different parts of the BRIR (binaural room impulse response) affect the degree of externalization, to investigate whether monaural cues from reverberation are sufficient for externalization when binaural HRTF cues are available in the direct sound, and to assess whether externalization at low frequencies also depends on dynamic binaural cues in a similar way as was shown for broadband and highpass filtered speech in Catic *et al.* (2013). Here, these effects were considered by modifications of measured BRIRs. Listening

tests were conducted where the listeners rated the degree of externalization for broadband, lowpass- and highpass filtered speech created using different types of modified BRIRs. Subsequently, the dynamic binaural cues that occurred in the speech signals were analysed and compared to the externalization ratings obtained in the listening tests.

3.2 Methods

3.2.1 Measurements of individual BRIRs

Individual BRIRs and headphone responses were measured in six listeners in order to simulate a distant sound source delivered via headphones. The measurement method, equipment, and test environment were the same as in Catic *et al.* (2013). The loudspeakers were placed at 0° and 30° azimuth, respectively, and the distance of the listeners' head to the loudspeaker was 1.7 m. The measurements were carried out using miniature microphones inside the listeners' ears. Subsequently, the listeners were asked to judge the location and the degree of externalization of the sound sources that were synthesized based on the measurements and delivered via headphones. The headphone-delivered sound sources were perceived as coming from the loudspeakers.

3.2.2 Modification of the measured impulse responses

A room impulse response can roughly be divided into the direct sound, the early reflections and the late reverberation. The duration of anechoic head related impulse responses has been shown to be about 2.5 ms (Møller *et al.*, 1995), and therefore the direct sound can be considered to be contained in the first 2.5 ms of the BRIR in a given room. The early reflections appear after the direct sound and up to about 50 to 80 ms, depending on the room, and reverberant energy arriving after this time is considered as the late part of the BRIR.

In order to examine the effect of BRIR duration on externalization, the measured BRIRs were truncated to the following durations: 2.5, 5, 10, 20, 30, 50, 80, 100, 200, and 300 ms. The full reference BRIR had a duration of 500 ms. Figure 1 shows an example of the truncation at 2.5 ms. A cosine squared falling ramp of duration 0.5 ms was applied to the BRIR at 2.5 ms and the resulting impulse response was zero padded to a length of 500 ms. The same ramp was also applied to the BRIRs with the other truncation times, which were also zero padded to a length of 500 ms . The condition considering the various truncation times is denoted as "TR" in the following. In order to examine the degree of externalization for a sound source that contains the correct binaural cues in the direct sound but does not contain any binaural information in the reverberant sound, the impulse response was windowed as indicated in the right panel of Figure 1. Here, the direct sound was concatenated

with the monaural reverberant part, where the transition between the two was made by applying a cosine squared falling ramp to the direct sound and by applying a cosine squared rising ramp at the point where the monaural reverberant part starts. The duration of the ramps was 0.5 ms.

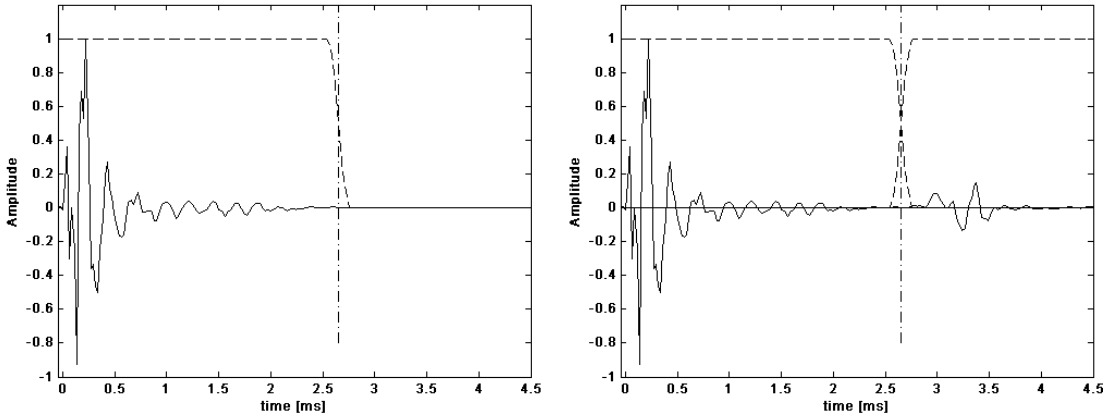

Figure 1: Truncation of BRIRs by windowing (left panel) and the concatenation of a binaural and monaural head related impulse response (right panel). The example is shown for the BRIR duration of 2.5 ms. Although the part of the impulse response below 2.5 ms is binaural, only one (left) channel is shown for clarity.

The monaural reverberant part consisted of identical reverberation in the left and right ear channels, in the following denoted as condition "MO" which was recorded in the right ear of the respective listener. Subsequently, the duration of the binaural part of the BRIR was gradually increased in order to assess the addition of binaural reverberant cues on externalization. The intersection points between the first and second part of the impulse responses matched the truncation times in the TR condition, as can be seen in Figure 1.

3.2.3 Listeners

Six normal-hearing listeners participated in the experiment concerning BRIR duration (condition TR), and five of those six listeners completed the experiment regarding identical reverberation in both ears (condition MO).

3.2.4 Stimuli

The modified BRIRs were convolved via the frequency domain with speech sentences from the "conversational language evaluation (CLUE)" test (Nielsen & Dau, 2009). In addition to the broadband speech condition, a lowpass and a highpass condition were considered, where the speech was filtered with a 4th order Butterworth filter at 1 kHz and 2 kHz respectively. The sounds were presented at a level of 62 dB SPL via Sennheiser HD580 calibrated headphones. Prior to stimulus presentation, individual headphone equalization was applied. The reference sound

was a broadband speech source convolved with an unmodified BRIR of 500 ms duration. All the speech sounds were highpass filtered at 150 Hz.

3.2.5 Procedure

A subjective rating scale as in Catic *et al.* (2013) was used with 4 possible externalization ratings: (0) "the sound is in my head"; (1) "the sound is closer to me"; (2) "the sound is closer to the loudspeaker"; and (3) "the sound is at the position of the loudspeaker". The listeners were instructed to ignore sound attributes that were not directly related to externalization (such as the unnatural frequency content in the lowpass and highpass filtered conditions), and to only focus on whether the sound was experienced as inside or outside the head and at which position within the given rating scale. The listening setup was the same as in Catic *et al.* (2013), but an additional sound source at 0° azimuth was considered in the present study. In total, four different tests were completed by the listeners, i.e. conditions TR and MO for the frontal and the lateral sound source, respectively. With the specified BRIR durations and three bandwidths, this resulted in 33 different sounds to be rated in condition TR. In the case of condition MO, a diotic control sound was also included, resulting in 36 sounds to be rated overall. In each trial, the reference sound was presented first, followed by a 400 ms silent gap and one of the modified sounds that the listeners were asked to rate. The sounds to be rated were presented in random order. The externalization rating for each presentation was taken as the average of 5 ratings.

3.2.6 Analysis of binaural cues

Binaural cues in reverberant environments do not only convey information about the direction of the active sound source, but also information about the room through the characteristics of the reflections. Due to the superposition of the direct and reverberant sound, the binaural cues become dynamic and their fluctuations depend on both the level and the direction of the reflections relative to the direct sound. As an example, if a reflection has a different ILD or ITD than that of the direct sound, variations of the binaural level cues occur (*e.g.* Catic *et al.*, 2013).

Here, the short-term interaural coherence (IC), was used as a potential indicator of externalization at low frequencies, together with the analysis of dynamic ILDs as in Catic *et al.* (2013), which are utilized for externalization mainly at high frequencies. Figure 2 illustrates the steps involved in the calculation and analysis of the binaural cues of a BRIR filtered speech source.

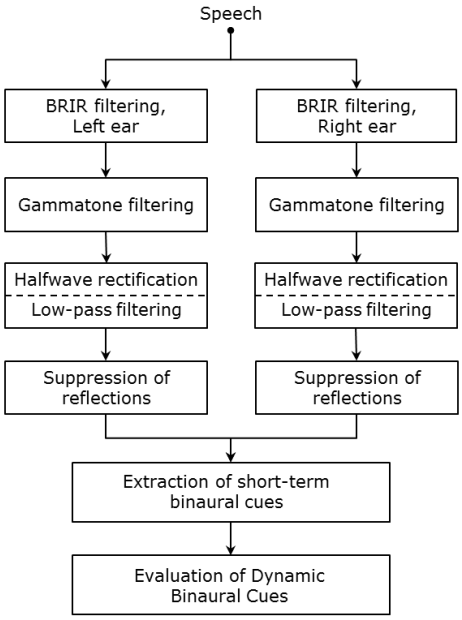

Figure 2: The steps involved in the calculation of the statistical ILD and IC measures.

3.2.6.1 Simulation of auditory periphery

First, the frequency analysis of the basilar membrane was simulated by passing the left and right ear signals through a Gammatone filterbank. These bandpass signals were then half-wave rectified and a second-order lowpass filter with cutoff frequency 1 kHz was applied.

3.2.6.2 Reflection suppression

The interaural fluctuations in Catic *et al.* (2013) were considered on the basis of long-term average binaural cues. However, the auditory system is known to suppress the localization information of reflections in a certain time window after the direct sound, (*e.g.* Litovsky *et al.*, 1999), often referred to as the precedence effect. The precedence effect may already be effective at a peripheral level of auditory processing (Hartung & Trahiotis, 2001) and when only monaural spectral shape cues are utilized (Litovsky *et al.*, 1997). Hence, such processing could happen before the computation of dynamic binaural cues and was assumed here to contribute to their calculation. Specifically, following Dizon & Colburn (2006), it was assumed that the auditory system does not only suppress the localization cues of reflections that arrive after the direct sound has diminished, but also the cues that persist during the direct sound in an ongoing manner. This type of processing can be expected to affect the dynamic binaural cues in continuous signals such as speech. Here it was assumed that the reflections up to 10-15 ms are suppressed and that the suppression is happening before the dynamic binaural cue calculation takes place. Since it is not known how the localization information suppression is achieved in sounds where the

direct component and the reverberation are present at the same time, in the present study the BRIRs were modified before the convolution with the speech source, such that the effect of reflections in the first 10 ms was diminished. This was done by multiplying the BRIR with a function that is zero in the first 10 ms after the direct sound, and applying a ramp that ranges from zero to one in the time interval from 10 ms to 15 ms. The remaining part of the BRIR was left unmodified.

3.2.6.3 Extraction of short-term binaural cues

The short-term interaural coherence was calculated from the normalized cross-correlation function. A running normalized cross-correlation was computed where the lags in the range [-1, 1] ms were considered. An exponentially decaying window with time constant τ was applied in the analysis, and the interaural coherence was taken as the maximum of the instantaneous cross-correlation function. The ILD was computed as the power difference at the two ears at the time instant where the instantaneous normalized cross-correlation was at its maximum. For the ILDs, an exponentially decaying window with time constant τ was also applied in the analysis.

The time constant τ was chosen to be 10 ms. A short time constant was chosen here as the utilization of dynamic cues for externalization was assumed to depend on the detection of rapid variations in the signal. Goupell & Hartmann (2007) found time constants on the order of 0-10 ms to be appropriate for detection of IPD and ILD fluctuations in noise, while large time constants were not applicable for that type of binaural detection.

3.2.6.4 Evaluation of the dynamic binaural cues

It was assumed here that it is the size of the fluctuation of the binaural parameters that dominates externalization. Therefore, the evaluation of the binaural cues was based on a measure that captures the span of ILD and IC in the sound sequence. This can be done by forming a histogram based on the short-term binaural cues collected over a specific sound sequence. However, since a single parameter would be preferred in order to easily relate externalization in each experimental condition to the corresponding dynamic binaural cues, the standard deviation was used here as an indicator of the size of the fluctuations in ILDs. In the case of IC, the 80th and 20th percentiles were used as measures as they both indicate the amount of variation in short-term IC and the absolute values of IC at times when IC is low and high, respectively. The absolute values were considered since the listeners' sensitivity to changes in IC is significantly higher for high IC values (close to unity) compared to lower IC values.

For both the ILD and IC measures, the short-term values were collected over the entire speech sentence, which was about 1 second in duration. This does not imply

that such a long time is required to evaluate externalization based on dynamic cues, as the changes in the short-time interaural parameters happens at a much higher rate.

3.3 Results

3.3.1 Experimental data

3.3.1.1 Truncated BRIR condition (TR)

Figure 3 shows the across-listener average externalization ratings, referred to as E-ratings in the following, for the truncated BRIR condition (TR). The data are shown for broadband speech located at 0° (solid line, squares) and 30° (dashed line, triangles). The E-ratings are strongly affected by the window duration up to about 80 ms and display a similar shape for the frontal (0° azimuth) and the lateral (30° azimuth) sound source. For short windows of 2.5 to 20 ms, the E-ratings were below 1, corresponding to a sound source being almost fully internalized or very close to the head. Increasing the window duration from 20 ms to 80 ms resulted in a substantial increase in externalization, where the perception of the sound source changed from being close to the head (E-rating around 1) to the sound source being perceived very close to the loudspeaker (E-rating around 2.5). For durations above 80 ms, the externalization ratings did not change further. Thus, the window durations between 20 and 80 ms contributed mostly to externalization. Hence, a given increase in BRIR duration did not result in accordingly higher externalization ratings.

A two-way ANOVA performed on the listener's E-ratings with window duration and angle as factors showed that the window duration had a significant effect on externalization, [$F(10,110) = 49.81$, $p<0.0001$] while the effect of source angle was not significant ($p>0.05$). The interaction between window duration and source angle was also not significant ($p>0.05$).

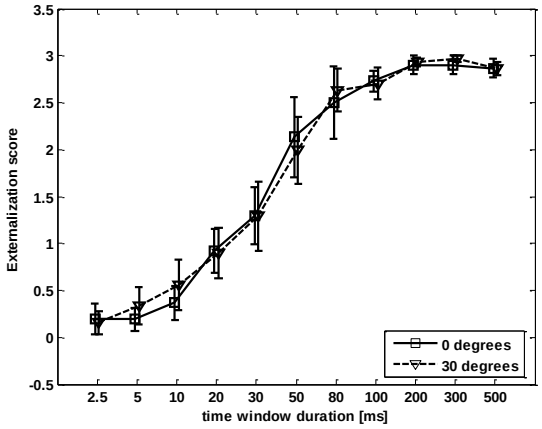

Figure 3: Across-listener mean externalization ratings for the truncated window condition (TR). The data for the sound source at 0° azimuth are indicated by the solid line and squares, and the data for the sound source at 30° azimuth are indicated by the dashed line and triangles. The error bars indicate the standard error of the mean.

The data further showed that the increase in window duration required to obtain well externalized sound images is different for different listeners. The results for the six listeners in the TR condition are shown in Figure 4, where the left panel represents the data obtained with the frontal source and the right panel indicates the data obtained with the lateral source. Listeners S2, S3, S5 and S6 showed poor externalization ratings for BRIR durations up to 10-20 ms, and after this point the ratings increased up to durations of 50-80 ms.

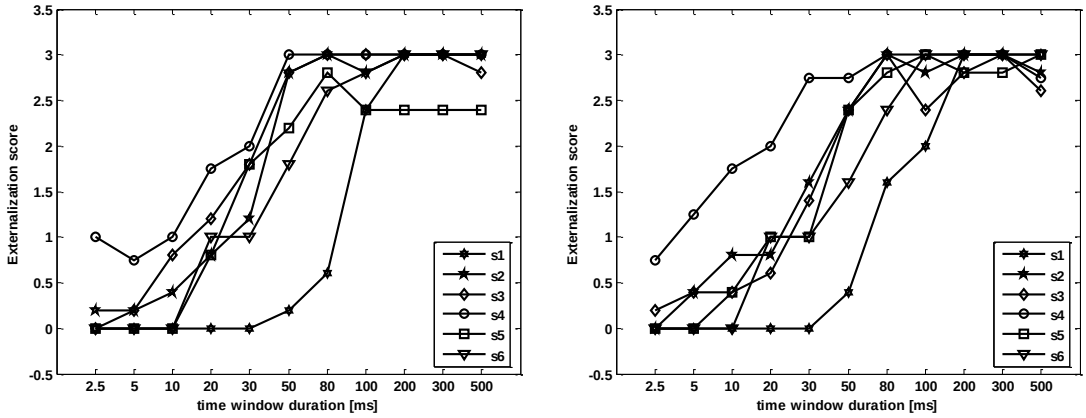

Figure 4: Individual externalization ratings for the six listeners for the truncated BRIR condition. The left panel shows the results for the source at 0° azimuth and the right panel shows the results for the source at 30° azimuth.

Listener S1 showed poor externalization ratings up to 30-50 ms. In contrast, listener S4 achieved higher externalization ratings with shorter windows compared to the average listener data. These individual characteristics were very similar for the two different measurements (i.e., the source at 0° and 30° azimuth). The individual differences observed at the "critical" window durations reflecting the transition from an almost internalized to a well externalized percept were also reflected in the large standard errors in Figure 3 for durations of 30-80 ms, while the errors were much smaller for the short or the very long durations.

3.3.1.2 Condition with monaural reverberation (MO)

The results for the MO condition with impulse responses containing identical reverberation parts in both ears are shown in Figure 5. The data for the frontal and lateral source are indicated by the solid line with squares and the dashed line with triangles, respectively. In contrast to the TR condition, the results were different here for the frontal and the lateral sound source. For the frontal source, the shape of the E-ratings followed that obtained in the TR condition. It can be seen that window durations between 10 and 50 ms gradually contribute to perceiving the source from being close to the head to being very close to the loudspeaker, whereas a further

increase in duration of the binaural part of the impulse response did not lead to a further increase of externalization. The E-ratings for the lateral sound source in this condition reached already a high level of about 1.7 for the 2.5 ms duration of the binaural part of the impulse response. Increasing this duration to just 5 ms resulted in a well externalized sound (E-rating around 2.4). Thus, these results showed large differences in externalization due to source direction.

A two-way ANOVA with the factors window duration and angle confirmed a significant main effect of window duration [$F(10,88) = 12.56$, $p<0.0001$] and source angle [$F(1,88) = 31.42$, $p<0.0001$]. The interaction between window duration and source angle was also significant [$F(10,88) = 3$, $p<0.005$], thereby reflecting the fact that externalization is affected by the source laterality in the MO condition for short and intermediate window durations.

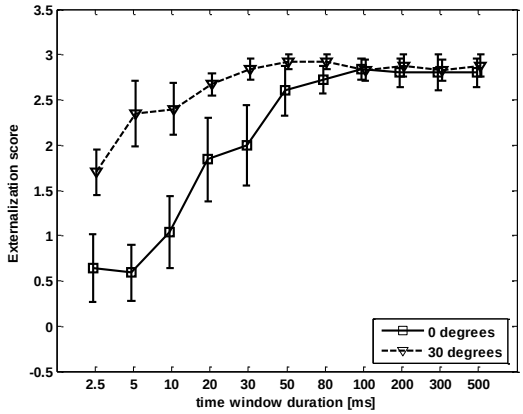

Figure 5: Across listener mean externalization ratings for the MO condition for broadband speech. The data for the sound source at 0° azimuth are indicated by the solid line and squares, and the data for the sound source at 30° azimuth are indicated by the dashed line and triangles. The error bars indicate the standard error of the mean.

3.3.1.3 Lowpass- and highpass filtered speech

The data obtained with lowpass and highpass filtered speech are shown in Figure 6. The top left panel shows the E-ratings for lowpass (LP) filtered speech condition TR. The data for the frontal and lateral source are indicated by the solid line with squares and the dashed line with triangles, respectively. The results for lowpass filtered speech in the TR condition showed low externalization ratings between 0 and 0.5 for BRIR durations below about 20 ms for both the frontal and lateral source. The degree of externalization then increased for window duration between 20 and 80 ms and longer durations did not produce any further increase in externalization. Thus, these patterns were similar to those obtained with broadband speech. However, the E-ratings for the filtered speech did not reach full externalization even for long window durations, as the maximum E-ratings were about 2.1, corresponding to a

source perceived close to the loudspeaker. The corresponding data for the highpass (HP) filtered speech are shown in the bottom panel of Figure 6. The results for HP filtered speech in the TR condition showed similar effects of window duration as the corresponding LP filtered conditions, although the increase in E-ratings was steeper for the frontal source.

A two-way ANOVA for the LP filtered speech showed that the effect of window duration was significant [$F(10,110) = 22.23$, $p<0.0001$], while the effect of angle was not significant ($p>0.05$). The interaction between the two factors was also not significant ($p>0.05$). For the highpass filtered source, the effect of window duration was significant [$F(10,110) = 18.06$, $p<0.0001$], while the effect of source angle was not significant ($p>0.05$) and the interaction between the two factors was also not significant ($p>0.05$). This confirms that, for truncated BRIRs, in both filtering conditions, the window duration affected externalization, but the effect did not depend on the angle of the sound source.

The upper right panel of Figure 6 shows the E-ratings for LP filtered speech in the MO condition. The results showed poor externalization ratings between 0.5 and 1 for BRIR durations below about 20 ms for the frontal source (solid line, square). For the lateral source (dashed line, triangles), the E-ratings for short window durations were higher (between 1 and 2). The externalization ratings for the lateral source were already close to their maximum value of about 2 at 20 ms window duration. In contrast, the E-ratings for the frontal source increased between 20 and 80 ms and reached their maximum value at 80 ms window duration. For the highpass filtered speech, shown in the lower right panel of Figure 6, the sound was perceived as being close to the head (E-rating $\approx$ 1) for short window durations, and an increase in externalization was observed up to 80 ms. In comparison, the lateral sound source was already well externalized for short window durations (E-rating $\approx$ 2).

A two-way ANOVA for the highpass filtered speech showed main significant effects of window duration [$F(10,88) = 2.48$, $p<0.02$], and angle [$F(1,88) = 4.34$, $p<0.05$]. The interaction between the two factors was not significant ($p>0.05$). For the lowpass filtered source, the effect of window duration was significant [$F(10,88) = 4.04$, $p<0.0001$], and the effect of source angle was also significant [$F(1,88) = 17.13$, $p<0.0001$]. The interaction between the two factors was not significant ($p>0.05$). This confirms that when the window duration of the binaural part of the BRIR was varied and identical reverberation was used for the remaining part of the BRIR, both the window duration and the source laterality affected externalization. This effect was found for both lowpass and highpass filtered speech.

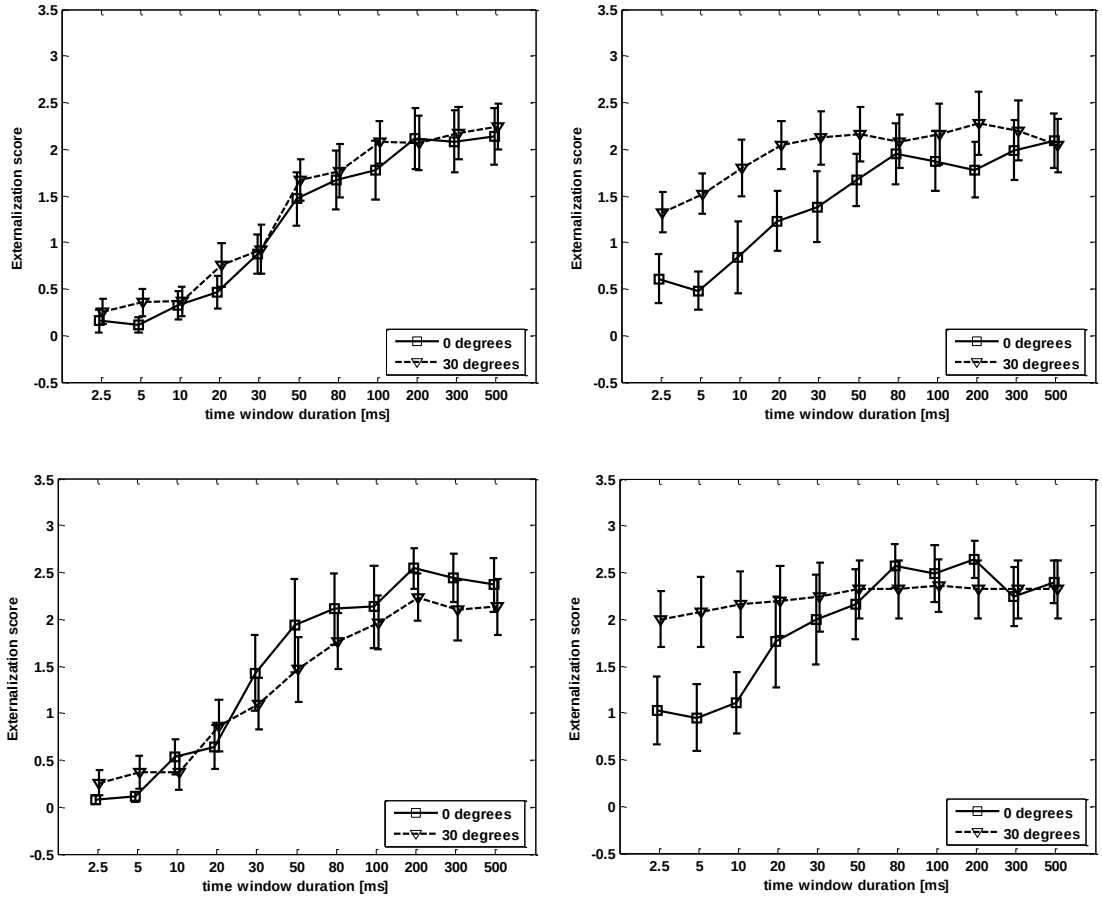

Figure 6: Across-listener mean externalization ratings for the lowpass (top panel) and highpass (bottom panel) filtered speech. The left panel shows the results for the TR condition and the right panel shows the results for the MO condition. The data for the sound sources at 0° azimuth are indicated by the solid line and squares, and the data for the sound sources at 30° azimuth are indicated by the dashed line and triangles. The error bars indicate the standard error of the mean.

3.3.2 Analysis of cues

3.3.2.1 Direct to reverberant energy ratio (D/R)

Figure 7 shows the D/R for the left ear for the conditions considered in the perceptual experiments. The D/R here was calculated as the ratio of the energy in the first 2.5 ms of the BRIR to the energy in the remainder of the BRIR. The truncated BRIR conditions are shown by the solid lines and the monaural reverberation conditions are indicated by the dashed lines. The frontal and lateral sources are represented by diamonds and squares, respectively. The differences between the D/R values for the frontal and the lateral sound sources are caused by the head shadowing effect, whereby the direct sound is larger in one ear compared to the other for the lateral source. As the left-ear signal was modified (while the right-ear signal was not changed) in the MO condition, the left-ear signal is shown here such that the effects of the modification on D/R can be seen directly.

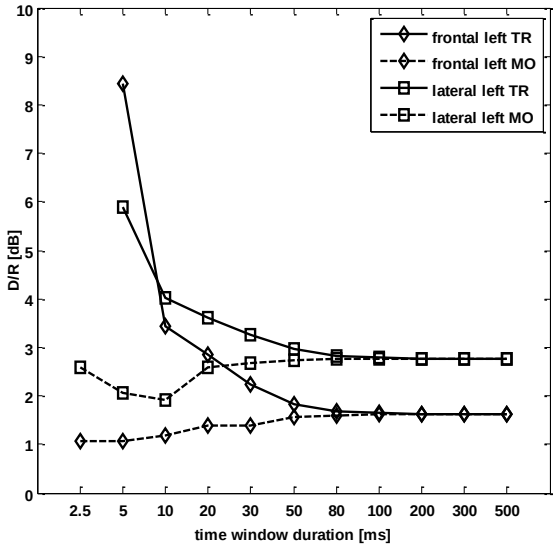

Figure 7: D/R for the left ear for the conditions considered in the perceptual experiments. The truncated BRIR conditions are indicated by the solid lines and the monaural reverberation conditions are indicated by the dashed lines. The frontal and lateral sources are indicated by diamonds and squares, respectively.

In the first 2.5 ms, the D/R takes on an infinite value in the TR condition (hence not indicated in Figure 7) and after this point it decreases as more reflected energy is added. The largest change can be seen below 20 ms and, in contrast to the experimental data, the D/R does not show a major change in the region of 20 to 80 ms as was observed in the externalization ratings. In the MO condition, the D/R is roughly constant for the different window durations (note that the 500-ms duration was the unmodified reference). There is a slight decrease in D/R for short window durations; about 0.5 dB for the frontal source and about 0.8 dB for the lateral source. This is due to the additional contribution of reflected energy from the modified very early reflections.

Comparing the D/R in the MO condition to the experimental data, it can be observed that, for the frontal source, the almost constant D/R across the different window durations does not result in constant externalization ratings. In contrast, the externalization ratings were low for short window durations where the D/R is either equal to or slightly lower than the reference D/R at 500 ms, while the sound images were well externalized for long window durations (note that a decrease in D/R should correspond to increased externalization ratings due to the increased proportion of reflected energy). In the case of the lateral source, the modification of the BRIR has a similar effect on D/R as for the frontal source. However, the experimental data showed a different trend in the externalization for the considered window durations, since the sound images were externalized even for short window durations for the lateral sound source.

3.3.2.2 Statistical ILD distributions

Figure 8 shows the distributions of short-term ILDs, similar to those used in Catic *et al.* (2013), in a single frequency band centred at 2.4 kHz with a bandwidth corresponding to one ERB. In order to illustrate the distribution of ILDs for the two source angles when identical reverberation was applied at both ears after the direct sound, the results are shown for the MO condition in the case of a 2.5 ms window duration (dashed line), for the TR condition in the case of a 2.5 ms window duration (solid line) and the full binaural impulse response reference condition (dotted line). The results for the frontal source are shown in the left panel while the right panel shows the results for the lateral source. For the frontal source, it can be seen that the distributions are narrow, as expected for the TR condition when only the direct sound is included. Adding diotic reverberation to the direct sound here also produces a narrow distribution, and the experimental data showed that the sound was perceived inside or close to the head in both these conditions. The full binaural impulse response, however, has a wider distribution, also consistent with the well externalized sound obtained in this condition.

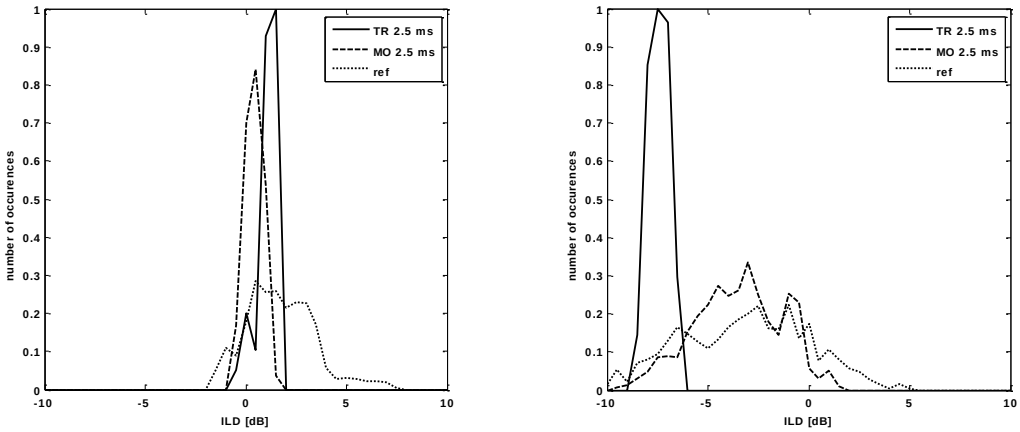

Figure 8: Distributions of short-term ILDs in a single frequency band centred at 2.4 kHz and a bandwidth of one ERB. The results for the frontal source are shown in the left panel while the right panel shows the results for the lateral source. The dashed lines indicate the results for the MO condition in the case of 2.5 ms window duration, the solid lines indicate the TR condition in the case of 2.5 ms window duration and the full binaural impulse response reference condition is indicated by the dotted lines.

In the case of the distributions of the lateral source, it can be seen that for the TR condition in the case of the 2.5 ms window duration, the ILD distribution is narrow, consistent with the low externalization rating obtained in this condition. Adding the diotic reverberation to the direct sound in this condition results in a much wider distribution, consistent with the experimental data for the MO condition, where the sound was perceived as more externalized for the lateral source. Here, the ILD distribution for the 2.5 ms window in the MO condition is slightly narrower than the

distribution for the full binaural reference, but it is much wider compared to the corresponding distribution in the case of the frontal source. Hence, the wider ILD distributions correspond to a higher degree of externalization of the sound images, while the narrow distributions correspond to lower externalization ratings.

3.3.2.3 The effect of reflection suppression

In order to illustrate the difference between ILD cues when suppression of localization cues is applied (as described in section 3.2.6) and the case when no suppression is applied, Figure 7 shows the ILD standard deviations (ILD SDs) for the considered windows in the truncated BRIR condition. Here, the circles indicate the ILD SDs when no suppression of reflections was applied and the squares indicate the results when suppression was applied in the first 15 ms after the direct sound. It can be seen that, when no suppression was applied, the ILD SDs rise rapidly for short window durations, i.e. for a window duration of 5 ms, the SD has already reached about half of its maximum value, after which it increased up to about 80 ms window duration. Compared to the experimental data in Figure 3 it can be seen that, although the externalization ratings also increased up to about 80 ms and stayed constant thereafter as observed for the ILD SDs without suppression, the ratings for the short windows below about 20 ms were low. Hence, the shape of the increase seen in the ILD SDs for windows higher than 2.5 ms was not generally observed in the experimental data. The individual data in Figure 4 showed that only one listener, S4 (circles), demonstrated an increase in externalization for short window durations, and, in addition, higher overall externalization ratings. The remaining listeners all showed poor externalization for short window durations but showed a transition towards well externalized images in the region from 20 to 80 ms. This trend is more similar to the ILD SDs when suppression of the very early reflections was applied, as here, the ILD SDs are low for window durations below 20 ms, show a transition towards high values between 20 and 80 ms, and no further increase for durations above 80 ms.

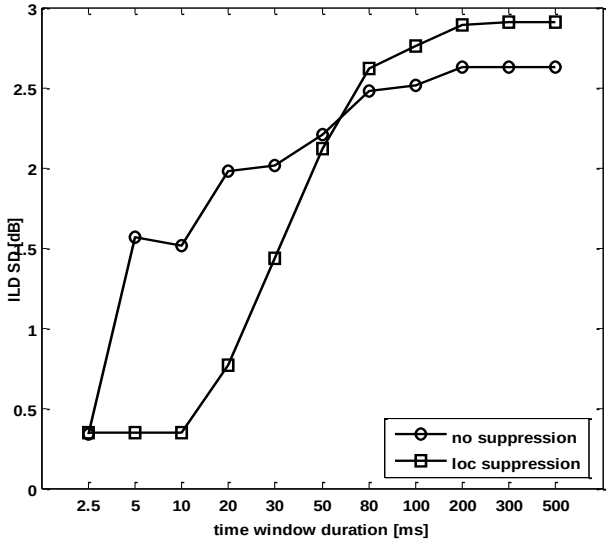

Figure 9: ILD standard deviations for the considered windows in the truncated BRIR condition for a sound source at 30° azimuth. The circles indicate the ILD SDs when no suppression of reflections in the BRIR is applied. The squares indicate the corresponding ILD SDs when suppression of reflections is applied in a window of 15 ms after the direct sound.

3.3.2.4 ILD standard deviations

Figure 10 shows the ILD SDs for all the experimental conditions considered in the present study. The left and right panels show the results for the frontal and the lateral source, respectively. The results are shown for one frequency channel with a centre frequency of 2.7 kHz and a bandwidth of one ERB, and the suppression of the very early reflections was applied. Here, ILD SDs for the two source angles are displayed separately because the dynamic binaural cues have an additional dependence on source laterality, which does not directly affect externalization, i.e. the range for the SDs for the frontal source is about 1.1 dB while it is 2.9 dB for the lateral source. The circles and diamonds indicate the results for the TR and MO conditions, respectively.

Considering the frontal source in the TR condition, for short window durations, the ILDs remain just below 0.3 dB, then increase up to about 1.1 dB with the transition being in the region of 20 to 80 ms duration, and do not increase further beyond 80 ms. This corresponds well to the shape of externalization ratings obtained in the experimental data for the broadband and highpass filtered speech in the TR condition. The ILD SDs for the MO condition are also low at short window durations (about 0.6), but higher than the corresponding ILD SDs in the TR condition. For intermediate window durations in the region 20-80 ms, the ILD SDs increase and, after this point, they do not increase further. In the corresponding experimental data for broadband and highpass filtered speech, the E-ratings followed this shape and it can also be observed that the slightly higher E-ratings obtained for short window

durations in the MO condition compared to the TR condition are reflected in the slightly higher ILD SDs in the MO condition. However, for the window duration of 30 ms, the ILDs SDs are almost the same in the TR and MO condition, while the data showed somewhat higher E-ratings in the MO condition compared to the TR condition.

For the lateral source in the TR condition, the ILD SDs display the characteristic shape seen for the frontal source (although with a less steep increase in the transition region), which corresponds well to the experimental data. In the MO condition, the ILD SDs are already high (about 2.1) for short window durations, then increase in the region 20-80 ms to about 2.9, and after this point they do not increase further. Considering the ILD SDs for the frontal and lateral source, it can be observed that the lateral source has relatively higher SDs for short window durations also when its larger dynamic range is considered. The higher ILD SDs even for short window durations here is reflected in the experimental data, where the sound source was already perceived as close to the loudspeaker in the MO condition.

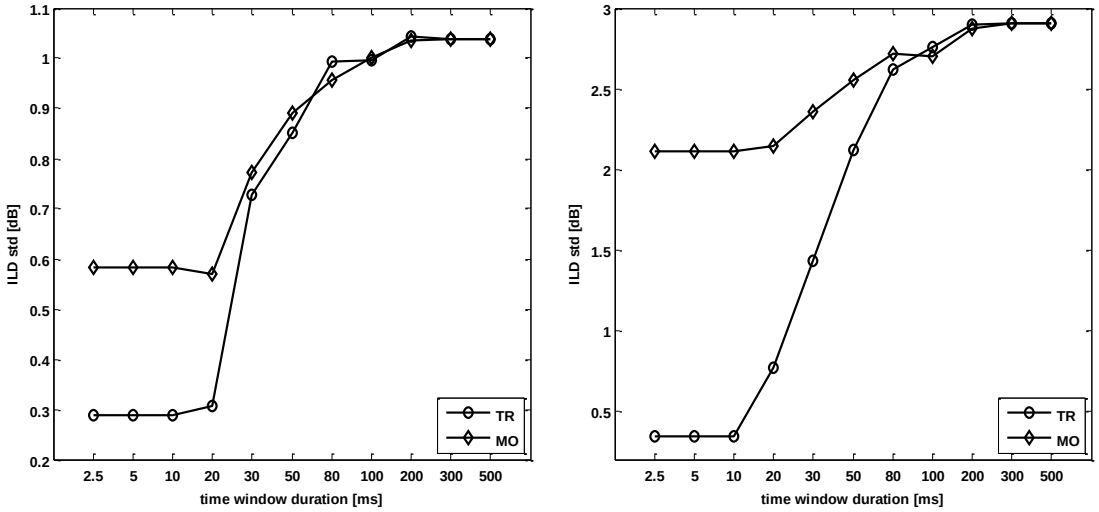

Figure10: ILD SDs for the TR and MO conditions for the frontal and lateral source. The left panel shows the results for the frontal source, where the circles indicate the TR condition and the diamonds indicate the MO condition. The right panel shows the results for the lateral source, where the TR and MO conditions are indicated by circles and diamonds, respectively.

3.3.2.5 Short-term interaural coherence statistics

Figure 11 shows the 80th and 20th short-term IC percentiles corresponding to the conditions shown for the ILD SDs shown in Figure 10. Here, the IC is shown for a low-frequency channel with a centre frequency of 830 Hz. The solid and dashed lines indicate the upper and lower percentiles in the TR condition, respectively, while the corresponding results for the MO condition are indicated by squares. It should be noted that, in the case of IC, not only the span of the IC values is relevant, but also

the IC absolute values, whereby lower ICs correspond to higher externalization ratings.

For the frontal source (left panel), both the upper and lower IC percentile are about 1 for both the TR and MO conditions, and hence the IC is both close to unity and its fluctuation is very small. This is consistent with the low externalization ratings obtained in these conditions for broadband and lowpass filtered speech. However, there is a small but consistent increase in externalization ratings (from 0.2 to 0.6) for short windows in the MO condition compared to the corresponding TR condition, and this is not observed in the IC. Both upper and lower IC percentiles decrease in the transition from window duration of 10 ms to 50 ms, and thereby both the IC fluctuation size increases and the IC absolute values decrease. Both of these effects can be related to the increase in externalization ratings in this window region. However, the lower IC percentile in the TR condition has a very steep decrease that occurs in the window region of 10-30 ms, and this is not reflected in the experimental data, which did not show such a steep increase in externalization.

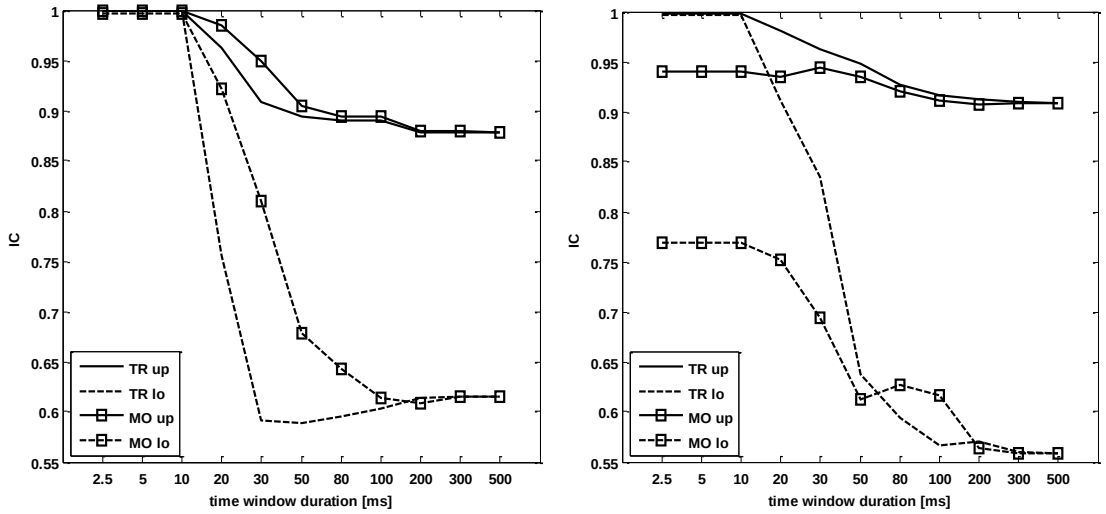

Figure11: IC 80th and 20th percentiles for the TR and MO conditions for the frontal and lateral source. The left panel shows the results for the frontal source, where the solid and dashed lines indicate the upper and lower percentiles, respectively, for the TR condition. The corresponding results for the MO condition are indicated by squares. The right panel shows the results for the lateral source, where the upper and lower percentiles are indicated by solid and dashed lines, and the corresponding results for the MO condition are indicated by squares.

For the lateral source, the IC lower and upper percentiles are around 1 for the TR condition for short window durations. Similarly to the frontal sound source, low externalization ratings were observed in the corresponding experimental data. However, the addition of diotic reverberation here results in a decrease of the absolute IC values and an increase in IC fluctuation size, where the upper percentile decreases from 1 to 0.94 and the lower percentile decreases from 1 to 0.77. This

change is reflected in the higher externalization ratings obtained for the lateral source in the MO condition. For intermediate and longer window durations, the upper and lower percentiles show the characteristic decrease followed by constant values for longer window durations in the TR condition. In the MO condition, the upper percentile remains roughly constant and the lower percentile decreases for intermediate window durations, which is also reflected in the small but additional increase in externalization ratings for the intermediate window durations in this condition. The effect that is not captured by the IC is the slight increase in externalization for the short window durations, as the IC values here are constant. The E-ratings increased from 1.7 to 2.4 from 2.5 ms to 5 ms window duration for broadband speech and, in the case of lowpass filtered speech; there was an increase in E-ratings from 1.4 to 1.7 in the range of 2.5 ms to 10 ms in the MO condition.

3.4 Discussion

3.4.1 The role of binaural and monaural reverberant cues

The present study investigated externalization perception in relation to binaural and monaural reverberant cues. Externalization based on monaural cues alone can be assessed either by selecting a monaural listening mode or by presenting identical signals to both ears that contain the details of the room impulse response and the monaural spectral HRTF cues. Such a presentation did not yield externalized sound images in Catic *et at.* (2013). However, this could have been due to the disturbance of the natural localization cues in the direct sound. Thus, a lack of externalization in such configurations does not necessarily imply that binaural reverberant cues are essential for externalization. In the current study, identical reverberation in both ears was added to the binaural direct sound part, and the results showed that externalization of sound in such a configuration depends on the laterality of the source, i.e. the azimuth of the direct sound incidence. The monaural reverberation cues were far more effective in increasing the externalization ratings of the sound source located at 30° azimuth compared to those of the sound source located at 0°. Hence, the same manipulation of the impulse responses at both angles resulted in significant differences in externalization. This suggests that binaural cues from reflections are more important for externalization when the direct sound itself contains only weak binaural cues, as it is the case for the frontal source. Binaural cues from reflections are less important when the direct sound contains stronger binaural cues that occur when the laterality of the source is increased.

Larsen *et al.* (2008) investigated listeners' ability to discriminate changes in D/R by a modification of measured BRIRs which allowed the realization of controlled variations in D/R. The measured BRIR was cut in a direct sound portion at 3 ms, and

the remaining part of the BRIR was considered reverberant. The reverberant part was then scaled relative to the direct sound in order to obtain the desired D/Rs. It was found that the changes in D/R are primarily discriminated based on monaural cues since monaural and binaural listening configurations were essentially equally effective for the discrimination of D/R. Thereby, it was shown that reverberation related cues can be perceived effectively based on monaural cues and that further binaural input does not improve performance. However, in the present study it was found that a binaural reverberation effect is still needed for robust externalization perception and that monaural cues alone are not sufficient, as externalization ratings were poor in the MO condition for the frontal source, but significantly improved when the source was moved to the side. The ILD distributions showed a clear difference for the two source azimuths, where a narrow distribution was observed for the frontal source, while a wider distribution was found for the lateral source. These differences can be explained by the disparities in the binaural cues of the direct and reverberant sound. Specifically, for the lateral source, the binaural cues in the direct sound interact with the diotic reverberation in such a way that stronger binaural fluctuations are created in the continuous speech signal due to larger disparities between the binaural cues in the direct and reverberant sound. Such disparities were much smaller for the frontal source, as the direct sound here contains only weak binaural cues and the diotic reverberation obviously does not contain any binaural differences. Hence, the interaction in this case resulted in small binaural fluctuations, and the corresponding ILD distribution was narrow. The observed relation between the width of the ILD distributions and the degree of externalization is consistent with the findings in Catic *et al.* (2013).

Although the source laterality produced differences in externalization when diotic reverberation was added to the truncated BRIR, the data demonstrated that it is not the source laterality *per se* that contributed to this difference. For the truncated BRIR condition, there was no significant effect of source azimuth for any of the considered BRIR durations, demonstrating that it is the interaction between the cues in the binaural part of the truncated BRIR and the added monaural reverberation that creates externalization for the lateral source.

Considering that reverberation improves externalization and that the discrimination of changes in reverberant energy can be achieved monaurally just as well as binaurally, it may be surprising that a binaural reverberation related component still is necessary to obtain robust externalization. Informal listening tests showed that, in a similar configuration as used in Larsen *et al*. (2008), changes in the reverberant energy can be discriminated using monaural listening. However, these changes were perceived on the basis of sound attributes different from externalization, such as the perceived amount of reverberation, or the width of the

sound image. Therefore, the results of the current study do not contradict previous studies, as the task in those studies was to identify the more reverberant sound. It appears that, although monaural cues affect spatial perception, the auditory system does not effectively utilize such cues for externalization. This observation does, however, not exclude that the monaural cues may be important for externalization in addition to the binaural cues.

3.4.2 Effects of BRIR duration and reflection suppression

The results in the current study have shown that the early reflections which arrive within the first 80 ms contribute to externalization of both broadband and band limited speech, while later reverberation does not have an effect. The importance of early reflections in this range for spatial perception is well known and has been linked to distance perception (Pellegrini, 2002) and externalization (Begault *et al.*, 2001). Begault *et al.* (2001) used auralization via headphones with simulated BRIRs. Three durations of the BRIRs were considered; an anechoic presentation, a presentation where only the first 80 ms of the BRIRs were included, and a full auralization with 1.5 s BRIRs. It was found that, while reverberation greatly improved externalization, there was no difference for the 80 ms BRIR duration and the 1.5 s duration where the late reverberation was included. This corresponds well to the findings in the present study, although the results cannot directly be compared since the externalization measure used in Begault *et al.* (2001) was limited to two degrees of externalization - a sound could either be inside or outside the head, which is different from the scale used in the present study where four degrees of externalization were considered. Since the intermediate window durations were considered in the present study, it was observed that not only are the first 80 ms of the BRIR sufficient to provide well externalized sound images, but also that the largest change in externalization occurs for BRIR durations of about 20 to 50 ms. It was found that, large changes occurred for both the ratio of direct sound energy to the reflected sound energy and the dynamic binaural cues already when a few of the very early reflections ($\leq$20 ms after the direct sound) are included. A further increase in window duration resulted in comparably smaller changes in the actual cues. Considering binaural cues, a single early reflection that arrives shortly after the direct sound can cause a large increase in the extent of the dynamic cues when it interacts with the direct sound due to its distinct binaural cues and high energy. The lack of increase in externalization in such conditions may be caused by a reflection suppression effect where the localization cues of the reflections are inhibited by the auditory system in a certain window after the direct sound. As a result, the binaural cues of the very early reflections should have little effect and thereby the resulting size of interaural fluctuations should not increase considerably. Since the suppression

diminishes for later arriving reflections, the reverberation for longer window durations is still effective and interacts with the direct sound, which generates the increase in the binaural fluctuations.

The individual differences that were observed for the critical window durations at which the sound image transitions occurred from a nearly internalized to a well externalized sound source may also be related to the precedence effect. Litovsky & Shinn-Cunningham (2001) found that the window duration after the direct sound where precedence is effective can vary substantially between listeners. This type of reflection suppression was included in the present study in the calculation of the dynamic binaural cues, where the effect of the very early reflections arriving in a window of 10-15 ms after the direct sound was suppressed. However, it should be noted that this simple approach is most likely not adequate for complex sounds considered in the present study. Much of the literature regarding the precedence effect had its focus on investigations based on simple configurations containing a single ideal reflection, (*e.g.* Litovsky *et al.*, 1999; Dizon & Colburn, 2006; Hartung & Trahiotis, 2001), and little is known about the mechanisms that contribute in complex scenarios in rooms where many reflections are active at the same time. Nevertheless, the inclusion of the simple suppression rule in the dynamic cue calculation without assumptions about the details of the mechanisms behind the precedence effect resulted in binaural cue measures that were able to capture the key characteristics in the experimental data. Therefore, it appears that the precedence effect may influence the evaluation of the dynamic binaural cues and externalization perception.

3.4.3 Externalization of lowpass- and highpass filtered speech

The results for the lowpass and highpass filtered speech showed similar trends in externalization ratings for both the TR and the MO conditions. This suggests that comparable mechanisms contribute to externalization in the two frequency regions. Compared to the broadband speech, which was fully externalized for window durations of 80 ms or longer, the bandlimited speech did not reach full externalization even for long window durations. Some listeners reported that it was more difficult to rate the filtered sound sources in general, but also that these sound sources were not perceived exactly at the loudspeaker, as was the case for the broadband source. Instead, the filtered sounds were often perceived as just in front of the loudspeaker (with an E-rating of 2) though still close to the loudspeaker for long window durations. Some of the listeners reported that the filtered sound sources did not produce the same perception of source depth as the broadband source, and therefore it was difficult to place them exactly at the loudspeaker.

Catic *et al.* (2013) investigated the effect of interaural fluctuations in different frequency regions and found no effect of the compression of ILD or ITD fluctuations at frequencies below 1 kHz. It was suggested that, for the low frequencies, the dynamic binaural cues arising from reverberation and direct sound interaction were most likely not involved in the perception of externalization and that a monaural component may be more relevant. However, in the present study, it was found that a binaural component was actually essential for externalization also for the lowpass filtered speech, as a significant effect of source direction in the MO condition was found. The short term interaural coherence was considered in the present study as an alternative measure for externalization at low frequencies and was found to be related to the experimental data to a large extent. Thus, it is possible that the short-term IC is utilized by the auditory system as the binaural reverberation related component in externalization perception. Furthermore, the dynamic ILD and IC cues were closely related, and it may be that a combined evaluation of several binaural cues plays a role at low frequencies. A strong correlation between the short-term binaural cues naturally exists, since the reflections that cause the interaural level difference fluctuations also cause the changes in interaural time differences and further decorrelate the sounds at the two ears. Thus, it can be difficult to clarify which cue is actually evaluated.

Goupell & Hartmann (2006) investigated the detection of incoherence in slightly decorrelated noise, and found that the value of the long-term coherence itself was an inadequate predictor of discrimination, and that incoherence was detected based on the size of fluctuations in interaural phase and level differences. Furthermore, Goupell & Hartmann (2007) found that the best model to describe the incoherence detection data for noise centred at 500 Hz was a model based on both the interaural phase and level fluctuations as measured by their standard deviations. Although the type of fluctuation generated due to HRTF filtering and reverberation is not the same as that of decorrelated noise (e.g. one of the differences is the presence of static binaural cues in addition to the dynamic binaural cues), this still supports the notion that the auditory system might evaluate a combination of different dynamic binaural cues for the perception of externalization. In Catic *et al.* (2013), the compression of ILD fluctuation size at low frequencies was less effective than that at high frequencies, and also less effective than the compression of the ITDs at low frequencies. In addition, the binaural cues were modified separately. Therefore, a binaural component was still available in the case of stimuli with modified binaural cues, and thus the findings in the present study for sounds at low frequencies are not inconsistent with the previous study.

3.4.4 The relation between the binaural dynamic cue measures and externalization ratings

The ILD SD measure corresponded well to the externalization ratings in most of the conditions considered in this study, although some differences in the transition region with intermediate window durations could be observed but can be attributed to the fact that the suppression window was not adjusted to each individual listener. In the MO condition, higher ILD SDs were found for both the frontal and lateral source compared to the TR condition, although relatively lower ILD SDs occurred for the frontal source in the MO condition. Even though the frontal source contains weak binaural cues, they are different from zero in many frequency channels due to the cue variability inherent in the HRTFs. Hence, some interaction between the direct sound and diotic reverberation occurs and generates an increase in the dynamic ILDs compared to the corresponding TR condition. This corresponded well to the externalization ratings that were slightly higher in the MO condition for short window durations for the frontal source. However, it was also observed that for the window duration of 30 ms, the ILD SDs for both TR and MO conditions had almost the same values, whereas the data showed slightly higher externalization ratings in the MO condition. This could be the case because, when the diotic reverberant tail is added to the truncated BRIR, the amount of suppression of reflections may decrease, and thereby the internal binaural fluctuation size would be higher than indicated by the considered ILD measure with the current suppression process. It could also be that, the added monaural reverberation cues, such as spectral variance and further lowpass filtering due to the addition of higher-order reflections, contribute to the small improvement in externalization observed in this condition.

The ILD SDs were shown in a single frequency channel in the analysis in section 3.3.2. Although there are some differences across channels, the ILD SDs in other frequency channels exhibit a similar overall shape also at low frequencies.

The size of the temporal changes in interaural coherence also corresponded well with the externalization ratings in most of the conditions considered here. In the case of the absolute values of short-term coherence, it was especially the upper percentile that followed the trends in the data, where values of about 1 and 0.9 were associated with low and high externalization ratings, respectively. A discrepancy compared to the ratings was observed in the TR condition for the frontal source for the intermediate window durations where the IC fluctuation size increases. Here, the decrease in the lower percentile was much more steep (i.e. full transition occurred between 10 and 30 ms) than the increase in externalization ratings. Furthermore, in the MO condition for the frontal source, the IC decrease was less steep than in the corresponding TR condition, whereas the data actually show a slight increase in externalization. The ILD SDs and the upper percentile of the IC did not show such a

steep transition, and this might have affected the ratings, especially if externalization is based on a combined evaluation of the binaural cues.

The ILDs and ICs were shown in single frequency channels in the analysis in section 3.3.2. While these cues showed a similar shape as a function of window in other frequency channels, there are certain differences in the cues across frequency. These differences can mainly be attributed to the variation in the HRTF related binaural cues across frequency, which means that ILD SDs can have different absolute values in different frequency channels. The influence of sound diffraction around the head and the distance between the ears results in differences in IC across channels as well. Thus, the relation to externalization perception may depend on an integration of cues across frequency, or there could be a specific weighting of channels, where some channels are weighted heavier than others. This could imply that a single-channel cue extraction may only coarsely relate to the actual data. An effective use of such cues would imply a mapping similar to the mapping of localization cues where all the HRTF variation is mapped into a single interpretation of a specific source direction. This means that the auditory system has to know which level of ILD and IC is appropriate for a specific frequency channel. This also explains why the two source angles yield different dynamic range of the ILD SD, while the externalization is the same. It could be that the size of the azimuth dependent binaural cues used for directional localization may have an effect on how the fluctuations are interpreted.

3.4.5 Limitations of the study

In the present study, speech signals were used as stimuli in the experimental data and in the binaural cue analysis since they are a relevant in everyday listening situations. However, for analysis purposes, it could have been a better choice to use noise, since noise has a more uniform distribution of energy across frequency and time, which would result in more stable estimates of the dynamic binaural cues.

The test sounds that only contained direct sound component were most often not externalized in this study, while the addition of reverberation to the direct sound resulted in externalized sound images. This does not imply that externalization in anechoic rooms is not possible. It is likely that the expectations of the listener are different in a reverberant room compared to an anechoic room, and that learning effects related to the acoustic environment occur due to listener exposure to a specific room. The available reverberation cues would be utilized which can change the weighting of the different cues that are used in the evaluation of externalization. It is also very likely that, if a listener is presented an anechoic sound in a reverberant room, the sound could be perceived as unnatural or interpreted as an extreme situation with a very close source as the D/R is practically infinite.

As the auditory system undoubtedly integrates the information received from both auditory and visual input, the visual cues in the experimental setup in the form of the loudspeakers at the positions of the sound sources could have affected the results. Still, it was deemed that the visual cue was important for identification of the sound source and as an anchor point for the two extremes of the externalization degree scale. Since the visual cues were kept constant during the course of the experiments, all the changes in externalization that were observed can be attributed to the modifications in the auditory stimulus. However, it can still be expected that the results were affected to some degree by visual cues.

The present study only considered static spatial scenarios, where head movements were not taken into account. In everyday environments, head movements and movements of the sound sources occur frequently and are a natural part of listening situations. Although well externalized sound images can be obtained in static listening situations, head movements may also have an effect on externalization of sound. Furthermore, as they also provide additional dynamic cues, this may improve externalization particularly for sound sources in the median sagittal plane, where externalization might be less robust due to the smaller binaural fluctuations.

The method of splitting of the direct and reverberant sound components as done in the present study and Larsen *et al.* (2008) and Zahorik (2002) may have affected the results at low frequencies, as it is generally not possible to perfectly isolate the direct sound response from the reflections at very low frequencies.

The BRIRs in this study were acquired using measurements in the experimental room. However, a method based on BRIR simulation could also have been used, which would have given control of the reflections and their binaural cues. Still, a measurement approach was used here, as there are certain disadvantages to the use of simulated BRIRs as well. In a simulated BRIR, the energy density of a specular reflected sound wave is much higher compared to the corresponding reflection in the measured room impulse response. The resulting energy distribution over time is much smoother in the measured BRIR compared to the simulated BRIR, and this has been shown to affect perception (Pellegrini, 2002).

4 THE EFFECT OF A VOICE ACTIVITY DETECTOR ON THE SPEECH ENHANCEMENT PERFORMANCE OF THE BINAURAL MULTICHANNEL WIENER FILTER

Abstract

A multi-microphone speech enhancement algorithm for binaural hearing aids that preserves interaural time delays was proposed recently. The algorithm is based on multichannel Wiener filtering and relies on a voice activity detector (VAD) for estimation of second order statistics. Here, the effect of a VAD on the speech enhancement of this algorithm was evaluated using an envelope-based VAD and the performance was compared to that achieved using an ideal error free VAD. The performance was considered for stationary directional noise and nonstationary diffuse noise interferers at input SNRs from -10 to +5 dB. Intelligibility weighted SNR improvements of about 20 dB and 6 dB were found for the directional and diffuse noise, respectively. No large degradations (<1 dB) due to the use of envelope-based VAD were found down to an input SNR of 0 dB for the directional noise and -5 dB for the diffuse noise. At lower input SNRs, the improvement decreased gradually to 15 dB for the directional noise and 3 dB for the diffuse noise.

This chapter is based on Catic *et al.* (2010)

4.1 Introduction

An increasing number of people suffer from hearing loss, a deficit that can limit them in their interaction with the surrounding world and often severely reduces their quality of life. The most common type of hearing loss is the sensorineural, caused by damage to the inner ear (cochlea). People with sensorineural hearing loss often find it difficult to understand speech in the presence of background noise, even when wearing their hearing aids.

Consequences of sensorineural hearing loss vary from one individual to another, but factors that often contribute are reduced audibility, loudness recruitment, reduced frequency selectivity, and reduced temporal resolution. Reduced audibility can be compensated for by a hearing aid through amplification, and loudness recruitment can to some extent be alleviated by compression. However, other contributing factors, such as reduced frequency selectivity or deficits in temporal processing cannot fully be compensated for by a hearing aid. Even if the hearing loss is located in the cochlea and the higher levels of the auditory system function well, the impaired ear may not be able to pass on the multitude of cues otherwise available in the incoming sound. The internal representation of the signals can then be incomplete and difficult to analyse.

It is well known that the intelligibility of speech is tightly connected to the signal to noise ratio (SNR) (Moore, 2003). Thus, the problem of speech intelligibility (SI) in noise can be approached by reducing the noise level. While normal-hearing (NH) people can have a speech reception threshold (SRT; the point where 50% of speech is intelligible) at SNRs in the range of -5 to -10 dB depending on the type of noise (Festen & Plomp, 1990), this threshold is typically 5-6 dB higher for hearing impaired (HI) people (Plomp, 1978). At SNRs comparable to the SRT, a small increase in SNR can improve the intelligibility scores drastically as a 1 dB increase can lead to an improvement of up to 15% (Nilsson *et al.*, 1994). This also implies that even a few dB of elevated SRT in HI listeners can cause substantial problems understanding speech compared to NH listeners. Thus, many HI listeners could benefit from a noise reduction of about 5 dB (Plomp, 1978), depending on the acoustical environment.

The noise reduction techniques used in hearing aids employ either a single microphone or multiple microphones. Single microphone techniques have been shown not to improve SI in noise but may improve listening comfort (Kates, 2008). On the other hand, multi-microphone techniques can exploit the spatial diversity of acoustic sources, ensuring that both temporal and spatial processing can be performed. Several microphone array processing techniques have been shown to improve SI in noise (Kates, 2008). Particularly, adaptive arrays can in certain

conditions reduce impressive amounts of noise. However, while the array benefit in hearing aid applications can be very large in the case of a single noise source in mild reverberation, it reduces considerably when several interfering sources are present or when the environment is reverberant (Bitzer *et al.*, 1999). This is due the use of small arrays with a limited number of microphones used in hearing aids, which limits the array performance. Nevertheless, as small improvements of a few dB might improve intelligibility significantly, a large SNR improvement is not always necessary.

One potential problem with microphone array processing is that it may affect the hearing aid user's sense of the auditory space. Some studies have shown that the users can localize sounds better when the directionality in their hearing aid is switched off (Van den Bogaert *et al.*, 2006; Keidsler *et al.*, 2006). Preserving the interaural localization cues can have a positive effect on speech intelligibility in complex acoustic environments, as the binaural processor in the auditory system can exploit additional information provided by the two ears.

Many HI people are able to take advantage of the low frequency interaural time delays (ITDs) almost as effectively as NH people (Bronkhorst & Plomp, 1989). Thus, a system that combines noise reduction with preservation of ITDs would be desirable. Such an algorithm has recently been proposed in Klasen *et al.* (2005), as a binaural extension of a multi-channel Wiener filter based speech enhancement algorithm proposed in Doclo & Moonen (2002). In Cornelis *et al.* (2010) it was shown theoretically that the binaural version preserves the interaural time delays (ITDs) and interaural level differences (ILDs) of the speech component. It was also shown that the ITDs and ILDs of the noise component are distorted in such a way that they become equal to those of the speech component. Therefore, in Klasen *et al.* (2007), the Binaural Multichannel Wiener Filter (BMWF) algorithm was extended to preserve the ITDs of the noise component. A parameter that can pass a specified amount of noise unprocessed, which is supposed to restore the binaural cues of the noise, was included into the calculation of the Wiener filters. Further, it was shown, using an objective cross-correlation measure, that the ITD cues of the noise component were preserved.

The BMWF algorithm has also been evaluated perceptually in terms of lateralization performance (Van den Bogaert *et al.*, 2008) and SRT improvements (Van den Bogaert *et al.*, 2009). The conclusion in Van den Bogaert *et al.* (2008) was that correct localization was possible with BMWF processing as long as a small amount of noise was left unprocessed. Regarding the SRT improvements in Van den Bogaert *et al.* (2009), it was concluded that the performance was as good as or better than that achieved with an adaptive directional microphone (ADM), a standard directional processing often implemented in hearing aids. The algorithm was developed for arbitrary array geometry with no need for any assumptions about the

sound source location or microphone positions, and as such it is robust against microphone gain and phase mismatch, as well as deviations in microphone positions and variation of speaker position (Doclo & Moonen, 2002). It only relies on the second order statistics of the speech and noise sources, which allows for an estimation of the desired clean speech component. The algorithm relies on a voice activity detection (VAD) mechanism for estimation of the second order statistics, i.e. the algorithm requires another algorithm that detects time instants in the noisy speech signal where the speech is absent. The studies evaluating the BMWF have used an ideal error free (perfect) VAD which is not available in practice. Generally, VAD algorithms only work well at moderate to high SNRs (Vary & Martin, 2006). It is therefore anticipated that the speech enhancement ability of BMWF in those conditions would not be degraded by using a practical VAD instead of a perfect VAD. However, for hearing aid applications, speech enhancement at low SNRs must be considered for two reasons: 1) the SNRs often found in the environment span the range of -10 to 5 dB and should therefore be included in the evaluation of algorithms for hearing aids (Ricketts, 2005) and 2) the SRT point, at which there is highest potential for improving intelligibility, is often found at negative SNRs.

In this study, it is investigated to what extent the noise reduction performance of the BMWF algorithm is affected by a *realistic* VAD compared to a perfect VAD. The BMWF is connected to an envelope based VAD and the combined system's noise reduction performance is assessed for different types of noise and different spatial configurations of noise sources. The evaluation is based on objective measures such as the intelligibility weighted SNR improvement. The paper is organized as follows. Section 4.2 provides an overview of the Binaural Multichannel Wiener Filter algorithm and the envelope-based VAD. Sections 4.3 and 4.4 describe the evaluation methods and present results with stationary directional noise and non-stationary diffuse noise. The non-stationary noise is derived from recordings in a restaurant to approach a real world situation. Section 4.5 provides a discussion of the potential use of this type of noise reduction processing in hearing aids based on the results obtained in this study.

4.2 System Model and Algorithms

4.2.1 System Model

A binaural hearing aid system is considered throughout the present study. There are two microphones on each hearing aid and it is assumed that the aids are linked, such that all four microphone signals are available to a noise reduction algorithm. The processor provides a noise reduced output at each ear.

It is assumed that the signals at each microphone, $y[k]$, at time *k*, consist of a speech (target) signal, $s[k]$, convolved with the impulse response, $h[k]$, from speech source to microphone, and some additive noise. The additive noise contains both the interfering sound source $v_n[k]$ convolved with the room impulse response from the source to microphone, $g[k]$, and the internal sensor noise $v_i[k]$, as indicated in Eq. (1) for the left and right hearing aid, respectively,

$$\begin{aligned} y_{R_m}[k] &= \left(h_{R_m}[k] \otimes s[k]\right) + \left(g_{R_m}[k] \otimes v_n[k] + v_{i_{R_m}}[k]\right) \\ y_{L_m}[k] &= \left(h_{L_m}[k] \otimes s[k]\right) + \left(g_{L_m}[k] \otimes v_n[k] + v_{i_{L_m}}[k]\right) \end{aligned} \qquad \textbf{Eq. (1)}$$

with m = 1,2 representing the microphone number index in the two hearing aids. It is assumed that the noise is uncorrelated with speech and is a short-term stationary zero-mean process.

4.2.2 Binaural Multichannel Wiener Filter

The BMWF algorithm proposed in Klasen *et al.* (2007) provides a Minimum Mean Square Error (MMSE) estimate of the speech component in the two front microphones. As depicted in Figure 1, two Wiener filters are computed to estimate the noise components $\tilde{v}_L[k]$ and $\tilde{v}_R[k]$ in the front left and right microphones, which are then subtracted from the original noisy speech signals $y_L[k]$ and $y_R[k]$ to obtain estimates $\tilde{x}_L[k]$ and $\tilde{x}_R[k]$ of the clean speech components.

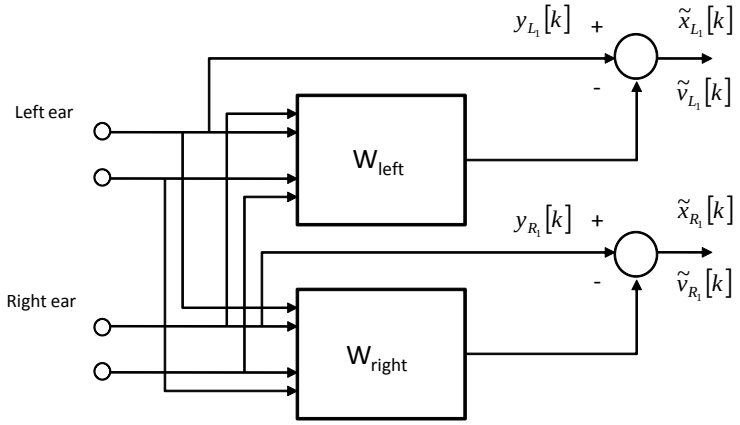

Figure 1 Structure of the BMWF algorithm. Clean speech components are obtained by computing two Wiener filters that estimate the noise component in the left and right front channels, which are subtracted from the received noisy signals.

Computation of the left and right Wiener filters requires spatio-temporal information about the speech and noise sources in the form of their second order statistics. Using the received microphone signals, an approximation of the second order statistics can be obtained from a block of input data of length *K*. For a filter of length *N* per channel, the input data vector $\mathbf{y}_{L_1}[k]$ for the left front channel is given in Eq. (2).

Accordingly, input data vectors are defined for the remaining channels. An input data vector $\mathbf{y}[k]$ for all microphone signals is constructed as expressed in Eq. (3), which is used for computing the correlation matrices of speech and noise.

$$\mathbf{y}_{\mathrm{L}_1}[\mathrm{k}] = \left[y_{L_1}[k] \;\; y_{L_1}[k-1] \;\cdots\; y_{L_1}[k-N+1] \right]^T \qquad \textbf{Eq. (2)}$$

$$\boldsymbol{y}[k] = \left[\boldsymbol{y}_{L_1}^T[k] \;\; \boldsymbol{y}_{L_2}^T[k] \;\; \boldsymbol{y}_{R_1}^T[k] \;\; \boldsymbol{y}_{R_2}^T[k] \right]^T \qquad \textbf{Eq. (3)}$$

The speech plus noise correlation matrix $\mathbf{R_{YY}}(\mathrm{m})$, given in Eq. (4), can be calculated directly from the input data vector in Eq. (3).

$$\mathbf{R_{YY}}(k) = \mathbf{R_{YY}}(k-1) + \boldsymbol{y}(k)\boldsymbol{y}(k)^T \qquad \textbf{Eq. (4)}$$

The noise components are not directly available, as they cannot be separated from the mixture of speech and noise in the received microphone signals in Eq. (2) and Eq. (3). Therefore, they need to be estimated in periods that only contain noise, in order to compute the second order statistics of the noise. Such an operation requires a voice activity detection (VAD) mechanism to identify the time instants in the received mixture signal that do not contain speech. At these time instants, denoted k^n, the noise correlation matrix $\mathbf{R_{vv}}(\mathrm{m})$ is calculated as expressed in the following:

$$\mathbf{R_{vv}}(k) = \mathbf{R_{vv}}(k-1) + \boldsymbol{y}(k^n) \left[y_{L_1}[k^n] \;\; y_{R_1}[k^n] \right] \qquad \textbf{Eq. (5)}$$

As the noise correlation matrix is constructed from q data samples collected at time instants k^n, the correlation matrices are scaled such that $\mathbf{R_{YY}} = \mathbf{R_{YY}}/K$ and $\mathbf{R_{vv}} = \mathbf{R_{vv}}/q$. The left and right Wiener filters $\mathbf{W_{LR}}$ are then calculated as shown in the following):

$$\mathbf{W_{LR}} = \left[\mathbf{W_{Left}} \; \mathbf{W_{Right}} \right] = \mathbf{R_{YY}}^{-1}\mathbf{R_{vv}} \qquad \textbf{Eq. (6)}$$

Since the speech signal is estimated in the left and right microphone channel, the BMWF processing inherently preserves the ITD cues of the speech component. However, ITD cues of the noise component are distorted (Cornelis *et al.*, 2010; Klasen *et al.*, 2007). In order to improve localization, some noise is left unprocessed at the output, by incorporating a parameter λ into the filter calculation in Eq. (6), as shown in Eq. (7).

$$\mathbf{W_{LR}} = \left[\mathbf{W_{Left}}\ \mathbf{W_{Right}} \right] = \lambda \mathbf{R_{YY}}^{-1} \mathbf{R_{vv}} \qquad \textbf{Eq. (7)}$$

The noise controlling parameter λ can take on values between 0 and 1, where λ=1 puts all effort on noise reduction with no attempt on preservation of localization cues, and λ=0 puts all effort on preserving localization cues and no noise reduction is performed, i.e., there is a trade-off between noise reduction and preservation of localization cues.

The BMWF algorithm uses no information for computation of the filter matrix other than the second order statistics determined by the VAD. It can be expected that the performance of the BMWF will degrade at some point due to VAD detection errors, leading to incorrect noise estimation. If speech is detected as noise, vectors containing speech samples will be added to the noise data matrix in Eq. (5), which leads to cancellation of parts of the speech signal. On the other hand, if too many actual noise samples are detected as speech, less noise vectors are added to the noise data matrix in Eq. (5) and a poorer noise estimate is obtained which leads to incorrect noise reduction. Generally, a multichannel Wiener filter can be decomposed into a minimum variance distortionless response MVDR beamformer followed by a (spectral) Wiener post-filter (Brandstein & Ward, 2001). Therefore, it can also be expected that the speech enhancement strongly depends on the spatial configuration of the noise sources. The adaptive beamformer is mostly effective at suppressing interference comprising fewer sources than the number of microphones, with the noise reduction decreasing fast as the number of noise sources increases. While the beamformer should not modify the target signal, the post-filter can attenuate the target signal, according to the amount of noise present at the output of the beamformer. Hence, as the Wiener post-filter trades off target distortion with noise reduction, the amount of target cancellation is expected to be small in the case of few noise sources, and high for many sources.

4.2.3 Voice Activity Detector

Speech has strong amplitude modulations in the frequency region of 2-10 Hz, such that its envelope fluctuates over a wide dynamic range. Many types of noise (e.g. such as traffic or babble noise where signals of many speakers are superimposed) exhibit smaller and more rapid envelope fluctuations compared to speech. These properties can be exploited for detection of time periods in a signal where speech is absent. Therefore, an envelope-based VAD developed for hearing aid applications is used, as proposed in Marzinzik & Kollmeier (2002). The algorithm adaptively tracks the dynamics of a signal's power envelope and provides speech pause detection based on the envelope minima in a noisy speech signal. This VAD has been shown to have a low rate of speech periods falsely detected as noise

even at low input SNR of -10 dB (Marzinzik & Kollmeier, 2002), which is desirable in order to avoid deteriorations of the speech signals in the noise reduction process. Also, in Marzinzik & Kollmeier (2002), the VAD was compared to the standardized ITU G.729 VAD by means of receiver operating characteristic (ROC) curves, and was found to outperform it for a representative set of noise types and SNRs. The VAD provides speech/noise classification by analysing time frames of 8 ms, using the following processing steps for each frame:

- A 50% overlap is used such that the processing delay is 4 ms. Each frame is Hanning windowed and a 256-point FFT is performed.
- Short term magnitude squared spectra were calculated. Temporal power envelopes are obtained by summing up the squared spectral components. Moreover, a low- and highband power envelope are calculated, by summing up the squared spectral components below a cutoff frequency f_C and above f_C. The envelopes of band-limited signals are considered since some noise types have stronger low (or high) frequency components. In that case, one of the band-limited envelopes may be less disturbed by the noise and provide more reliable information for speech pause decision. The envelopes are smoothed slightly using a first order recursive lowpass filter with a release time constant τ_E.
- The maxima and minima of the signal envelope are obtained by tracking the peaks and valleys of the envelope waveform. This is done with two first order recursive lowpass filters with attack and release time constants τ_{raise} and τ_{decay}. The differences between the maxima and minima are calculated to obtain the current dynamic range of the signal.
- The decision for a speech pause is based on several requirements regarding the dynamic range of the signal and the current envelope values for the three bands. As the complete decision process is described in Marzinzik & Kollmeier (2002), it will not be outlined here, i.e. only the general concepts are provided. The criterion for the envelope being close enough to its minimum is determined by the free parameters β and η and the current dynamic range of the signal. The threshold parameter η represents the threshold for determining whether the current dynamic range of the signal is low, medium or high. The parameter β can take on values between 0 and 1 and is used in comparisons of whether a fraction (β) of the current dynamic range is higher than the difference between the current envelope and its minimum. The settings of β and η determine how strict the requirements for detecting a speech pause are, and they can be adjusted to make the VAD more or less sensitive to detecting speech pauses. By increasing one or both

of the parameters, the algorithm will detect more speech pauses, but at the same time, it will also detect more speech periods as noise.

4.3 Evaluation Setup

The speech enhancement performance of the system was evaluated for SNRs in the range from -10 to +5 dB, as this range is most important for hearing aid applications (see Section 4.1). Since the performance of microphone arrays strongly depends on the spatial characteristics of the interfering noise, the system was evaluated both in conditions of directional and diffuse noise. Further, two noise types were considered: a stationary noise with low modulation index and a non-stationary noise with strong envelope fluctuations.

4.3.1 Performance measures

The noise reduction performance was evaluated using the intelligibility weighted SNR improvement, $\mathrm{SNR_{INT}}$, defined in Greenberg *et al.* (1993). This is a measure of noise reduction that incorporates basic factors related to speech intelligibility in noise. The signals were split into *i* third octave bands where the SNR (in dB) was calculated for each band *i*, as shown in Eq.(8) for the input and output of the noise reduction algorithm, respectively. Here, *P(f)* represents power spectral density, with the subscripts *S* and *N* denoting the speech and noise components, respectively. As different frequency bands do not contribute equally to the intelligibility of speech, each band with centre frequency f_i^c was weighted with a weight $\mathrm{I_i}$ according to its importance for speech intelligibility. The centre frequencies and weights are defined in ANSI S.3.5-1997. The weighting function has roughly a bandpass characteristic, with a passband of 1-3 kHz. Since the improvement in SNR after processing is of interest, $\Delta\mathrm{SNR_{INT}}$ was calculated as expressed in Eq.(9), where the input SNR was subtracted from the output SNR

$$\mathrm{SNR}_{i,\mathrm{in}} = 10\ \log_{10}\left(\frac{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{S,\mathrm{in}}(f)df}{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{N,\mathrm{in}}(f)df}\right)$$

$$\mathrm{SNR}_{i,\mathrm{out}} = 10\ \log_{10}\left(\frac{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{S,\mathrm{out}}(f)df}{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{N,\mathrm{out}}(f)df}\right) \qquad \textbf{Eq. (8)}$$

$$\Delta\mathrm{SNR_{INT}} = \sum_i I_i\left(\mathrm{SNR}_{i,\mathrm{out}} - \mathrm{SNR}_{i,\mathrm{in}}\right) \qquad \textbf{Eq. (9)}$$

Several studies on microphone arrays for hearing aids have found good agreement between the weighted SNR improvement and changes in SRTs for normal-hearing individuals (Peterson *et al.*, 1990; Hoffman *et al*. 2009). In Laugesen & Schmidtke (2004), a close agreement between the AI weighted directivity index (AI-DI) (in the case of diffuse noise and frontal incidence of target, the $\Delta\mathrm{SNR}_{\mathrm{INT}}$ approaches the AI-DI) and SRTs for hearing impaired listeners was reported. Although it can be expected that an improvement in SNR in the frequency regions important for speech intelligibility should improve speech recognition, this measure is not considered as a substitute for speech intelligibility tests with hearing impaired listeners.

Cancellation of speech can occur when the VAD erroneously detects speech periods as noise periods, due to speech samples being added to the noise data correlation matrix in Eq. (5). Speech cancellation can also occur due to the BMWF algorithm processing. This effect may not always be reflected in the SNR improvement, since the noise can be reduced accordingly. The speech cancellation ($\mathrm{SC}_{\mathrm{INT}}$) was therefore calculated as the ratio of the speech signal output power to speech signal input power, frequency weighted and averaged in dB, similar to the intelligibility weighted SNR calculation described above:

$$\mathrm{SC}_i = 10\ \log_{10}\left(\frac{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{S,\mathrm{out}}(f)df}{\int_{-2^{1/6}f_i^c}^{2^{1/6}f_i^c} P_{S,\mathrm{in}}(f)df}\right) \qquad \textbf{Eq. (10)}$$

$$\mathrm{SC}_{\mathrm{INT}} = \sum_i I_i\ \mathrm{SC}_i$$

4.3.2 Reference system

In order to quantify the degradation of the BMWF system performance due to the integration of a realistic VAD mechanism in the noise estimation method, it was necessary to have a reference VAD that performs "perfectly". Ideally, a VAD should detect all the noise samples without cutting parts of speech. The reference VAD sequence was derived by running the implemented envelope-based VAD algorithm on the speech material used for target speech, mixed with a very low-level noise signal (speech weighted noise at -35 dB SNR) to ensure correct speech/noise classification, as shown in Figure 2. This VAD sequence was used as the reference VAD here and is from now on referred to as "perfect" VAD, while the VAD running on the actual signals is referred to as envelope-based VAD. The noise reduction obtained with BMWF using the perfect VAD can be regarded as the optimum for the considered acoustic scenarios.

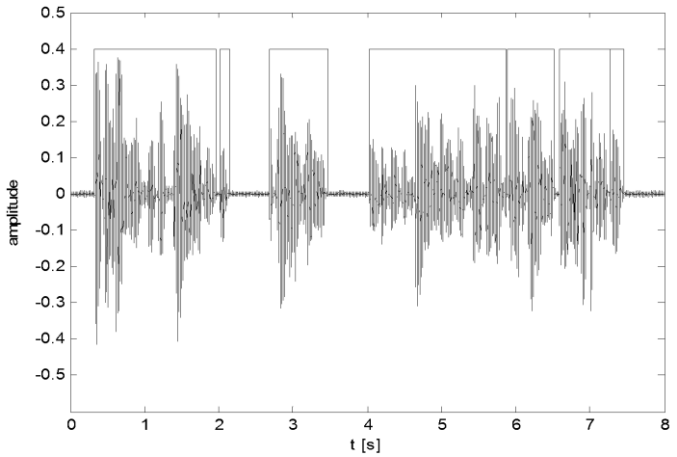

Figure 2 Target speech waveform accompanied by the binary sequence representing the perfect VAD. The selected speech pauses are indicated by zeroes in the binary sequence.

4.3.3 Experimental setup

The measurements of speech and noise were carried out in an acoustically highly damped room. The speech and noise sources were recorded separately on behind the ear (BTE) hearing aids with omni-directional microphones, mounted on a dummy head which was placed in the centre of the room. The speech waveform is shown in Figure 2. The 8 seconds long speech segment is a male speaker on BBC news, where an additional speech pause was added to the waveform in the intervals from 3.5 to 4 seconds and 7.5 to 8 seconds. This was done since there are very few natural speech pauses in the newsreader speech, and because the BMWF relies on presence of speech pauses for noise estimation. It is assumed that, in a more natural conversation, several speech pauses would be present in the waveform. The speech was played through a loudspeaker located at 0° azimuth relative to the dummy head. The stationary noise used was speech-shaped noise, which is a steady noise with the same long term average spectrum as (typical) speech. The noise was recorded at the House Ear Institute in Los Angeles. In order to generate directional noise, this recording was played through a loudspeaker positioned at an azimuth of 90° relative to the dummy head. The non-stationary noise used was diffuse multi-talker babble noise. Further recording were made in a restaurant at 8 different locations. These recordings were played from 8 different loudspeakers located in the corners of the room. This artificial diffuse sound field is assumed to mimic a "cocktail party" situation, and was chosen to assess the performance of BMWF combined with envelope-based VAD in a realistic and challenging acoustical environment.

The sampling frequency was 24.414 Hz and the BMWF filter length per channel was 64. The filters in Eq. (7) were calculated using the whole signal. The output speech and noise signals were generated by filtering the clean speech and noise signals separately with the obtained filter coefficients. The input SNRs were

calculated using the VAD sequence shown in Figure 2 in order to exclude the noise-only samples indicated by zeroes from the calculation.

In order to investigate the combined systems' noise reduction performance, including the effect of the noise controlling parameter λ that trades off noise reduction with preservation of ITDs, two different settings of λ were used: λ=1, corresponding to full effort on noise reduction, and λ=0.8, corresponding to adding a small amount of unprocessed noise to the output. These values were chosen since it was found in Van den Bogaert *et al.* (2008) that by passing a small amount of unprocessed noise (λ=0.8), the localization can be preserved also for the noise component, while λ=1 distorts the localization of the noise component but provides more noise reduction. The λ parameter was kept fixed in all situations, i.e., it was assumed that the hearing aid user does not adjust this according to the acoustical situation. The algorithmic parameters for the VAD used in the current implementation were determined empirically in Marzinzik & Kollmeier (2002) based on tests employing several noise types, speech signals, and input SNRs. However, since these parameters were adjusted to yield a low false alarm rate (which consequently results in a low hit rate), two additional values of β were considered here, as an increase in β yields a larger speech pause hit rate. This also allowed the investigation of different combinations of speech and noise classification errors. The complete list of VAD parameters is shown in Table 1.

Table 1: List of parameters used in VAD implementation

frame length T	8 ms
no of FFT points N_0	256
sampling frequency f_S	24.414 kHz
cutoff frequency f_C	2 kHz
smoothing time constant τ_E	32 ms
minima tracking time constant τ_{decay}	3 s
maxima tracking time constant τ_{raise}	3 s
threshold parameter η	5 dB
threshold parameter β	0.1, 0.2 and 0.3

4.4 Results

4.4.1 Speech and noise classification

In this section, the speech and noise classification performance of the envelope-based VAD for the three settings of β is presented. The percentages of correctly detected samples were calculated for the scenarios described in the experimental setup in section 4.3. Hence, the noise reduction and speech cancelation obtained for each scenario in Sections 4.4.2 and 4.4.3 can directly be related to this particular classification performance. The correct scores were calculated with respect to the perfect VAD sequence from Figure 2 (Section 4.3). Note that the length of the entire signal was 8 seconds of which about 2 seconds were noise and so the amount of speech and noise is not equal.

In Figure 3 the percentages of correct scores are shown for the diffuse multi-talker babble noise for β=0.1 (solid curve), β=0.2 (dashed curve) and β=0.3 (dotted curve). The left and right panels show the correct scores for the speech and noise periods, respectively. For β=0.1, the amount of correctly detected speech samples is at least 95% at all input SNRs. However, only about 15-20% of the actual noise samples are detected as noise. This is partly due to the way the VAD tracks the minima in the envelope, and due to the threshold settings used to obtain a speech pause decision. The multi-talker babble noise fluctuates strongly, such that its envelope is rarely as close to its minimum as is required in the algorithm for a speech pause decision. Increasing β improves the classification of noise, which is mostly pronounced at higher SNR, but this comes at the expense of more speech being classified as noise. It should be noted, that some of these errors occur at time instants when the speech signal is weak, and hence may not always be detrimental.

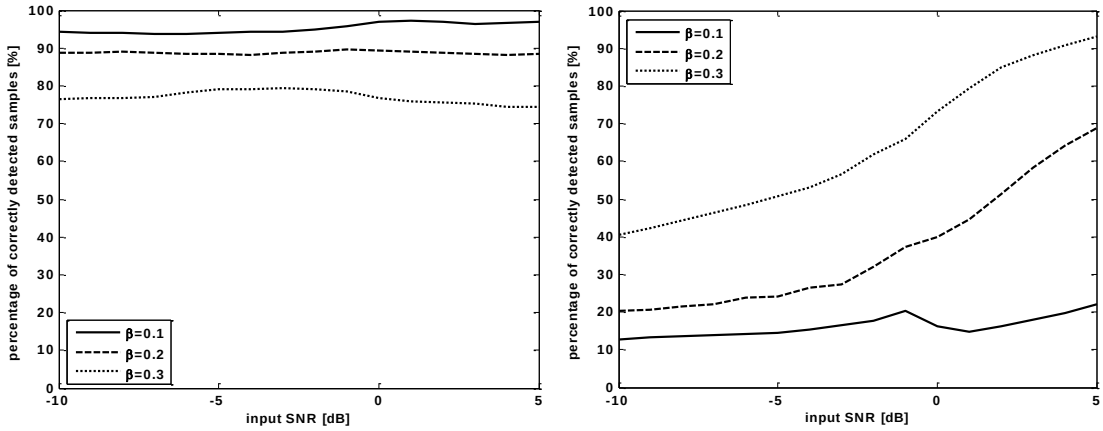

Figure 3 Percentage of correctly detected samples for diffuse multi-talker babble noise as interferer, at different SNR and for β=0.1, 0.2 and 0.3. Left panel: speech period, right panel: noise period.

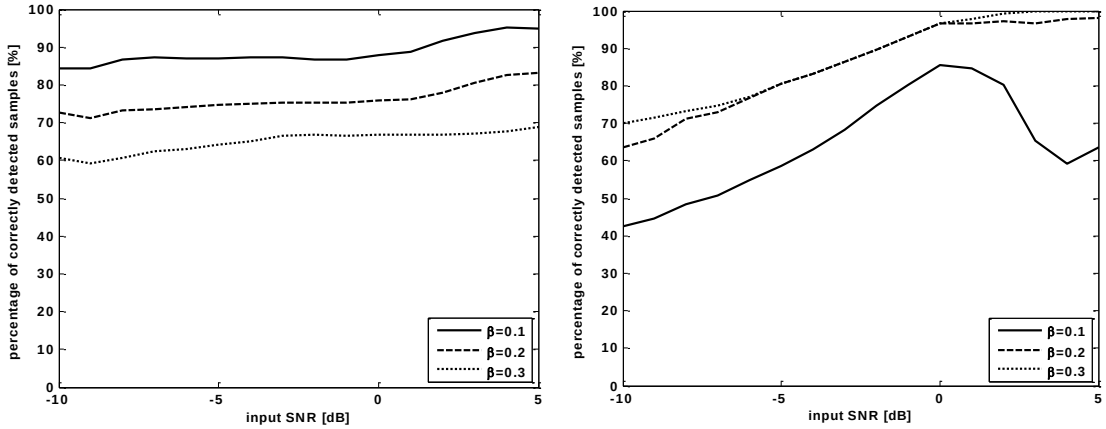

Figure 4 Percentage of correctly detected samples for directional speech-shaped noise as interferer, at different SNR and for β=0.1, 0.2 and 0.3. Left panel: speech period, right panel: noise period.

In Figure 4 the percentages of correct scores are shown for the directional speech shaped noise for β=0.1 (solid curve), β=0.2 (dashed curve) and β=0.3 (dotted curve). The left and right panel show the correct scores for speech and noise period, respectively. For β=0.1, the amount of correctly detected speech samples is at least 85% at all SNRs. Compared to the multi-talker babble noise, the speech shaped noise exhibits smaller fluctuations of the envelope. Thus the VAD demonstrates significantly better detection of the actual noise frames, but also a higher amount of incorrectly classified speech. Increasing β from 0.1 to 0.2 improves the overall noise classification, with correct scores on the order of 98% down to an input SNR of 0 dB. Below this point, the amount decreases gradually to 64%. Further increase of β to 0.3 only slightly improves the noise classification, but each increase in β results in an increased error in speech classification.

4.4.2 Stationary directional noise

Figure 5 shows the intelligibility weighted SNR improvement ΔSNR_{INT} for stationary directional noise when the perfect VAD is used for the noise estimation (solid curve), and when the envelope-based VAD is used with β=0.1 (dashed curve), β=0.2 (dotted curve), and β=0.3 (solid curve with cross markers). The left panel and right panel show the results for λ=1 and λ=0.8, respectively. For β=0.2 and β=0.3, the noise reduction performance does not degrade due to VAD down to an input SNR of 0 dB, where an improvement of about 20 dB SNR is obtained. This can be related to the speech and noise classification shown in Figure 4, as a high amount of noise is correctly detected for the two β settings down to an input SNR of 0 dB. In this condition, the setting β=0.1 yields less improvement, which is also consistent with the 15-30% lower detection rate for noise observed in Figure 4. In this context, the increased misclassification of speech due to increasing β does not have a negative impact on noise reduction performance. Below an input SNR of 0 dB, the noise

suppression gradually decreases for all β settings, and eventually amounts to roughly 15 dB at an input SNR of -10 dB.

The right panel of Figure 5 shows that reducing λ from 1 to 0.8 (to preserve ITD cues of the noise component) leads to SNR improvement of about 13 dB for all considered SNR conditions when utilizing perfect VAD. This is substantially less than the 20 dB obtained with the λ=1 setting. However the degradation of noise reduction performance due to employing envelope-based VAD is smaller when the noise estimate is scaled, such that an average gain of 10 dB is found.

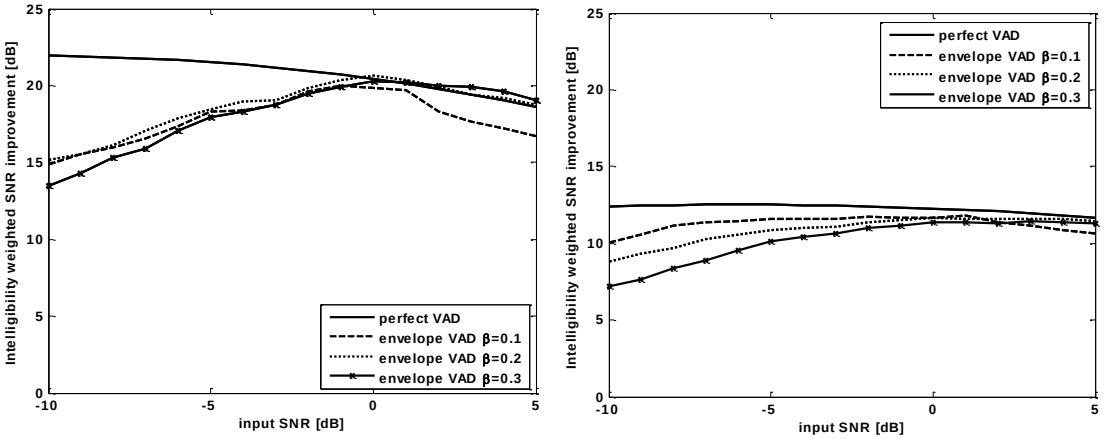

Figure 5 Intelligibility weighted SNR improvement for directional speech shaped noise at different SNRs for perfect VAD and envelope based VAD with β=0.1, 0.2 and 0.3. Left panel: λ=1 and right panel: λ=0.8.

Figure 6 shows the intelligibility-weighted speech cancellation SC_{INT} for the same conditions as for the ΔSNR_{INT} in Figure 5. (Note that a smaller number indicates higher target cancelation) The SC_{INT} ranges from 0.2 to 1 dB when the perfect VAD is employed. When envelope-based VAD is employed, the SC_{INT} increases, with higher β resulting in increased cancellation, as more speech is classified as noise. This increase is modest at higher input SNR but becomes progressively greater at lower SNR.

Results in the right panel of Figure 6 show that setting λ=0.8 reduces the amount of target cancellation by up to 1.5 dB.

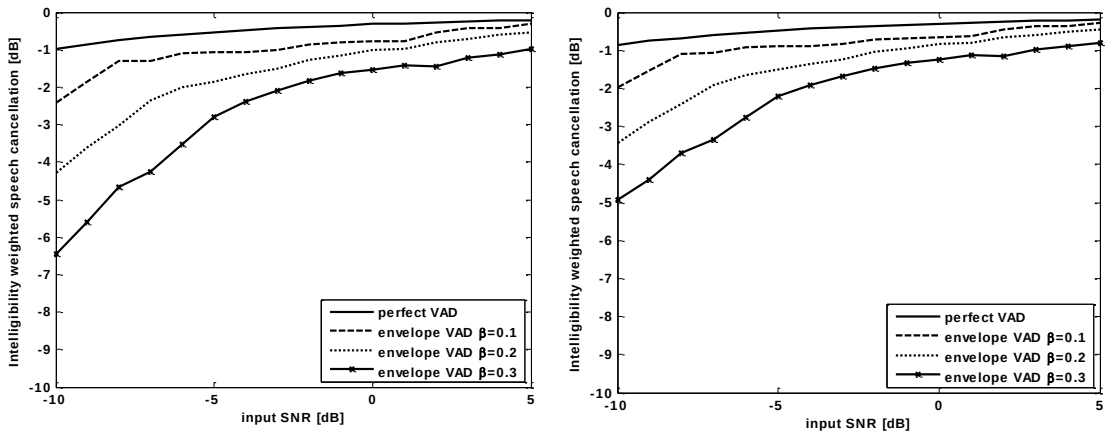

Figure 6 Intelligibility weighted speech cancelation for directional speech shaped noise at different SNRs for perfect VAD and envelope based VAD with β=0.1, 0.2 and 0.3. Left panel: λ=1 and right panel: λ=0.8.

4.4.3 Diffuse and fluctuating noise

Figure 7 shows the intelligibility-weighted SNR improvement for a diffuse multi-talker babble scenario with the same conditions as for stationary noise (section 4.4.2). The noise suppression is around 6 dB with a slight decline below input SNR of -5 dB when the perfect VAD is employed. Using the envelope-based VAD does not result in large degradations (<1 dB) down to an input SNR of -5 dB, at least for the β=0.3 setting (this β value yields the highest noise reduction). Below -5 dB, the noise reduction degrades gradually to about 3 dB at -10 dB.

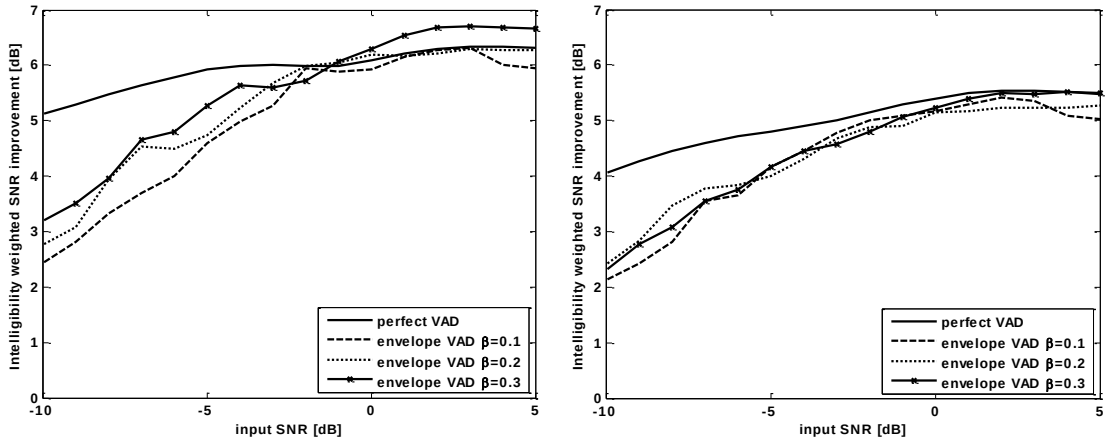

Figure 7 Intelligibility weighted SNR improvement for diffuse multi-talker babble noise at different SNRs for perfect VAD and envelope based VAD with β=0.1, 0.2 and 0.3. Left panel: λ=1 and right panel: λ=0.8.

The detection rates for noise displayed in Figure 3 show that, as the input SNR decreases, the VAD classifies a higher amount of noise as speech. But this is not the only reason for reduced performance. Figure 3 shows that the VAD detection rates are quite similar at and below -5 dB input SNR, yet the SNR improvement decreases.

The noise reduction performance does not only depend on the VAD error rates, but also on the quality of the noise estimate and this is especially pronounced at very low SNRs in non-stationary noise. The non-continuous collection of noise data introduces inaccuracies in the noise correlation matrix since it is estimated only in limited periods of time in the entire signal waveform. Thus, the filter coefficients differ from those that could have been obtained if the speech and noise correlation matrices were estimated at the same time. While the improvement for directional speech shaped noise in Figure 5 actually increases with decreasing SNR when employing a perfect VAD, this is not the case for diffuse babble noise (Figure 7), where a 1 dB decrease is seen. Therefore, frequent sampling of the fluctuating noise is even more important at lower SNRs.

The right panel of Figure 7 shows that a setting λ=0.8 in diffuse noise results only in a very small decrease in SNR improvement (on average 1 dB).

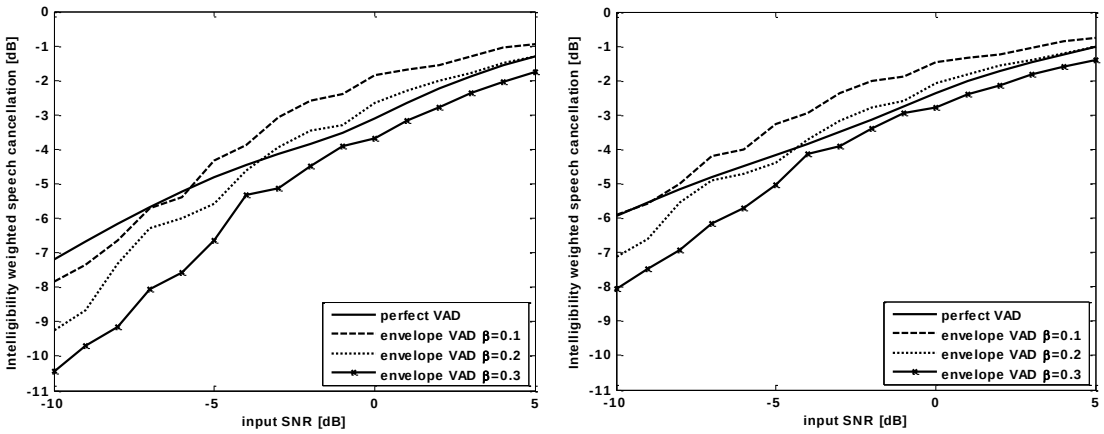

Figure 8 Intelligibility weighted speech cancellation for diffuse multi-talker babble noise at different SNRs for perfect VAD and envelope based VAD with β=0.1, 0.2 and 0.3. Left panel: λ=1 and right panel: λ=0.8.

The target cancelation for the multi-talker babble interferer is shown in Figure 8. Most of the target cancellation occurs due to the BMWF processing, which ranges from 1.5 to 7 dB depending on the input SNR. Since the noise is diffuse, the data-dependent spatial filter is not as effective as in the case of a few noise sources, and consequently the spectrum-dependent post-filter attenuates the signal in the effort to reduce the considerable amount of residual noise at the output of the spatial filter. The additional target cancelation due to VAD errors is around 3 dB at most and in some cases the SC_{INT} is actually lower than that obtained with the perfect VAD. Thus the amount of cancellation for diffuse babble noise due to VAD errors is limited. The right panel of Figure 8 shows that scaling the noise estimate by setting λ=0.8 reduces the target cancelation by up to 2.5 dB.

4.5 Discussion

The noise reduction results showed that for stationary directional noise an average SNR improvement of 20 dB (see left panel of Figure 5) can be achieved when using perfect VAD for noise estimation in the BMWF system. The effect of incorporating a realistic VAD for this scenario is minimal (<1 dB) as long as the input SNR is at or above 0 dB. Although noise reduction performance deteriorated with decreasing SNR, a robust gain of about 15 dB is still obtained at -10 dB input SNR. When trading off some noise reduction in order to preserve ITD cues of the noise component (i.e., setting λ=0.8, shown in right panel of Figure 5), an adequate improvement in SNR of 10 dB on average can still be obtained. This means that in such a situation, the user could, in addition to the benefit from auditory release from masking (that also improves speech intelligibility), also benefit from the microphone array processing. While an adequate amount of noise reduction can be obtained for the case of stationary directional interferer, the noise recorded in a restaurant is a more realistic condition that often would be encountered by hearing aid users. In this scenario, a limited amount of noise reduction of about 6 dB was obtained by the BMWF system in the optimal case (i.e. with perfect VAD), as can be seen in Figure 7. Furthermore, the setting λ=0.8 reduced the SNR improvement by 1 dB. It could be argued that this reduction is not necessary since in a diffuse noise environment no directional localization cues for the noise are available. In the present study, it was assumed that the hearing aid user does not adjust the λ setting according to the acoustical environment, but in principle it should be possible that this adjustment is made in the hearing aid according to the acoustical environment with the sound classifiers installed in modern hearing aids.

When using the envelope-based VAD, the performance is not degraded by more than 1 dB down to an input SNR of about -5 dB compared to the optimal case. At this point (for β=0.3), the correct classification of speech was about 78% and the correct classification of noise was about 50% (see Figure 3). Thus, it is not necessary for the BMWF system that the VAD shows satisfactory performance (i.e. a low error rate), but rather that the error rate is not excessive (e.g. higher than 50%), and therefore only small effects of VAD are observed in relatively adverse conditions. It should be noted, that even a small weighted SNR improvement of 3-6 dB found for diffuse babble noise can lead crucial speech recognition increase, if the improvement is found at SNRs comparable to the SRT. In Wagener & Brand (2005), for example, sentence intelligibility in different types of noise for hearing impaired listeners was investigated. The average SRTs for speech shaped noise and fluctuating noise were -3.3 dB and -2.1 dB, respectively, with improvements in speech recognition of 16 and 11 percent for each 1 dB increase in SNR. This means that for a typical hearing

impaired individual the SNR range of understanding almost nothing to understanding almost everything is -7 to 3 dB for sentences in fluctuating noise. In much of this SNR range (down to -5 dB), the BMWF performance does not degrade much due to VAD errors and an SNR improvement of 5-6 dB is found. Hence, the BMWF with envelope-based VAD might provide a significant improvement in speech recognition of more than 50%.

In very adverse conditions, e.g. at -10 dB SNR, which may also be encountered in the environment, the SNR improvement reduced to about 3 dB when using envelope-based VAD for noise estimation, which is comparable to that of a directional microphone. A first order directional microphone, consisting of two closely spaced microphones has an AI weighted directivity index as measured on KEMAR (which is equivalent to our measure of weighted SNR improvement in diffuse noise) of around 3 dB (*e.g.* Hamacher *et al*., 2005; Kates, 2008). Apart from that the performance of the present system is reduced to that of a two microphone system (in diffuse noise) for some of the noise levels that can be expected in a real environment, an obvious problem for this system arises if the interference is a single speaker or only a few speakers. In such situations, the temporal fluctuations of the noise interferer are very similar to the target fluctuations and thus, the VAD cannot discriminate between both. In consequence, no significant suppression of the interferers can be achieved.

The purpose of this work was primarily to investigate the effect of a realistic VAD on BMWF, more specifically, to identify the range of SNRs where the VAD has minimal effect on noise reduction performance compared to the case when VAD errors are not taken into account, and to quantify the degradation in performance for the conditions where the VAD has significant influence. Therefore, the following aspects can be subject to further research. The analysis presented has employed block processing where the statistics of speech and noise were calculated using the entire signal of 8 seconds of which about 2 seconds were noise. It is likely that head movement and movement of noise sources will degrade algorithm performance. In this context, the performance of the algorithm will not only be influenced by the type of adaptation used, but by the filters only being updated during speech pauses. Obviously, this impedes tracking of fast movement, as the filters can be frozen for seconds to the previous scenario. Also, VAD classification errors can lead to slower convergence of the filters. Due to the directional properties of the BMWF, this degradation is more likely to be significant in a simple (directional) noise source setup, than if the noise scenario is complex i.e. spatially diffuse.

Although it can be expected that an SNR improvement in frequency regions important for speech recognition would result in higher speech recognition, the gains obtained in intelligibility weighted SNR can only be related to the potential of this

system to improve intelligibility. This is particularly critical when individual hearing impairments (e.g. limitation in audibility, spectral resolution, or temporal fine structure processing) are considered. The effect of hearing impairment on speech intelligibility might be addressed by using modifications to the speech weighted SNR measure such as those proposed in (*e.g.* Pavlovic *et al.*, 1986) for the Articulation Index. However, in order to demonstrate the true benefit of the BMWF system in complex scenarios, speech intelligibility tests with hearing aid users need to be ultimately conducted. Also, the quality of the processed speech could be addressed.

5 OVERALL SUMMARY AND DISCUSSION

In this thesis, different aspects of spatial hearing and, in particular, externalization were investigated and discussed. Chapter 2 focused on externalization perception in relation to binaural cues that occur due to a combined effect of HRTF filtering and reverberation. Binaural room impulse responses were acquired using an artificial head at different positions in a room and it was shown that the dynamic level differences that arise in reverberant environments depend on the amount of reverberant energy relative to the direct sound energy. This dependence was shown to be related to changes in the width of the ILD distributions that were based on short-term ILDs collected over time. A method was developed that allowed for controlled modifications of short-term ILDs such as to reduce the size of their fluctuation and thereby the resulting width of the distribution. Psychoacoustic experiments were performed where the ILDs at each time instant were modified in such a way that the ILD distribution of the processed speech was reduced in width while the mean ILD was kept unchanged. It was found that the modifications of the dynamic ILDs affected the degree of externalization when the speech source contained high frequencies, i.e. for broadband and highpass filtered sound. The reduction in the degree of externalization was related to the extent of compression of the fluctuation size. However, in the case of lowpass filtered speech with a cut-off frequency of 1 kHz or lower, the modifications of the ILDs had no effect on the degree of externalization. An investigation of the bandwidth effects furthermore showed that unprocessed lowpass- and highpass- filtered speech with different cut-off frequencies was well externalized. This implies that the cues that contribute to externalization are about equally effective in both low and high frequency regions and for sounds with limited bandwidth. Informal listening tests indicated that externalization at low frequencies was not affected by compression of ITD fluctuation size.

Due to these findings, it was anticipated that monaural reverberation related cues could be responsible for the well externalized sounds at low frequencies. Larsen *et al.* (2008) found that the discrimination of changes in direct-to-reverberant sound energy was equally good in monaural and binaural listening mode and it was concluded that monaural cues, such as spectral variance and spectral envelope shape, were dominating D/R discrimination. Thus, since externalization perception is improved by reverberation and since the percept of reverberation is dominated by monaural cues, it could be assumed that monaural reverberation cues also strongly

affect externalization. However, the monaural presentation of sound via headphones did, indeed, not result in externalized sound images in this study. Still, this could have been the case due to the unnatural representation of HRTF cues used for localization in the direct sound in such a configuration. Thus, it could not be excluded that the monaural cues from the reverberant part of the sound contribute to externalization at low frequencies, whereas a binaural component might not be necessary.

In order to clarify these aspects, chapter 3 investigated the effect of using identical reverberation in both ears, while the direct sound was unmodified in order to preserve the binaural localization cues. It was found that the resulting degree of externalization was influenced by source laterality. A lateral source at 30 degrees was well externalized even though only monaural reverberation was present in the sound. The frontal source, however, was poorly externalized and required an increased amount of reflections containing binaural cues in order for the degree of externalization to increase. This was explained by differences in the dynamic binaural cues for the two source azimuths that arise from the interaction of the direct sound with the reverberant component. Since the frontal source contained only weak binaural cues in the direct sound, the addition of diotic reverberation resulted in small fluctuations of binaural cues. In the case of the lateral source, which contained stronger binaural cues in the direct sound, stronger fluctuations occurred, as the disparities in the binaural cues of the direct sound and the diotic reverberation became larger. The externalization ratings for the modified speech stimuli considered in chapter 3 corresponded well with the notion that the ILD distribution width was related to the degree of externalization. Thus, it was concluded that monaural reverberation cues are not sufficient for externalization and that a component related to binaural reverberation was essential for robust externalization. These findings do not contradict the results found in Larsen *et al.* (2008), as the task in their experiments was to identify the more reverberant sound. Thus, it appears that the cues used to perceive reverberation itself are different from the reverberation related cues involved in the perception of externalization.

Chapter 3 further focused on the effects of BRIR duration on the degree of externalization and dynamic binaural cues. When the BRIR duration was varied, it was found that, while full externalization occurred when about 80 ms of BRIR was included in the headphone auralization, the largest change in externalization occurred for BRIR durations between 20-50 ms. An analysis of the spatial cues, however, showed that large changes in the cues occurred for window durations within the first 20 ms after the direct sound, while a further increase in window duration resulted in relatively smaller changes. This discrepancy between the externalization degree ratings and the spatial cues calculated for the corresponding conditions was

explained by a suppression of reflections in a certain time window after the direct sound, related to the precedence effect (Litovsky *et al.*, 1999; Dizon & Colburn, 2006). The inclusion of a simple reflection suppression rule in the calculation of the binaural dynamic cues resulted in a dynamic binaural cue measure that better accounted for such aspects of the experimental data.

Furthermore, chapter 3 investigated the effects of the BRIR duration and the binaural cues from reflections for lowpass- and highpass filtered speech. The findings for the filtered speech were similar to those of broadband speech, suggesting that the mechanisms that are utilized in externalization perception are similar across frequency. Particularly, a binaural effect of reverberation was also found for frequencies below 1 kHz which is in contrast to the results from chapter 2 (Catic *et al.*, 2013) where no effect of dynamic ILDs and ITDs was found. Therefore, in chapter 3 of this thesis, the short-term interaural coherence (IC) was considered as an additional measure in the analysis of binaural cues. This measure was shown to be correlated with the externalization ratings in most conditions. Thus, it may be that the short-term IC is utilized by the auditory system as the reverberation-related binaural component in externalization perception. Furthermore, the dynamic ILD and IC cues were shown to be closely related, and there were indications that a combined evaluation of several binaural cues may play a role at low frequencies. In chapter 2 (Catic *et al.*, 2013), the compression of ILD fluctuation size at low frequencies was less effective than that at high frequencies, and also less effective than the compression of the ITDs at low frequencies. In addition, the binaural cues were modified separately. Therefore, a binaural component was still available in the case of stimuli with modified binaural cues and, thus, the findings in chapter 3 for sounds at low frequencies are not inconsistent with the results from chapter 2.

Since binaural cues affect externalization and spatial perception in general, it would be beneficial for the improvement of hearing aid user's sense of auditory space if these cues would be considered in the processing strategies for their hearing aids. Often, the algorithms developed to improve speech intelligibility lead to a distortion of binaural cues in the process, *e.g.* Van den Bogaert *et al.* (2006). In Klasen *et al.*, (2007), an algorithm that was designed to preserve binaural cues was proposed and subsequently evaluated in terms of localization performance (Van den Bogaert *et al.*, 2008) and speech intelligibility improvements (Van den Bogaert *et al.*, 2009). An improvement in localization performance was found in their study when compared to alternative processing strategies. However, all the evaluations were performed using an ideal error free voice activity detector (VADs). As the performance of a VAD degrades in adverse acoustic conditions, such as in the presence of a multi-talker babble interferer or in acoustic conditions with low signal-to-noise ratios (SNRs), an evaluation of the effect of a realistic VAD on algorithm

performance was performed in Chapter 4 (Catic *et al.*, 2010). It was found that, for a directional stationary noise interferer, the intelligibility-weighted SNR improvement was 20 dB, which decreased to 15 dB at low SNRs. Thereby, for stationary directional noise, the algorithm was found to provide large SNR improvements even in adverse conditions. In the case of the diffuse multi-talker babble noise, the improvement was found to be 6 dB, which degraded to 3 dB at low SNRs.

The ability of an algorithm to improve speech intelligibility should not only be evaluated in relation to the SNR improvement itself, but also the range of SNRs where this improvement occurs. Even a small improvement of 3-6 dB which was found for the nonstationary noise can lead to significant improvements in speech intelligibility if it occurs at SNRs close to the speech reception threshold (SRT). Hearing-impaired listeners often have SRTs at higher SNRs than normal-hearing listeners. As an example, Wagener & Brand (2005) found SRTs for hearing-impaired listeners at about -2 dB for fluctuating noise. Hence, it was concluded that, although the realistic VAD degraded the SNR improvement at low SNRs, there may still be a possible benefit for the hearing-impaired listeners since a large degradation did not occur in the region of their SRTs. Moreover, if a noise reduction algorithm would preserve the binaural cues, it may provide additional benefits for understanding speech in the presence of interferers in a spatial setting, due to spatial release from masking.

Overall, in this thesis, the reverberation-related binaural cues were investigated and found to be important for externalization. Externalization is a multi-dimensional percept which encompasses both distance and localization aspects, as it implies the identification of a specific sound source in space. Therefore, the process of evaluating the degree of externalization may be affected by several auditory cues. As an example, while it was found in this study that monaural reverberation cues were not sufficient to produce externalized sound images, it might still be that they are necessary in addition to the binaural cues. The weighting of different cues may also be flexible, such that the most reliable cues are selected in a given situation. The identification of the cues that contribute to externalization can be important for hearing aids or applications where the simulation of externalized sound sources is desired. Although individualized HRTFs yield the best performance in localization tasks, other spatial features, such as distance perception, are not affected by the use of non-individualized HRTFs.

Moreover, there are indications that, for certain spatial aspects, such as the perception of distance and the room acoustic environment, the perceptual effects of manipulating the properties of the reverberant part of the sound are much larger than the effects due to spatial rendering with non-individualized HRTFs (Zahorik, 2000; Zahorik, 2009). The knowledge of reverberation related properties that are essential

for achieving an externalized sound percept could therefore be useful for applications where exact HRTF measurements are not feasible. This could be simulations of externalized sound images without the need for the tedious measurements of individualized HRTFs, or for hearing aid applications. In hearing aids, the microphones are often placed behind the ear, which disturbs the HRTFs. While this is a static change which the listener can get acclimatized to, the use of adaptive signal processing changes the spatial cues adaptively according to each acoustic scenario, which can be difficult to get used to. Hence, an enhancement of the relevant binaural cues may be useful for a better externalization perception in such applications.

The method as described in Chapter 2 (Catic *et al.*, 2013) could be used for the enhancement of dynamic binaural cues, where a time dependent gain adaptation for the α parameter could be applied in order to avoid changes of HRTF cues at certain time instants, e.g. at the onsets of the sound where localization cues are important. In this thesis, an auditory-based approach was used employing the gammatone filterbank, but due to the shallow slopes of this filterbank and resulting issues with filter overlap, it is possible that a rectangular filterbank would be more appropriate in such an application.

The present study investigated externalization perception in normal-hearing listeners, where the available visual cues did not change and the listening scenarios were static. Future research may focus on the perception of externalization in hearing-impaired listeners, considering also non-static listening environments. Furthermore, audio-visual integration effects on externalization perception would be relevant and exciting to investigate.

BIBLIOGRAPHY

Alais, D., & Burr, D. (2004). "The ventriloquist effect results from near-optimal bimodal integration". *Curr. Biol.*, pp. 257–262.

ANSI S.3.5-1997. (1997). *American National Standard Methods for Calculation of the Speech Intelligibility Index.* Acoust. Soc. Am.

Ashmead, D. H., LeRoy, D., & Odom, R. D. (1990). "Perception of the relative distances of nearby sound sources". *Percept. Psychophys.*, pp. 326–331.

Begault, D. R., Wenzel, E. M., & Andersson, M. R. (2001, October). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". *J. Audio Eng. Soc.*(10), pp. 904–916.

Bitzer, J., Simmer, K. U., & Kammeyer, K. (1999). "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2965-2968.

Blauert, J. (1997). "Spatial hearing: The psychophysics of human sound localization". Cambridge, MA: MIT press.

Bradley, J. S., Sato, H., & Picard, M. (2003). "On the importance of early reflections for speech in rooms". *J. Acoust. Soc. Am.*, pp. 3233-3244.

Brandstein, M., & Ward, D. (2001). Microphone Arrays – Signal Processing Techniques and Applications. New York: Springer.

Bronkhorst, A. W., & Houtgast, T. (1999). "Auditory distance perception in rooms". *Nature*, pp. 517-520.

Bronkhorst, A. W., & Plomp, R. (1989, October). "Binaural speech intelligibility in noise for hearing-impaired listeners". *J. Acoust. Soc. Am*(no. 4), pp. 1374-1383.

Brungart, D. S., & Rabinowitz, W. M. (1999). "Auditory localization of nearby sources. I. Head-related transfer functions". *J. Acoust. Soc. Am.*, pp. 1465–1479.

Brungart, D. S., & Scott, K. R. (2001). "The effects of production and presentation level on the auditory distance perception of speech". *J. Acoust. Soc. Am.*, pp. 425–440.

Catic, J., Dau, T., Buchholz, J. M., & Gran, F. (2010). The Effect of a Voice Activity Detector on the Speech Enhancement Performance of the Binaural

Multichannel Wiener Filter. *EURASIP Journal on Audio, Speech, and Music Processing.* .

Catic, J., Santurette, S., Buchholz, J. M., Gran, F., & Dau, T. (2013). The effect of interaural level difference fluctuations on the Externalization of Sound. *Journal of the Acoustical Society of America*(2), pp. 1232-1241.

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

Coleman, P. D. (1962). “Failure to localize the source distance of an unfamiliar sound”. *J. Acoust. Soc. Am.*, pp. 345–346.

Cornelis, B., Doclo, S., Van den Bogaert, T., Wouters, J., & Moonen, M. (2010, Feb.). “Theoretical analysis of binaural multi-microphone noise reduction techinques”. *IEEE Transactions on Audio, Speech and Language Processing*(no. 2), pp. 342-355.

Dizon, R. M., & Colburn, S. H. (2006, May). "The influence of spectral, temporal, and interaural stimulus variations on the precedence effect". *J. Acoust. Soc. Am.*(5), pp. 2947-2964.

Doclo, S., & Moonen, M. (2002, Sep). “GSVD-based optimal filtering for single and multimicrophone speech enhancement”. *IEEE Transactions on Signal Processing, 50*(no. 9), pp. 2230-2244.

Festen, J. M., & Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing". *J. Acoust. Soc. Am.*, pp. 1725-1736.

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

Glasberg, B. R., & Moore, B. C. (1990). “Derivation of auditory filter shapes from notched-noise data". *Hear. Res.*, pp. 103-138.

Goupell, M. J., & Hartmann, W. M. (2006, June). "Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects". *J. Acoust. Soc. Am.*(6), pp. 3971-3986.

Goupell, M. J., & Hartmann, W. M. (2007, August). "Interaural fluctuations and the detection of interaural incoherence. III. Narrowband experiments and binaural models". *J. Acoust. Soc. Am.*(2), pp. 1029-1045.

Greenberg, J. E., Peterson, P. M., & Zurek, P. M. (1993, Nov). “Intelligibility-weighted measures of speech-to-interference ratio and speech system performance”. *J. Acoust. Soc. Am.*(no. 5), pp. 3009-3010.

Griesinger, D. (1997, July/ August). The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acta Acustica united with Acustica*(4), pp. 721-731.

Hamacher, & et al. (2005). “Signal Processing in High-End Hearing Aids: State of the art, Challenges, and Future Trends”. *EURASIP Journal on Applied Signal Processing*, pp. 2915-2929.

Hartmann, W. M., & Wittenberg, A. (1996). “On the externalization of sound images”. *J. Acoust. Soc. Am.*, pp. 3678–3688.

Hartung, K., & Trahiotis, C. (2001, September). "Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli". *J. Acoust. Soc. Am.*, pp. 1505-1513.

Hoffman, M. W., Trine, T. D., Buckley, K. M., & Van Tasell, D. J. (2009, August). “Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement”. *J. Acoust. Soc. Am*(no. 2), pp. 759-770.

Hohmann, V. (2002). “Frequency analysis and synthesis using a gammatone filterbank". *Acustica/Acta Acust.*, pp. 433–442.

Ihlefeld, A., & Shinn-Cunningham, B. G. (2011). “Effects of source spectrum on sound localization in an everyday reverberant room”. *J. Acoust. Soc. Am.*, pp. 324–333.

Kates, J. M. (2008). *"Digital Hearing Aids".* San Diego: Plural Publishing INC.

Keidsler, G., Rohrseitz, K., Dillon, H., Hamacher, V., Carter, L., Rass, U., & Convery, E. (2006). “The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers”. *International Journal of Audiology*, pp. 563-579.

Kirkeby, O., Nelson, P. A., Hamada, H., & Orduna-Bustamante, F. (1998). “Fast deconvolution of multichannel systems using regularization”. *IEEE T. Speech Audi. P. 6*, pp. 189–194.

Klasen, T. J., Moonen, M., Van den Bogaert, T., & Wouters, J. (2005). “Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction”. *ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing, 3*, pp. 29-32.

Klasen, T. J., Van den Bogaert, T., Moonen, M., & Wouters, J. (2007). "Binaural Noise Reduction Algorithms for Hearing Aids that Preserve Time Delay Cues". *IEEE Transactions on Signal Processing*(no 4), pp. 1579-1585.

Kopčo, N., & Shinn-Cunningham, B. G. (2011). “Effect of stimulus spectrum on distance perception for nearby sources”. *J. Acoust. Soc. Am.*, pp. 1530–1541.

Kulkarni, A., & Colburn, H. S. (1998). “Role of spectral detail in sound source localization”. *Nature*, pp. 747–749.

Larsen, E., Iyer, N., Lansing, C. R., & Feng, A. S. (2008). “On the minimum audible difference in direct-to-reverberant energy ratio”. *J. Acoust. Soc. Am.*, pp. 450–461.

Laugesen, S., & Schmidtke, T. (n.d.). “Improving on the speech-in-noise problem with wireless array technology”. *News from Oticon*.

Litovsky, R. Y., & Shinn-Cunningham, B. G. (2001, January). "Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression". *J. Acoust. Soc. Am.*, pp. 346-358.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). "The precedence effect". *J. Acoust. Soc. Am.*, pp. 1633-1654.

Litovsky, R. Y., Rakerd, B., Yin, T. C., & Hartmann, W. M. (1997, April). "Psychophysical and physiological evidence for a precedence effect in the median sagittal plane". *J. Neurophysiol.*, pp. 2223-2226.

Little, A. D., Mershon, D. H., & Cox, P. H. (1992). “Spectral content as a cue to perceived auditory distance". *Perception*, pp. 405–416.

Macpherson, E. A., & Sabin, A. T. (2007, June). Binaural weighting of monaural spectralcues for sound localization. *J. Acoust. Soc. Am.*(6), pp. 3677-3688.

Marzinzik, M., & Kollmeier, B. (2002, Feb). “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics”. *IEEE Transactions on Speech and Audio Processing*(no. 2), pp. 109-118.

Mershon, D. H., & King, L. E. (1975). “Intensity and reverberation as factors in auditory-perception of egocentric distance”. *Percept. Psychophys.*, pp. 409–415.

Møller, H., Sørensen, M. F., Hammershøi, D., & Jensen, C. B. (1995). "Head-related transfer functions of human subjects". *J. Audio Eng. Soc.*, pp. 300-321.

Moore, B. C. (2003). "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms". *Speech Communication*(no. 1), pp. 81-91.

Muller, S., & Massarani, P. (2001). “Transfer-function measurement with sweeps”. *J. Audio Eng. Soc.*, pp. 443–471.

Nábelek, A. K., & Robinette, L. (1978). "Influence of precedence effect on word odentification by normally hearing and hearing-impaired subjects". *J. Acoust. Soc. Am.*, pp. 187-194.

Nielsen, J. B., & Dau, T. (2009). “Development of a Danish speech intelligibility test”. *Int. J. Audiol.*, pp. 729–741.

Nielsen, S. H. (1993). “Auditory distance perception in different rooms”. *J. Audio Eng. Soc.*, pp. 755–770.

Nilsson, M., Soli, S., & Sullivan, J. (1994, Feb). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise". *J. Acoust. Soc. Am.*(no. 2), pp. 1085-1096.

Ohl, B., Laugesen, S., Buchholz, J., & Dau, T. (2010). “Externalization versus internalization of sound in normal-hearing and hearing-impaired listeners”. *Fortschritte der Akustik – DAGA 2010*, pp. 633–634.

Patterson, R. D., & Moore, B. C. (1986). “Auditory filters and excitation patterns as representations of frequency resolution”. In *Frequency Selectivity in Hearing* (pp. 123–177). London: Academic Press.

Pavlovic, C., Studebaker, G., & Sherbecoe, R. (1986, July). “An Articulation Index Based Procedure for Predicting the Speech Recognition Performance of Hearing-Impaired Individuals”. *J. Acoust. Soc. Am*(no. 1), pp. 50-57.

Pellegrini, R. S. (2002). *A Virtual Reference Listening Room as an Application of Auditory Virtual Environments.* Berlin: dissertation.de.

Peterson, P. M., Wei, S. M., & Rabinowitz, W. M. (1990). “Robustness of an adaptive beamforming method for hearing aids,”. *Acta. Otolaryngol*, pp. 85-90.

Plomp, R. (1978, Feb). “Auditory handicap of hearing impairment and the limited benefit of hearing aids”. *J. Acoust. Soc. Am.*(no. 2), pp. 533-547.

Ricketts, T. A. (2005, July/August). “Directional Hearing Aids: Then and now”. *Journal of Rehabilitation Research and Development*(no. 4), pp. 133-144.

Sanvad, J. (1999). Auditory perception of reverberant surroundings. *J. Acoust. Soc. Am.*(2), p. 1193.

Shinn-Cunningham, B. G., Kopčo, N., & Martin, T. J. (2005). “Localizing nearby sources in a classroom: Binaural room impulse responses”. *J. Acoust. Soc. Am.*, pp. 3100–3115.

Shinn-Cunningham, B. G., Santarelli, S., & Kopčo, N. (2000). “Distance perception of nearby sources in reverberant and anechoic listening conditions: Binaural vs. monaural cues”. *Assoc. Res. Otolaryngol. Abs.*

Van den Bogaert, T., Doclo, S., Moonen, M., & Wouters. (2008, Jul.). ”The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids”. *J. Acoust. Soc. Am*(no. 1), pp. 484-497.

Van den Bogaert, T., Doclo, S., Moonen, M., & Wouters, J. (2009, Jan.). ”Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids”. *J. Acoust. Soc. Am*(no. 1), pp. 360-371.

Van den Bogaert, T., Klasen, T., Van Deun, L., Wouters, J., & Moore, M. (2006, Jan). "Horizontal Localization with Bilateral Hearing Aids: Without is Better than With". *J. Acoust. Soc. Am.*(no 1), pp. 515-526.

Vary, P., & Martin, R. (2006). "Digital Speech Transmission – Enhancement, Coding and Error Concealment". Wiley.

Wagener, K. C., & Brand, T. (2005). “Sentence Intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement

procedures and masking parameters". *International Journal of Audiology*(no.3), pp. 144-156.

Wiggins, M. A., & Seeber, B. U. (2012). "Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners". *Ear Hear.*, pp. 399–410.

Zahorik, P. (2000). "Distance localization using non-individualized head-related transfer functions". *J. Acoust. Soc. Am.*, p. 2597.

Zahorik, P. (2001). "Estimating sound source distance with and without vision". *Optometry Vision Sci.*, pp. 270–275.

Zahorik, P. (2002, November). "Direct-to-reverberant energy ratio sensitivity". *J.Acoust.Soc.Am*(5), pp. 2110-2117.

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

Zahorik, P. (2009, August). "Perceptually relevant parameters for virtual listening simulations of small room acoustics". *J. Acoust. Soc. Am.*(2), pp. 776-791.

Zahorik, P., & Wightman, F. L. (2001). "Loudness constancy with varying sound source distance". *Nature Neurosci.*, pp. 78–83.

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

CONTRIBUTIONS TO HEARING RESEARCH

Vol. 1: Gilles Pigasse, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: Olaf Strelcyk, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: Eric R. Thompson, Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: Tobias Piechowiak, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: Jens Bo Nielsen, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: Helen Connor, Hearing aid amplification at soft input levels, 2010.

Vol. 7: Morten Løve Jepsen, Modeling auditory processing and speech perception in hearing impaired listeners, 2010.

Vol. 8: Sarah Verhulst, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: Sylvain Favrot, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: Sébastien Santurette, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: Iris Arweiler, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: Filip Munch Rønne, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: Claus Forup Corlin Jespersgaard, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: Rémi Decorsière, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: Søren Jørgensen, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: Simon Krogholt Christiansen, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 17: Márton Marschall, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 18: Jasmina Catic, Human sound externalization in reverberant environments, 2014.