Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 29

Jens Cubick

Investigating distance perception, externalization and speech intelligibility in complex acoustic environments

Hearing Systems
Department of Electrical Engineering

Investigating distance perception, externalization and speech intelligibility in complex acoustic environments

PhD thesis by
Jens Cubick

Preliminary version: June 2, 2017

Technical University of Denmark

2017

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by the Centre for Applied Hearing Research (CAHR) consortium with Oticon, WIDEX, and GN ReSound (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Prof. Torsten Dau
Assist. Prof. Sébastien Santurette
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Dr Søren Laugesen
Interacoustics Research Unit
Kgs. Lyngby, Denmark

Abstract

Spatial hearing, together with other sensory input, greatly helps us humans to build up an internal model of the world that surrounds us. While our sense of vision is limited to the frontal hemisphere, our sense of hearing allows us to perceive and localize sounds from all directions. This ability is essential both in situations where we need to react to our surroundings (e.g., in traffic situations) and for speech communication in the presence of several talkers (e.g., in a crowded environment).

Complete localization of a sound source not only requires an estimate of the direction of a sound source, but also of its distance. This can be accomplished based on the acoustic properties of the signals at the two ears. The main cues that are considered crucial for this estimation are the sound pressure, the energy ratio between the direct and the reverberant sound, and the spectral content of the stimuli. However, the perception of distance is not only determined by the acoustic properties of the stimuli. This thesis investigated whether the perceived distance of a sound source depends on the room in which the experiments are performed. It also investigated whether the playback room has an influence on the externalization of sound images, i.e., the perception of sounds outside the head, when signals recorded at the ears of the listeners are presented through headphones. Furthermore, the thesis analyses whether this influence is due to a mismatch between the acoustic properties of the recordings and the playback room or between the recordings and the visual impression of the room.

Even though the room in which experiments are conducted can affect auditory perception, it is still desirable to run experiments in the laboratory for reasons of control, repeatability, and convenience. In recent years there has been a demand for using realistic acoustic scenarios with high ecological validity for listening experiments in the laboratory, especially for testing "aided" hearing, i.e., the effects of hearing-aid signal processing on perception. An experiment is described that aimed to validate a loudspeaker-based room auralization system, which allows for the generation of complex acoustic scenes inside the laboratory. It was tested how well both acoustic measures and the results of speech intelligibility experiments match the results from the corresponding physical room that was the basis for the room simulation in the lab. Finally, the influence of hearing aids on the spatial perception of a listening scenario with spatially

separated sound sources was studied. In particular, it was investigated, whether a distorted spatial perception could explain degraded performance in a speech intelligibility task.

Overall, the results described in this thesis provide new insights into the processing and perception of spatial sounds in realistic acoustic environments and could be valuable for applications related to sound reproduction techniques and signal processing strategies in hearing instruments.

Resumé

Rumlig hørelse hjælper sammen med andre sensoriske input os mennesker med at opbygge en indre model af den verden der omgiver os. Mens vores synsans er begrænset til området foran os gør vores hørelse os i stand til at lokalisere lyd fra alle retninger. Denne evne er både væsentlig i situationer hvor vi har behov for at reagere på vores omgivelser (f.eks. i trafikken) og for at kunne kommunikere via tale, når flere taler på samme tid (f.eks ved større forsamlinger)

For fuldstændigt at kunne lokalisere en lydkilde er det nødvendigt både at kunne estimere lydkildens retning og afstanden til den. Et sådan estimat er baseret på signalernes akustiske egenskaber som de optræder ved de to ører. De vigtigste cues for dette estimat er lydtryk, forholdet mellem energien fra den direkte lyd og dens efterklang samt lydens frekvensindhold. Dog er opfattelsen af afstand ikke kun bestemt af lydens akustiske egenskaber. Denne afhandling undersøgte om opfattelsen af afstand til en lydkilde afhænger af lokalet, hvor eksperimenterne foregår. Afhandlingen undersøgte desuden hvorvidt lokalet hvor lyden bliver afspillet har betydning for eksternaliseringen af lydbilledet, dvs. hvorvidt lyden opfattes som værende udenfor hovedet, når signalet optaget ved lytterens øre præsenteres via høretelefoner. Derudover, analyserer afhandlingen om denne effekt skyldes et misforhold mellem optagelsen og afspilningslokalets akustiske egenskaber eller et misforhold mellem optagelserne og det visuelle indtryk af lokalet.

Selvom lokalet hvor forsøget foretages kan påvirke den auditoriske opfattelse, er det stadig ønskeligt at lave sådanne forsøg i laboratoriet for at opnå bedre kontrol og mere reproducerbare resultater og for bekvemmelighed. I de seneste år har der været efterspørgsel på realistiske akustiske scenarier til lytteeksperimenter i laboratoriet, især til at teste, hvordan høreapparaters signalbehandling påvirker opfattelsen af lyd. Afhandlingen indeholder en beskrivelse af et eksperiment, der har til mål at validere et højtaler-baseret system til rumliggørelse af lyd, der gør det muligt at generere komplekse akustiske scener i et laboratorie. Dette eksperiment undersøgte, hvor godt akustiske mål og resultater fra et taleforståelseseksperiment matchede resultater målt i lokalet som laboratoriesimulationerne var baseret på. Til slut undersøgtes der, hvordan høreapparater påvirker rummelig opfattelse i en lyttesituation med rummeligt adskilte lydkilder. Mere specifikt blev det undersøgt om en forvrænget rumopfattelse kunne forklare nedsat taleforståelse.

Generelt set, giver resultaterne beskrevet i denne afhandling ny indsigt i processering og opfattelse af rummelig lyd i realistiske akustiske miljøer og konsekvenser for tekniske anvendelsesmuligheder relateret til lydreproduktions teknikker og signalbehandlingsstrategier i høreapparater

Acknowledgements

Four years of PhD project are over, a long period of time that involved about 15,000 km by bike, a lot of learning, challenges, new experiences, and development, both work-related and in private. Looking back, two projects are particularly vivid in my memory. I had the opportunity to help shaping the new laboratory facilities of the Hearing Systems Group, in particular the Audiovisual Immersion Lab. This is certainly an experience that I will not forget. Getting the chance to dream up an optimum solution for a new state-of-the-art research facility – and then realizing it, was very rewarding, and the result is certainly something to be proud of. The other highlight was my external research stay at the National Acoustic Laboratories in Sydney, Australia. Both getting to work with a different group and new colleagues, and getting a glimpse of this incredible continent and its friendly people was a great experience.

I want to thank my main supervisor Torsten Dau for giving me this opportunity and for his trust, advice, and support. Thanks also to my supervisors Sébastien Santurette and Søren Laugesen for the good discussions we had. I also want to thank my colleagues at NAL for making me feel welcome, and of course my colleagues in the Hearing Systems and the Acoustic Technology groups, especially of course my dear office mates. I would not have spent almost eight years in and around building 352, if it were not for you guys.

I also want to thank my parents and my sisters, who never stopped believing in me and who always supported me in my decisions, even if that meant that I would be living far away and that visits were few and far between.

Finally, I want to thank my old friends back in Germany who have not forgotten me and all the the new ones I met in Denmark and Australia. Thanks for all the good times. Thank you Maria, my dearest and most unique friend, for continually challenging me, for helping me learn so much about myself and

about life, and for helping me to become much stronger along the way. What a journey! And last but not least, thank you Els for being such an amazing person, for caring for me, and for being such a huge support despite living half the world away.

Related publications

Journal papers

- Cubick, J., and Dau, T. (**2017**). "Distance perception and externalization," Article submitted to Acta Acustica united with Acustica.
- Cubick, J., Buchholz, J.M., Best, V., Lavandier, M., Dau, T. (**2017**). "Spatial perception and speech intelligibility with hearing aids," Article submission in preparation to The Journal of the Acoustical Society of America.
- Gil Carvajal, J. C., Cubick, J., Santurette, S., and Dau, T. (**2016**). "Spatial Hearing with Incongruent Visual or Auditory Room Cues," Scientific Reports **6**, Article number: 37342
- Cubick, J., and Dau, T. (**2016**). "Validation of a Virtual Sound Environment System for Testing Hearing Aids," Acta Acustica united with Acustica **102**(3), 547-557
- Jørgensen, S., Cubick, J., and Dau, T. (**2015**). "Speech Intelligibility Evaluation for Mobile Phones," Acta Acustica united with Acustica **101**(5), 1016–1025

Conference papers

- Cubick, J., Sánchez Rodríguez, C., Song, W., MacDonald, E. N. (**2015**). "Comparison of binaural microphones for externalization of sounds," Proceedings of 3rd International Conference on Spatial Audio 2015, Graz, Austria.
- Cubick, J., Santurette, S., Laugesen, S., Dau, T. (**2015**). "The influence of visual cues on auditory distance perception," Fortschritte der Akustik - DAGA 2015, Nuremberg, Germany.

- Cubick, J., Santurette, S., Laugesen, S., Dau, T. (**2014**). "Influence of high-frequency audibility on the perceived distance of sounds," Proceedings of 7th Forum Acusticum, Krakow, Poland.
- Cubick, J., Santurette, S., Laugesen, S., Dau, T. (**2014**). "Influence of High-Frequency Audibility on Distance Perception," Fortschritte der Akustik - DAGA 2014, Oldenburg, Germany.
- Cubick, J., Favrot, S., Minnaar, P., Dau, T.(**2013**). "Validation of a Virtual Sound Environment System for Hearing Aid Testing," AIA-DAGA International Conference on Acoustics 2013, Merano, Italy.

Published abstracts

- Cubick, J., Dau, T. **(2016)** "Objective and Perceptual Evaluation of a Virtual Sound Environment System" Poster session presented at DAGA 2016, Aachen, Germany.
- Gil Carvajal, J.C., Santurette, S., Cubick, J., Dau, T. **(2016)** "The Influence of Visual Cues on Sound Externalization" Poster session presented at 39th midwinter meeting of Association of Research in Otolaryngology, San Diego, CA, United States.
- Gil Carvajal, J.C., Santurette, S., Cubick, J., Dau, T. **(2015)** "Effects of incongruent auditory and visual room-related cues on sound externalization" Poster session presented at Tenth anniversary symposium of the international laboratory for Brain, Music, and Sound Research, Montréal, Canada.

Contents

1

General introduction

Hearing is in several ways the most important of the human sensory modalities. Unlike vision, we are not able to shut it down at will, therefore hearing is our primary alarm sense. Hearing a loud and unfamiliar noise will easily wake us up and alert us in the middle of the night. Unlike our sense of vision, hearing is not limited to the frontal hemisphere, but we can hear sounds from all directions. From an evolutionary point of view, the alarm function of hearing is especially important, because being alerted in time by sounds made by, e.g., a predator allows for more time to escape or fight back and thus, eventually, might decide about life and death. In the same vein, the ability to determine where a sound is coming from is equally important. The higher the acuity of localization, the better the chances of looking or running in the right direction. Even nowadays where we rarely run away from predators, spatial hearing greatly helps us in creating a working model of the world around us, e.g., when we can hear an approaching car, even though we are not looking in its direction.

Hearing is also essential for human communication. Most of our direct interaction with other humans relies on speech communication. But this speech communication is often disturbed. Noise in the environment, other talkers, and reverberation all make speech harder to understand. Also here, spatial hearing plays an important role because it greatly facilitates the process of segregating the noisy mixture of different voices and environmental sounds that we encounter every day into separate streams that we can attend to more easily.

Sec. 1.1 gives a brief introduction to the mechanisms and cues involved in spatial hearing. Sec. 1.2 focuses on speech intelligibility and ways to measure it. Finally, Sec. 1.3 provides an overview of the main chapters of this thesis.

1.1 Spatial hearing

1.1.1 Localization

Much research has focussed on the localization of sounds in terms of the azimuth angle and in terms of elevation. For localization in azimuth, the hearing system mostly relies on the comparison of the incoming sound signals from the two ears. If a sound source is, for example, positioned to the left of a listener, the sound will arrive at the left ear before the right ear due to the difference in distance that the sound needs to travel. This difference in arrival time is called the interaural time difference (ITD). The sensitivity of the hearing system to these differences is remarkable. Mills (1958) found the minimum audible angle difference for frontal directions to be about 1° for pure tone pulses. The minimum detectable ITD was found to be as small as 10 μs (Moore, 2003). For pure tones, this time difference between the ears translates to a phase difference. This phase difference works well for low frequencies, whereas it becomes ambiguous towards higher frequencies, where the wavelength of the signal is in the same range or shorter than the path difference between the ears.

Another difference between the acoustic signals at the two ears is that the sound pressure level of the signal at the right ear will be lower, because the direct path to the farther ear is obstructed by the head. At low frequencies, the sound will be diffracted around the listener's head and not be significantly attenuated. At high frequencies, on the other hand, the size of the head becomes large compared to the wavelength of the sound and the presence of the head in the sound field leads to a significant attenuation of the acoustic signal at the far ear. Here, the interaural level difference (ILD) can become as large as about 20 dB (Moore, 2003).

Listeners are particularly sensitive to ITDs at low frequencies, and when low frequency information is available, it seems to dominate the localization percept (Wightman and Kistler, 1992). ILDs serve as a localization cue mainly at high frequencies, where the nervous auditory system is too slow to 'phase-lock' to the fast fluctuations of pure tones, and where the phase relation of pure tones between the ears becomes ambiguous. This distinction has been known for a long time and was termed the duplex theory by Lord Rayleigh (Strutt, 1907; Moore, 2003; Plack, 2005).

While ITDs and ILDs allow for localization in the horizontal plane (Blauert, 1997), they cannot account for localization in the median plane, i.e., for sources that are equally far from the two ears. It was found that localization in this case mostly relies on spectral coloration, caused by the shape of the pinnae and the resulting reflections and interference patterns that create a filtering effect that is highly dependent on the incidence angle of the sounds (e.g., Butler and Belendiuk, 1977). The resulting patterns seem to be learned and it has been shown that listeners can even adapt if the patterns are changed. Hofman et al. (1998) asked listeners to wear ear-moulds that changed the shape of their pinnae. Initial experiments showed that they could not reliably estimate sound source elevation when wearing the moulds. However, after wearing the moulds for several weeks, the localization performance with the moulds was almost as good as without, a remarkable example for plasticity in the brain.

Head-related transfer functions

In reality, the signals at the two ears contain more information about the direction of the sound source than can be explained by constant ITDs and ILDs. Interference of the incoming sound waves with reflections on the torso and in the pinnae generates a complex sound with characteristic resonances and cancellations that depend on the incidence angle of the sound. If the transmission path from the sound source to the ear is assumed to be a linear and time-invariant system, it is completely described by its impulse response or, equivalently, by the corresponding transfer function in the frequency domain. A measurement of this spectral shaping in a free-field environment, e.g., an anechoic chamber, is called a head-related transfer function (HRTF). Each set of HRTFs for the two ears uniquely describes a certain direction in the space around the listener. HRTFs can be seen as the combined description of ITDs and ILDs for all frequencies.

However, most of the environments humans commonly encounter are not anechoic. In such environments, the transfer path between a sound source and a listener's ears not only contains the direct sound generated by the source, but also a mixture of reflections from the surrounding surfaces. If a transfer function is measured in such environments, it is usually referred to as a binaural room transfer function, or, more commonly, its time-domain equivalent, the

binaural room impulse response (BRIR). HRTFs and BRIRs can be used in binaural technology to simulate real-world listening situations through headphones, a topic that has gained much attention in recent years with the dramatically increased popularity of headphones that came with portable media players and smart phones. When a set of BRIRs is convolved with an anechoic sound signal, the result can lead to a very convincing simulation of a real-life sound source via headphones that can be virtually indistinguishable from a real sound source at the simulated position.

1.1.2 Distance perception

The estimation of the azimuth and elevation angle of sound incidence is important, but in order to fully determine the location of a sound source, also an estimate of its distance is needed. Even though there is less literature on auditory distance perception than on the estimation of direction, distance perception has been studied for a long time (Thompson, 1882) and for distances from few centimetres (e.g.,Brungart et al., 1999; Kopčo and Shinn-Cunningham, 2011; Parseihian et al., 2014) to hundreds of metres (Fluitt et al., 2014).

Auditory cues for distance perception

Various reviews are available that describe the main cues that are utilized for estimating the distance of a sound event (e.g., Coleman, 1963; Zahorik et al., 2005; Kolarik et al., 2015). The first cue that has been associated with distance perception is the intensity of a sound (Thompson, 1882). In a free field environment, the sound pressure generated by a monopole sound source in the far field obeys the inverse square law, i.e., the sound pressure level decreases by 6 dB when the distance to the sound source is doubled. Therefore, softer sounds are usually perceived farther away than louder sounds. However, this usually requires either familiarity with the source or a comparative judgement.

The second cue for distance perception is the direct-to-reverberant sound energy ratio (D/R). The D/R describes the ratio between the energy of the direct sound and the energy of the sound that arrives at the listener's ears after it has been reflected on surfaces, e.g., the walls of a room. Whereas the direct sound level decays with increasing distance from the sound source, the reverberant sound field in a room is often assumed to be more or less diffuse, i.e., to

have constant energy throughout the room. Therefore, the D/R decreases with increasing distance from the sound source in a room. von Békésy (1938) experimented with a setup that allowed him to vary the level of a microphone that picked up the direct sound of a loudspeaker and a microphone that recorded the reverberation generated by the same loudspeaker in a reverberant room, and hence the D/R. However, even though von Békésy (1938) described some influence of the mixing ratio between the two signals on the perceived distance, he was not convinced that this setup actually generated a true impression of a sound source being farther from the listener. Mershon and King (1975) showed that D/R indeed provides a salient cue for distance perception and that this quantity, in contrast to the intensity cue, can serve as an "absolute cue" for distance. Akeroyd et al. (2007) showed that NH listeners could reliably judge whether a sound source was farther or closer than a reference sound source when the level differences were compensated for. HI listeners, in contrast, performed at chance level, indicating that D/R may not be a reliable cue for them.

The third cue for auditory distance perception is the spectrum of the signal. Most notably, the level of the high frequencies decreases with increasing distance, due to absorption in the air. This effect only becomes noticeable for very large distances. In rooms, the natural reverberation changes the spectrum of the sound with increasing distance between the sound source and the listener. This is mostly based on two effects. First, most materials have a larger absorption coefficient at high frequencies than at low frequencies, such that with each reflection on a surface the high-frequency sound energy will be reduced. Second, repeated reflection also means that, eventually, the travel distance of the sound waves will be long enough for air absorption to become noticeable. Thus, D/R and a loss of high frequencies in the sound co-vary with distance. The spectrum of the sound at both ears also changes at very close distances between the sound source and the listener. Due to near-field effects, the low-frequency energy content at the ear facing the source grows disproportionately and generates an additional difference cue between the two ears which might aid distance perception (Brungart et al., 1999).

1.1.3 Externalization

The term externalization describes the fact that humans usually perceive the sound emitted by a sound source to be outside their heads (i.e., externalized).

However, in some listening conditions, this externalized percept breaks down and sounds are perceived inside the head (i.e., internalized). This internalized perception is most commonly experienced in headphone listening but has also been reported when sounds are presented through loudspeakers in an anechoic chamber (Toole, 1970), or when a listener with a head-tracker is placed within a loudspeaker ring and the signals are panned between the loudspeakers according to the information from the head-tracker, such that the direction of the sound stays constantly in front of the listener (Brimijoin et al., 2013). Over the years, there has been much discussion about which cues are responsible for internalization (see Blauert, 1997 for a summary). Laws (1973) found that the fraction of stimuli that were perceived as externalized despite headphone presentation could be greatly increased when a filter circuit was inserted into the reproduction chain that equalized the average difference of the frequency response at the listeners' ears between the headphone and the loudspeaker reproduction. Nowadays, it is typically assumed that a natural listening experience with externalized auditory images and correct localization in space can be achieved when the signals at the ears of a listener with headphones are identical to what they would be in the corresponding real listening situation (Møller, 1992; Blauert, 1997; Hammershøi and Møller, 2005).

Importance of reverberation

It has frequently been observed that the presence of reverberation helps with externalizing sounds. Sakamoto et al. (1976) described a condition, where adding reverberation (by combining dummy head recordings of a loudspeaker from an anechoic chamber and from a reverberation chamber) resulted in externalization of the sounds, whereas the anechoic recording was perceived as internalized. Also the availability of "true" binaural information seems crucial for externalization. Monaural presentation of the stimuli always resulted in internalized images in Catic et al., 2013. Catic et al. (2015) demonstrated that the first 80 ms of individual, but truncated BRIRs were required for complete externalization, consistent with findings from an earlier study (Begault et al., 2001). When Catic et al. (2015) replaced the late part of the BRIR beyond the truncation point by a diotic version of the impulse response, the results were unchanged for frontal sources. For a source at 30° azimuth, about 20 ms of the binaural part of the impulse response were sufficient to achieve full externalization in this case.

Importance of head movements

Even though internalization often occurred with early recordings from dummy-head microphones, it has been reported that listeners could externalize the signals if the dummy head was moved in unison with the listener's head. More recently, Loomis et al. (1999) described that listeners perfectly externalized sounds when wearing a pair of highly sound insulating headphones that reproduced the signals of microphones mounted outside the ear cups in real time. Apparently, the correct reproduction of the ITD and ILD changes with head movements was sufficient to allow for externalization despite the lack of any pinna cues. Finally, Brimijoin et al. (2013) demonstrated that when head movements were permitted, keeping the position of the sound source constant with respect to the listener's head systematically increased the percentage of externalized stimuli compared to normal (un-tracked) headphone presentation of binaural signals.

Externalization in HI listeners

Recently, some evidence was provided that hearing-impaired (HI) listeners perceive externalization differently compared to normal-hearing (NH) listeners. Ohl (2009) and Ohl et al. (2010) varied the amount of head-related binaural information available to listeners during headphone reproduction and found that most HI listeners, on average, were less sensitive to changes in the amount of binaural information with respect to externalization. Boyd et al. (2012) found that HI listeners, on average, externalized sounds less than NH listeners in a condition with full BRIR cues. However, when the stimuli were lowpass-filtered at 6.5 kHz, the externalization performance of NH listeners dropped to that of the HI listeners, indicating that the reduced externalization of the HI listeners might be explained by their reduced sensitivity to high frequencies. In a condition where the head-related cues were removed, NH listeners mostly internalized the stimuli. In contrast, the HI listeners perceived the sounds to be further out in the room. The two conditions suggested that HI listeners have a narrower dynamic range of the externalization percept than NH listeners.

1.2 Speech Intelligibility

1.2.1 Masking of speech

NH listeners usually have no problems understanding speech in quiet. However, quiet conditions have become rare in our everyday lives, and we are often surrounded by, e.g., traffic noise, music from a radio or other people talking. In these conditions, understanding speech becomes substantially more difficult, because the background noise can mask the speech. Speech intelligibility usually depends on the signal-to-noise ratio (SNR), i.e., on the difference in sound level between the target speech and the masking noise. However, the spectrum of the masker relative to the target speech is also important. It is usually observed that a masker is more efficient at masking the target speech the more similar its spectrum is to that of the speech. Furthermore, also the temporal structure of the sounds is important. In general, stationary noise is a more effective masker than fluctuating noise, because fluctuating noise allows the listener to "listen in the dips". It has also been observed that masking not only occurs in the time- and frequency domain, but that also the envelope fluctuations inherent in the target speech can be subject to masking. This explains why understanding becomes more challenging in reverberant environments (Houtgast et al., 1980). Here, the dips in the speech signal are partly 'filled' by the reverberant energy in the room, i.e., the modulation depth of the target signal is reduced. Also the modulations inherent in a masker signal can have a detrimental effect on the perception of the modulation that is present in the envelope of the speech signal. This has led to the concept of the SNR in the modulation domain (e.g., Dau et al., 1999 for modulation detection, Jørgensen and Dau, 2011 for speech) which has proven to be a powerful predictor for speech intelligibility in various conditions.

1.2.2 Spatial release from masking

All masking effects that have been described above can be observed when the target and the interferer are presented to the same ear or when presented diotically. However, it is also well-known that it is much easier to understand a target speaker in a listening situation with noise maskers or interfering talkers, if the interferers are spatially separated from the target speech in terms of their azimuth angle (Plomp, 1976; Hawley et al., 2004) or distance (Westermann

and Buchholz, 2015b). This effect has been referred to as spatial release from masking (SRM). Contributors to this effect have been found in terms of the "better ear effect", where listeners can simply focus on the ear with the better SNR, as well as in terms of "true" binaural processing, often referred to as binaural unmasking. Binaural unmasking can be considered as a de-noising operation in the central auditory system and has been modelled as an "equalization-cancellation" process (Durlach, 1963). For more than one interferer it has been found that proximity of the target and the nearest interferer is more limiting for the intelligibility of the target signal than the number of interferers. (Hawley et al., 1999; Lőcsei et al., 2016).

1.2.3 Informational masking

It has also often been argued that, apart from the signal-inherent masking described here, there might be a more cognitive component to masking, often referred to as informational masking (IM, see Kidd et al., 2008). There has been much controversy around IM, but the concept is based on the observation that when stimuli with high information content are used, especially speech, the amount of masking observed is higher than in the case of stimuli with lower information content, such as stationary noise. The amount of masking observed with low-information stimuli has commonly been referred to as energetic masking (EM). The difference between the total amount of masking observed with high-information stimuli, and EM, has been considered as IM.

1.2.4 Measurement methods

To measure speech intelligibility, mainly two approaches have been considered. The first approach relies on the repeated presentation of the target speech and the noise at different, previously defined SNRs. This method is referred to as the method of constant stimuli. Here, the outcome measure is the percentage of correctly understood speech tokens for a certain SNR. The advantage of this method is that it not only provides information about intelligibility at the tested SNRs but also allows for the estimation of the slope of the underlying psychometric function. The second approach represents adaptive methods that vary the level of the target speech or the masker depending on the result of the previous presentation according to a pre-defined tracking rule. The outcome of such experiments is the speech reception threshold (SRT), the SNR at which a

listener understands a certain percentage, often 50%, of the presented speech tokens, words or sentences.

1.3 Overview of the thesis

This thesis broadly addresses two topics related to the human perception of sound: speech intelligibility and spatial hearing. While the two following chapters are concerned with speech intelligibility in spatial settings, chapters 4–6 deal with topics related to basic aspects of spatial hearing, i.e., distance perception and externalization. In the following, a brief overview of the individual chapters is provided.

Chapter 2 presents a study that aims to validate a loudspeaker-based virtual sound environment system for hearing research. In order to evaluate the performance of the acoustic simulation, speech intelligibility was measured in a real classroom and in its simulated counterpart inside a loudspeaker array. NH listeners were tested with and without hearing aids, and with omnidirectional and directional microphone processing. Additionally, the room simulation was evaluated using the room acoustic parameters reverberation time, clarity, and interaural cross-correlation coefficient. Finally, the directivity of the hearing aids with omnidirectional and directional processing was measured inside the real room, the simulated room, and in an anechoic chamber.

Chapter 3 presents a study that investigates speech intelligibility in NH listeners with HAs in a similar spatial setting as in chapter 2. This study focused on the question whether worse speech understanding of NH listeners with HAs than without might be due to a degraded spatial perception of the scene, which might make the separation of the target speech from the spatially distributed interferers more difficult. Speech intelligibility was measured with NH listeners with and without hearing-aids in a setting with target speech from the frontal direction and three interferers. The interferers were either collocated with the target speech or spatially distributed around the listener, and were either interfering talkers, or stationary speech-shaped noises with the same frequency content. The spatial perception of the listeners was tested by asking them to sketch their perception of the different sound sources, in particular with respect to their spatial distribution and their width.

Chapter 4 focuses on a more basic aspect of spatial hearing and describes two experiments on auditory distance perception. Literature on externalization suggests that externalization is reduced when binaural stimuli, presented via headphones, are lowpass-filtered. In the first experiment described here, it was investigated whether the same influence could be found in a distance perception experiment with binaural stimuli presented via headphones. The second experiment evaluated the influence of different playback rooms. The second experiment included a subset of the original listeners, and was performed in a listening booth instead of the workshop room where the BRIRs had been recorded.

Chapter 5 presents a study that compared five different commercially available binaural microphones and the built-in microphones of a head-and-torso simulator in terms of externalization. Furthermore, it compared the amount of externalization achieved with individual and with generic BRIRs, i.e., BRIRs measured on a dummy head.

Chapter 6 presents a study that investigated whether the playback room has an influence on on the externalization of binaural signals presented via headphones, as often reported anecdotally. Individual BRIRs were measured in a listening room for all (blindfolded) listeners, who were then asked to rate the distance, angle, and compactness of the auditory image during binaural presentation via headphones. Much care was taken to be able to disentangle the influence of auditory cues from the playback room and the visual impression of the playback room.

Chapter 7 summarizes the main findings of this work, and discusses their implications, and gives a perspective on potential future research within the fields of distance perception, externalization, and speech intelligibility in realistic environments.

Finally, **Appendix A** describes earlier work on the intelligibility of speech transmitted through mobile phones. In the mobile phone industry and -research, speech intelligibility is not commonly considered. The more commonly considered outcome is speech quality. This seems counterintuitive for a device originally designed for speech communication. In a black-box approach, this

study investigated speech intelligibility with stimuli recorded through three different commercially available mobile phones and a reference microphone. Apart from the listening experiments, it was also tested how well three different well-established speech intelligibility models were able to predict the measured outcomes.

2

Validation of a Virtual Sound Environment System for Testing Hearing Aids[a]

Abstract

In the development process of modern hearing aids, test scenarios that reproduce natural acoustic scenes have become increasingly important in recent years for the evaluation of new signal processing algorithms. To achieve high ecological validity, such scenarios should include components like reverberation, background noise, and multiple interfering talkers. Loudspeaker-based sound field reproduction techniques, such as higher-order Ambisonics, allow for the simulation of such complex sound environments and can be used for realistic listening experiments with hearing aids. However, to successfully employ such systems, it is crucial to know how experimental results from a virtual environment translate to the corresponding real environment. In this study, speech reception thresholds (SRTs) were measured with normal-hearing listeners wearing hearing aids, both in a real room and in a simulation of that room auralized via a spherical array of 29 loudspeakers, using either Ambisonics or a nearest loudspeaker method. The benefit from a static beamforming algorithm was considered in comparison to a hearing aid setting with omnidirectional microphones. The measured SRTs were about 2-4 dB higher, and the benefit from the beamformer setting was, on average, about 1.5 dB smaller in the virtual room than in the real room. These differences resulted from a more diffuse sound field in the virtual room as indicated by differences in measured directivity patterns for the hearing aids and

[a] This chapter is based on Cubick and Dau, 2016.

interaural cross-correlation coefficients. Overall, the considered VSE system may represent a valuable tool for testing the effects of hearing-aid signal processing on physical and behavioural outcome measures in realistic acoustic environments.

2.1 Introduction

Hearing aid (HA) users often have difficulties following a conversation in challenging listening situations that involve multiple talkers, background noise and/or reverberation (Bronkhorst, 2000), even though they typically benefit from their HAs in simple acoustic situations, such as a one-to-one conversation in a quiet room. The processing power of HAs has increased dramatically over the last 10 years and advanced signal processing strategies have been applied to help the users, particularly in complex listening situations. To assess and evaluate the performance of modern HAs, the test scenarios should therefore be as realistic as possible. Until recently, however, most testing has been done either in very basic conditions with simple loudspeaker setups in acoustically dampened rooms, or in field studies where the end users wear certain types of HAs for some time and report back via questionnaires after the testing period. The first approach offers much control over the test conditions but provides only very limited flexibility regarding the acoustic conditions and does therefore not reflect the challenges that HA users face in their everyday life. In field tests, representing the second approach, the participants experience the HAs in the environments where they would actually use them but the experimental conditions are difficult to control. The simulation of realistic acoustic scenes under controlled and repeatable conditions in the laboratory would combine the advantages of the two approaches.

One well-known method to provide such simulated scenes are headphone-based reproduction systems that use binaural technology (Møller, 1992) to reproduce the correct sound pressure at the listeners' ear. However, even though the results obtained with this method can be very convincing, headphone-based systems have some disadvantages. The simulation is most convincing if it is based on head-related transfer functions that are measured for each listener individually, and if head tracking is used to keep the auditory image position stable, even if the listener moves his/her head. Measuring impulse responses

for all incidence angles requires an enormous measuring effort and makes testing difficult. Furthermore, using HAs under headphones is impractical, as the acoustics under earphone cups are very different from a free field situation. These problems can be avoided with loudspeaker-based technologies that try to reproduce a desired sound field in a room. Sound field reproduction techniques, like wave-field synthesis (Berkhout et al., 1993), higher-order Ambisonics (HOA; Gerzon, 1973; Daniel et al., 2003; Vorländer and Summers, 2008), directional audio coding (Pulkki, 2007), or direct mapping of reflections to the nearest loudspeaker (Seeber et al., 2010), make it possible to render realistic and reproducible virtual sound environments (VSEs) in the laboratory, including room reverberation and multiple sound sources. In the case of HOA, the system aims at reproducing the sound field correctly at the listener's location in the virtual room around the "sweet spot" in the centre of the loudspeaker array. The presence of the listener thus ideally generates exactly the same acoustic effects as it would in the real sound field. Head rotations are allowed and, unlike in headphone-based systems, listeners are able to wear HAs in a VSE. In a HOA-based system, however, the spatial resolution of the reproduced sound field is limited by the Ambisonics order which, in turn, depends on the number of loudspeakers in the array (Daniel et al., 2003).

Such a HOA-based system has been realized at the Technical University of Denmark (DTU). It comprises a spherical array of 29 loudspeakers mounted in an acoustically highly dampened room (see Figure 2.1). The VSEs are based on simulations using the room acoustic modelling software ODEON (Christensen, 2013). A 3-dimensional model of a room is generated and the absorption and scattering properties of all surfaces are defined, as well as all source positions and the receiver position and direction. Even though such a geometrical acoustics-based simulation has limitations, especially in the low frequencies and with small rooms, it is very easy to model very well-defined complex listening scenarios. The simulation results are then processed by the loudspeaker-based room auralization toolbox (LoRA; Favrot and Buchholz, 2010). Using either HOA or a method where each reflection is mapped to the nearest loudspeaker (NLS), a multi-channel room impulse response is generated, which, when convolved with an anechoic source signal, yields the driving signal for the loudspeakers. Several studies have been conducted to evaluate the performance of this system. One study compared the common room acoustic

Figure 2.1: Photograph of the 'Spacelab' at DTU. A spherical array of 29 loudspeakers allows for the auralization of acoustical scenes in virtual rooms. Photo: Joachim Rode.

parameters, defined in EN ISO 3382-1, 2009 and derived from the LoRA output, with the corresponding values provided by the underlying ODEON simulation (Favrot and Buchholz, 2010). Considering different seats in a classroom and a concert hall, it was found that the variation of the room acoustic parameters for small head movements was mostly within 1-2 difference limens (Cox et al., 1993; Bradley et al., 1999) of the ODEON results. In another study (Favrot and Buchholz, 2009b), speech intelligibility in noise was measured for different rendering methods. The highest speech intelligibility was found when NLS coding was used, whereas it was lower in the case of 4^{th}-order HOA and even lower in the case of 1^{st}-order Ambisonics. In a third study (Favrot and Buchholz, 2009a), distance perception in the VSE was investigated and no significant difference was found between the LoRA system and a test based on binaural recordings. A study with a technically comparable auralization system at the HA manufacturer Oticon (Minnaar et al., 2011) compared speech intelligibility and listening effort of hearing-impaired listeners in different virtual rooms, a 'dry' room, a lecture hall, and a very reverberant basement. Another study, using a similar system, tested speech intelligibility in a 'complex' cafeteria environment with multiple talkers, and in a 'standard' anechoic environment (Best et al., 2015). Finally, two

very recent simulation studies investigated the applicability of multichannel loudspeaker-based reproduction chains for testing HAs (Grimm et al., 2015; Oreinos and Buchholz, 2015).

However, in all above studies, the VSE systems were evaluated either by comparing theoretical quantities, or room acoustical measures between the VSE and the underlying ODEON simulation, or by comparing results of behavioural measurements obtained inside the system. Only a few studies actually compared the listening performance measured in a VSE with the performance in the corresponding real environment. Few studies used simulation-based auralizations presented via headphones and compared speech intelligibility in this setup with the one measured in the real rooms, (e.g., Yang and Hodgson, 2007; Hodgson et al., 2008; Rychtáriková et al., 2011), or overall listening experience (Schoeffler et al., 2015). One early study compared speech intelligibility in a loudspeaker-based auralization system and in a real room using binaural technology (Kleiner, 1981), and, to the knowledge of the authors, only one study has compared perceptual measures obtained in a loudspeaker-based VSE directly to the same measures obtained in the corresponding real room (Koski et al., 2013). To successfully employ the system for HA testing, it is crucial to know how well experimental results from a VSE translate to real-life situations.

Specifically, the present study investigated whether the reproduction of a VSE in the LoRA-based system captures the acoustic properties of a 40-seat classroom accurately enough, such that the effects of HA processing in the VSE can be considered to be the same as, or very close to, the real environment. To achieve this goal, three requirements need to be fulfilled: (1) The ODEON simulation must be well calibrated to capture the key acoustical properties of the classroom. To assure this, the simulation results for the common room acoustic parameters reverberation time, T_{30}, and clarity for speech, C_{50}, (EN ISO 3382-1, 2009) from ODEON were compared to the values measured in the classroom; (2) The LoRA processing must be transparent to preserve these properties. To test the transparency of the LoRA processing, the same room acoustic parameters were calculated from room impulse responses measured inside the VSE, using either HOA or NLS rendering; and (3) The HA performance in the VSE and the real room needs to be comparable. To assess the HA performance, directivity patterns were measured both in the classroom and the VSE, using

omnidirectional microphones and a static beamforming (BF) program (Dillon, 2001).

If these requirements are fulfilled, the performance of the listeners in behavioural tasks in the VSE and the real room may be assumed to be comparable. To evaluate this, speech intelligibility was considered as an outcome measure in the present study since it represents one of the most important performance indicators in the HA development process. Speech reception thresholds (SRTs) were measured both in the classroom and its virtual counterpart with normal-hearing listeners, either with or without HAs. Testing normal-hearing listeners with HAs might seem counterintuitive but was chosen here as a first step in the evaluation process of the VSE system; normal-hearing listeners typically show more "homogeneous" results than hearing-impaired listeners and the main focus of the present study was to study the effect of basic features in the HA settings on the selected outcome measures in the real versus the simulated environments. The SRT benefit from a static BF algorithm relative to a HA setting with omnidirectional microphones was tested. This algorithm has been shown to yield a speech perception benefit of up to 8.5 dB in optimized conditions, when the test was performed in a sound-insulated booth with noise presented from 180° azimuth, (Valente et al., 1995), or up to 3.9 dB in more realistic scenarios with a noise source at 90° azimuth in a room with a reverberation time of 0.45 s (Wouters et al., 1999).

It was hypothesized in the present study that inaccuracies in the sound field reproduction should decrease the effectiveness of the BF and the associated gain in the effective signal-to-noise ratio (SNR) for frontal sources, which should result in higher SRTs. It was assumed that the room simulation can be considered sufficiently authentic if (1) the SRTs measured in the VSE are close to those obtained in the corresponding real room and if (2) threshold differences between the two HA settings are similar in the two situations.

2.2 Methods

2.2.1 Auralization technique

The acoustical data for the VSEs in the system under test were generated based on a room simulation in the commercial room acoustic simulation software ODEON (Christensen, 2013). This software uses a hybrid method for the calculation of the room acoustic parameters (Rindel, 2000; Rindel and Christensen, 2003). The image source and ray tracing methods (Vorländer and Summers, 2008) are combined to calculate the reflections up to a certain order. Above this transition order, the secondary source method is used to compute the late part of the room impulse response (RIR). The ODEON simulations in this study were run with 8000 early rays, 8000 late rays, a maximum reflection order of 2000, an impulse response resolution of 1 ms and a transition order of 3. The virtual sound sources were modelled to have the same directivity in the horizontal plane as that measured in an anechoic chamber for the Dynaudio BM6P loudspeaker used as the target source in the listening experiments. The simulation results, i.e., the reflectogram, containing information about the delay, direction and frequency content of each early reflection up to the transition order, and the energy decay curves, were exported from ODEON and processed by the LoRA toolbox (Favrot and Buchholz, 2010) to generate the driving signals for the loudspeaker array.

Due to the precedence effect (Blauert, 1997; Litovsky et al., 1999), the localization of a sound source in a room is mostly determined by the direct sound, whereas the late reflections in the rather diffuse reverberant tail of the RIR cannot be resolved individually (Buchholz et al., 2001). Following these properties of human sound localization, the LoRA toolbox splits the RIR into the direct sound, the early reflections, and the late reflections. The direct sound and the early reflections up to the transition order are rendered with the highest possible resolution, i.e., by either employing the highest possible HOA order for a given loudspeaker array, or by mapping it to the nearest loudspeaker available (NLS). The late reflections are provided by ODEON as the vectorial intensity and the envelope of the energy. These data are interpreted as a 1^{st} order Ambisonics signal and are decoded correspondingly. The resulting envelope for the late reflections is then multiplied with uncorrelated noise for each loudspeaker (Favrot and Buchholz, 2010). Summing up the parts of the decoded RIR generates a

multi-channel RIR, and convolution of this RIR with an anechoic signal forms the driving signal for the loudspeakers.

The VSE in the listening tests was played back through the spherical array of 29 Dynaudio BM6P loudspeakers in the 'Spacelab' shown in Figure 2.1. The array consists of a horizontal ring of 16 loudspeakers at ear height of a sitting listener at a distance of 1.8 m, two rings of 6 loudspeakers at ±45° elevation and one loudspeaker on the ceiling above the centre of the array. It is placed in an acoustically dampened room with a reverberation time of 0.16 s in the 125-Hz octave band and below 0.1 s in all frequency bands above 125 Hz. All loudspeakers were equalized to a flat frequency response relative to an omni-directional B&K 4092 microphone in the centre of the array using 1114-tap FIR filters. In the listening tests, 4^{th} order three-dimensional HOA rendering was used.

The room chosen for the VSE in this study was "Room 019", a lecture room at DTU with 40 seats and a volume of about 180 m^3. The ODEON model was carefully matched to the reverberation time and clarity values measured at the listening position shown in Figure 2.2 by assigning materials with appropriate absorption and scattering coefficients to the model surfaces. In addition to T_{30}, Clarity was considered an important criterion for the model calibration, because this early-to-late energy ratio is related to speech intelligibility (EN ISO 3382-1, 2009).

2.2.2 Physical evaluation

Room acoustic parameters

For the physical validation of the VSE, the common room acoustic parameters reverberation time, T_{30}, clarity for speech, C_{50}, and the interaural cross-correlation coefficient, $IACC$, were calculated according to EN ISO 3382-1 (2009) from RIRs measured with logarithmic sine sweeps (Müller and Massarani, 2001). This was done both in the classroom and the corresponding VSE. All impulse responses were measured both with an omni-directional measurement microphone B&K 4192 and a B&K 4100 head and torso simulator (HATS) at the listening position. Impulse responses were measured for 32 positions with the same Dynaudio BM-6P loudspeaker that was used as the speech target source in the listening experiments. For the evaluation, the results were averaged over

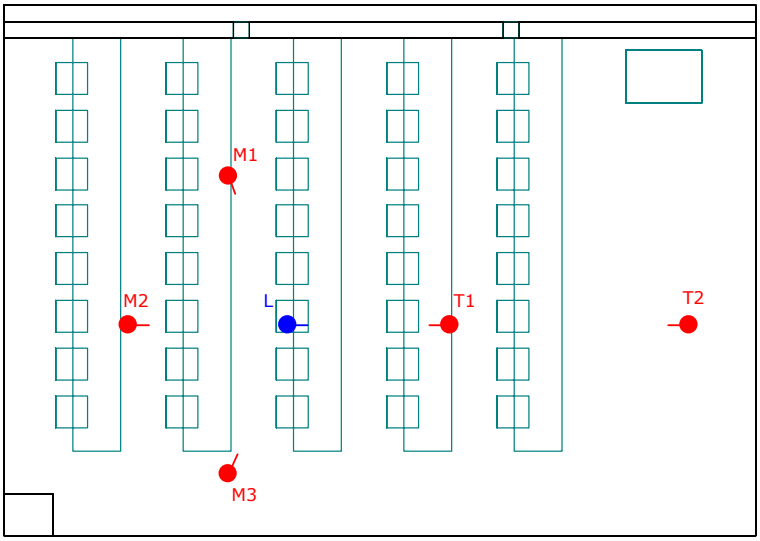


Figure 2.2: Top view of the room model with the listening position (L), the three maskers (M1, M2, M3), and the target speech sources T1 at 2 m and T2 at 5 m.

the 25 source positions for which the measurement distance was 2 m or larger.

Hearing aid directivity

Deviations of the auralized sound field from the original one were assumed to decrease the efficiency of the BF, which relies on the input from the two microphones, and, in turn, to decrease speech intelligibility. To assess the directional characteristics of the HAs, transfer functions were measured with the HA used in the test on the right ear of a B&K 4128 HATS. This was done for all incidence angles in steps of 10° at a distance of 2 m in an anechoic chamber, in the classroom, and in the VSE with each rendering method. All transfer functions were computed relative to the response of the HA in the omnidirectional program, measured on a B&K 4157 ear simulator with an outer-ear simulator DB 2012 for 0° incidence angle in an anechoic chamber. To reduce the strong magnitude fluctuations in the room transfer functions, their magnitude was smoothed with a 1/3-octave wide moving average filter.

2.2.3 Perceptual evaluation

Listeners

Eight normal-hearing native Danish speaking listeners (6 male, 2 female) with an average age of 27 years participated in the study and were paid an hourly wage. They were given written as well as oral information about the experiment and signed a consent form. The experiment was approved by the Danish Science-Ethics Committee (reference H-3-2013-004). The listeners were instructed in the use of the HAs as to changing the program and inserting or taking out the HAs after instruction. They were supplied with regular production receiver-in-the-ear Oticon Ino HAs providing a linear gain of 15 dB across the frequency range of the HA. In the HAs, an omnidirectional microphone and a static beamformer program could be selected. The HAs were coupled to the ears with mushroom-shaped silicone Oticon power domes, such that no individual earmoulds were needed. All adaptive features of the HAs, like noise reduction and feedback cancellation, were turned off.

Stimuli

SRTs were measured using the Danish Dantale II speech-in-noise test (Wagener et al., 2003), the Danish version of the Swedish Hagerman test (Hagerman, 1982). This speech corpus is a matrix test spoken by a female talker that consists of 160 five-word sentences with an identical syntax of “name + verb + numeral + adjective + object”. All sentences are permutations of the 50 words of a base list with 10 sentences, which makes the sentences hard to memorize and allows for reusing them within the same test session (Wagener and Brand, 2005). The masking noise was the corresponding Dantale II speech-shaped noise, produced from the test sentences that were superimposed with random pause durations for each sentence (Wagener et al., 2003). The target speech was embedded in clips of the noise file with a random start sample, such that the noise started 0.9 s before the sentence onset and ended 0.5 s after the end of the sentence. The on-and offset of the noise was windowed with 200 ms hanning ramps.

Experimental procedure

Before the actual measurements, the listeners were trained with 80 sentences, both with and without HAs and with both HA programs. The test conditions

were counterbalanced across all listeners and the sentence lists were randomized with the constraint that no list could be re-used within seven runs. For each test condition, the SRT, representing the SNR at which 50% of the words were understood correctly, was determined in an adaptive procedure using two lists, i.e., 20 sentences. The level of the speech-shaped noise was kept constant at 70 dB SPL in all unaided conditions, and 62 dB SPL in all HA conditions, resulting in roughly equal loudness across the two conditions. The speech level was adjusted using an adaptive maximum-likelihood procedure (Brand and Kollmeier, 2002). The test was conducted in the patient-based, closed-set version (Pedersen, 2007), where the listener had to choose the correct words from all possible alternatives in a Matlab-GUI on an iPad. The target speech source was placed at 0° at distances of 2 m and 5 m, respectively, as shown in Figure 2.2. Three noise sources were placed at angles of ±112.5° and 180° at a fixed distance of 2 m. All loudspeakers were placed with their acoustic centre at ear level, i.e., about 120 cm above the ground.

An overview over the test conditions can be found in Table 2.1. All listeners were tested in the classroom and in the VSE with both NLS and HOA rendering for the target distances of 2 m and 5 m. This was done without HAs as well as with the two HA programs. Half of the participants were first tested in the VSE, the other half of the participants was first tested in the classroom. During the SRT measurement, the listeners were asked to sketch the perceived position and extent of the sound sources in each experimental run on a response sheet with a schematic drawing of the listening test setup. The listeners were encouraged to describe any peculiarities they observed orally to the experimenter. Even though no formal evaluation was performed on these responses, the descriptions were expected to provide some hints regarding potential weaknesses of the auralization procedure or to allow for some exploration in the case of unexpected results. The experiments were divided into two sessions of about two hours.

Room	Distance	HA
R019	2 m	Unaided
VSE-NLS	5 m	Omni
VSE-HOA		BF

Table 2.1: Overview over listening test conditions. All listeners performed the experiments in all combinations of the listed conditions.

2.3 Results

2.3.1 Physical evaluation

Room acoustic parameters

Figure 2.3 shows T_{30} (left panel) and C_{50} (right panel) measured in the classroom (square symbols) and in the VSE using NLS (crosses) and HOA rendering (circles). The symbols indicate the average values measured at the listening position shown in Figure 2.2 for the 25 source positions with a minimum distance of 2 m. The average value of T_{30}, determined as the average of the values for the 500 Hz and 1 kHz octave bands according to EN ISO 3382-1 (2009), was 0.49 s in the classroom and 0.53 s in the VSE with both rendering methods. The values in the classroom varied between 0.48 s at 1 kHz and 0.6 s at 2 kHz and dropped to 0.44 s at 8 kHz. In the lowest two frequency bands, no meaningful values could be determined in the classroom due to distinct room modes. Considering the limited frequency range of hearing aids, these frequency bands were not considered crucial and the values were omitted in the figure. The ODEON simulation results for T_{30} were essentially identical with the ones measured in the VSE, and thus omitted in the figure for clarity. This indicates that the reverberation time is well-preserved by the LoRA processing and that the playback room does not provide additional reverberation, which is in good agreement with Favrot and Buchholz (2010), where similar measures were computed from the multichannel RIR. The values measured in the VSE differ from the ones in the classroom by less than 0.1 s. This deviation corresponds to the calibration error of the ODEON model. An even closer match between room model and reality would have required the use of materials that are highly absorbent in very narrow frequency bands, which would have compromised the plausibility of the room model.

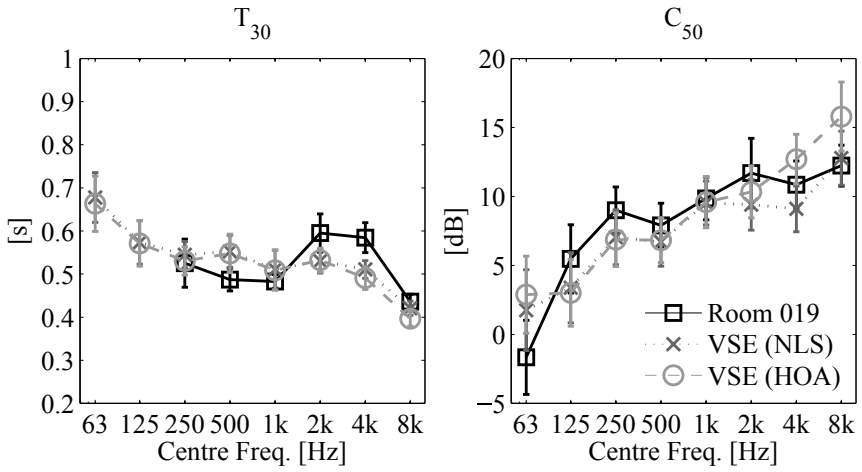


Figure 2.3: Average reverberation time T_{30} and clarity for speech C_{50} at the listening position for 25 source positions. The values were measured in the real classroom (square symbols) and in the VSE (crosses and circles).

Since the clarity for speech C_{50} represents the ratio of acoustic energy between the first 50 ms and the remaining part of the impulse response, it shows the opposite trend compared to the reverberation time. Apart from the two lowest frequency bands, the values ranged from 8 dB to 12.2 dB in the classroom. The values in the VSE tended to be slightly lower with a maximum deviation of 2.3 dB at 2 kHz. Bradley et al. (1999) argued that a just noticeable difference of 3 dB for clarity represents a realistic value in real listening situations. Thus, the match between the room acoustic simulation and the real room may be sufficient for a convincing auralization. However, in the 125 Hz frequency band, the values measured in the VSE are about 5 dB lower than the simulated values obtained with ODEON. This difference is most likely caused by the playback room, which is not fully anechoic and produces some reflections in this frequency band. At the highest two frequencies, the clarity values for the HOA rendering method are markedly higher than the ones for NLS. Favrot and Buchholz (2010) found a similar trend for the microphone position in the centre of the loudspeaker array. They explained this deviation by the energy regularization decoding method that is used in the frequency bands above the upper frequency limit imposed by the limited number of loudspeakers with HOA to preserve the total energy in the sweet spot.

Figure 2.4 shows the $IACC$ measured at the listening position in the classroom (square symbols), the VSE using NLS (crosses) and HOA coding (circles), for the two target source positions at 2 m (left panel) and 5 m (right panel) as a function of frequency. Two main trends can be observed: First, the $IACC$ for the 5-m target distance is lower than the corresponding value for the 2-m distance in nearly all room conditions. Second, in most cases, the IACC measured in the classroom is higher than in the VSE. Lower coherence values for larger distances were expected, because the sound field in a room becomes increasingly dominated by the reverberant sound with increasing distance. The lower values found in the VSE compared to the classroom may reflect the spatial 'jitter' introduced by the NLS technique and the imperfect reproduction of the sound field at the two ears with HOA coding. The pronounced dip in the curves at 500 Hz coincides with the decoupling frequency described by Lindevald and Benade (Lindevald and Benade, 1986). They stated that the spatial average of the correlation function between the two ear signals in a room is well described by a modified sinc function with the first zero at about 500 Hz, representing the decoupling frequency. Below this frequency, the signals at the two ears are highly correlated, whereas above it, the signals are essentially two independent samples of the sound field. Lower $IACC$ values in the VSE might indicate a more diffuse sound field than in the real room, which would make a BF algorithm less effective.

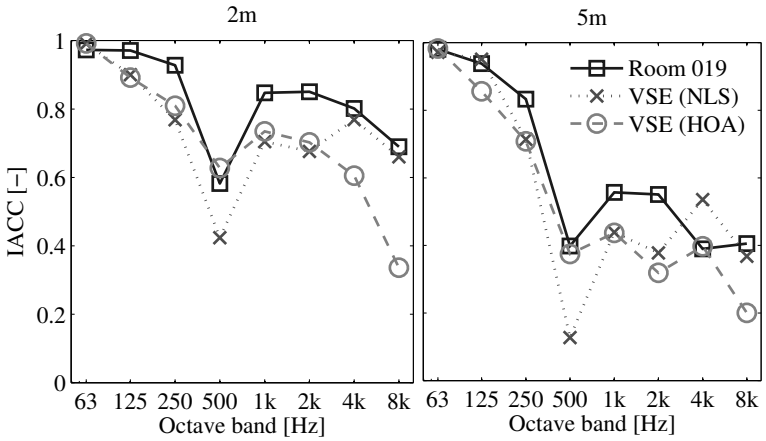


Figure 2.4: Interaural cross-correlation coefficient (IACC) measured in the real room (squares) and the VSE with NLS (crosses) and HOA rendering (circles) at a target source distance of 2 m (left panel) and 5 m (right panel).

Hearing aid directivity

Figure 2.5 shows the directivity patterns measured for the HA in the anechoic chamber (upper panels), Room 019 (middle panels), and the VSE with HOA rendering (bottom panels). The left column shows the directivity pattern for the omnidirectional program, the right column shows the pattern for the BF program. In the anechoic chamber (top row), the head shadow and the interference patterns on the contralateral side of the head are clearly visible as dark areas. In addition, the BF results clearly show the zeros of the BF at about −100° and +120°, especially at the lower frequencies up to about 2 kHz. In Room 019 (right middle panel), remainders of the pattern can still be found, but the dynamic range between the highest and the lowest sensitivity is strongly reduced. This was expected since, unlike in an anechoic chamber where all the sound energy arrives from the direction of the source, the sound that arrives at the HA in a room also contains reflected energy from the different surfaces, which makes the sound field more diffuse. Even if a zero in the BF sensitivity pattern would perfectly eliminate the direct sound, e.g., generated from a noise source in the room, the microphone would still pick up most of the reflected sound. Using HOA rendering of the VSE, the dynamic range is further reduced, especially when comparing the values for a given frequency across the different incidence angles, i.e., values lying on a horizontal line in the plots. The zeros at the low frequencies can hardly be observed anymore. This indicates that the sound field inside the VSE might be even more diffuse than the one in Room 019. The results for NLS coding are not shown here because they are very similar to the results obtained for HOA.

2.3.2 Speech intelligibility

Figure 2.6 shows the mean value and standard deviation of the measured SRTs for the conditions listed in Table 2.1, i.e., the three HA conditions ‘unaided’, ‘omni’ and ‘BF’ measured in the three room conditions ‘R019’, ‘VSE-NLS’ and ‘VSE-HOA’ for target source distances of 2 m and 5 m. For the target source distance of 2 m (black symbols), the SRTs for the unaided conditions were found at -13.8 dB in the real room (R019, left panel), -11.8 dB in the VSE with NLS coding (middle panel), and -9.4 dB with HOA coding (right panel). The higher SRTs obtained with HOA compared to NLS coding are consistent with findings reported in Favrot and Buchholz (2009b). Using HAs in the omnidirectional

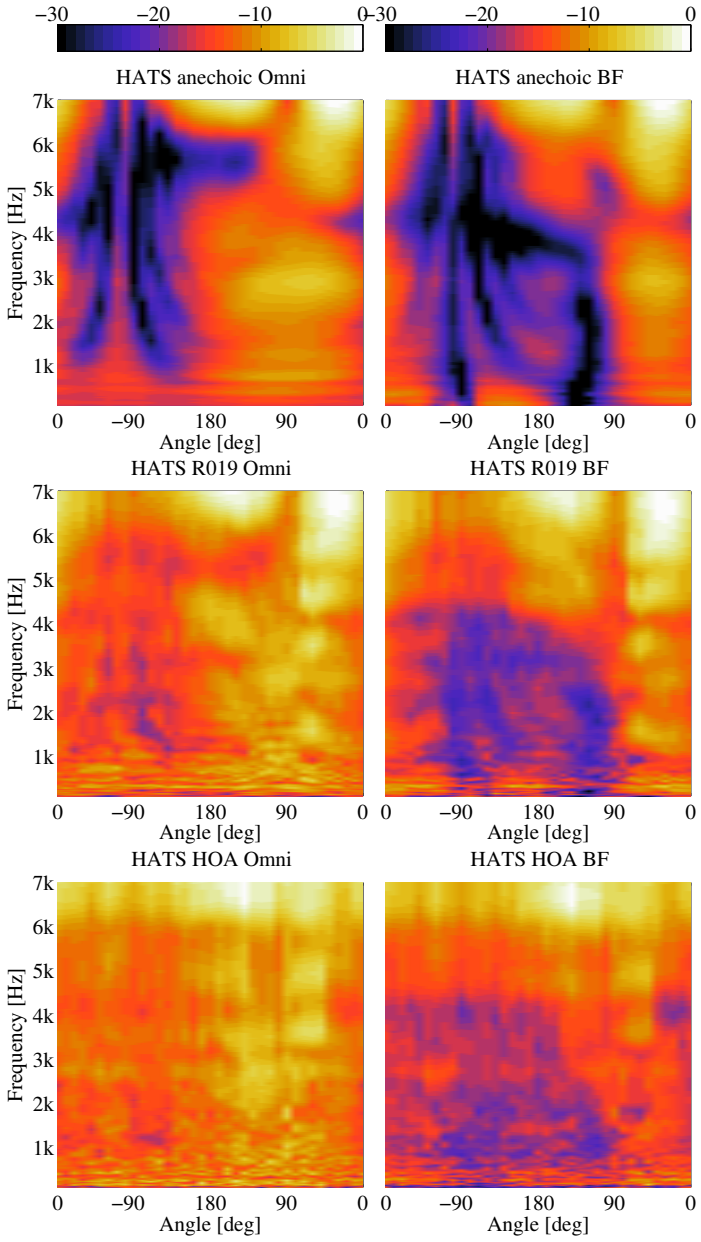


Figure 2.5: Directivity patterns of the HA measured on the right ear of a B&K HATS 4128 in an anechoic chamber (top row), the classroom (middle) and the VSE (bottom row). The left column shows the results for the omnidirectional program, the right column shows the results for the beamformer. All transfer functions are computed relative to the Omni-program for frontal (0°) incidence measured on an ear simulator B&K 4157 under anechoic conditions.

microphone setting generally increased the average SRT compared to the unaided condition by up to 4 dB in the real room, whereas using HAs in the BF program lowered it by up to 2.7 dB with HOA coding. For the target source distance of 5 m (grey symbols) in Room 019 (left panel), the listeners showed an increase in SRT of about 3 dB in all HA conditions compared to the results obtained at 2 m. This was expected since the direct-to-reverberant sound ratio in a room usually decreases with increasing distance, which is generally assumed to have an adverse effect on speech intelligibility (Bradley et al., 2003). Compared to the results for the 2-m distance, the SRTs measured for the 5-m distance showed a considerably larger spread in the real room. At this distance, small head movements subjectively had a larger effect on the SRT than at 2 m and some listeners might have utilized them more successfully than others. This might be due to wave phenomena like standing waves and local interference patterns. This would also explain, why this effect is not seen in the VSE, because the ODEON model is based on geometrical acoustics and hence cannot capture wave phenomena.

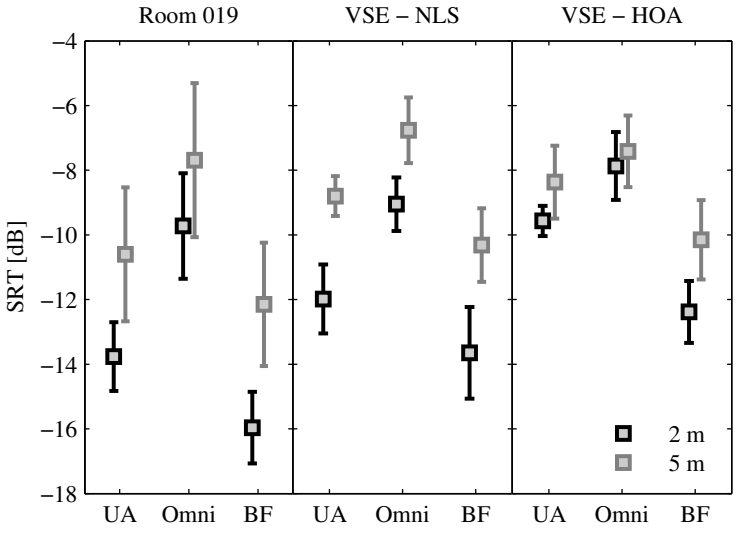


Figure 2.6: Average SRTs measured in Room 019, the VSE with NLS rendering and the VSE with HOA rendering for each of the HA conditions Unaided (UA), Omni, and Beamformer (BF), and for a distance of 2 m (black symbols) and 5 m (grey symbols). The error bars indicate ± one standard deviation.

For statistical analysis, a linear mixed model was fitted to the data with

'Room', 'Distance', and 'HA condition' as fixed factors and 'Listener' as random factor. In an Analysis of Variance (ANOVA), all factors and all two-factor interactions showed significant effects, indicating that there are differences between the results measured in the classroom and in the VSE. When only the data from Room 019 were considered, only the two main effects 'Distance' and 'HA condition' were significant, whereas their interaction was not. To address which VSE rendering method yields results that are more comparable to the real room, two ANOVAs were performed to compare the results of each rendering method to the ones measured in Room 019. In both cases, all main effects were highly significant, including the factor 'Room', which indicates that the measured SRTs measured in the room are different from the ones in the classroom. However, all two-factor interactions showed significant effects in the case of HOA rendering, but not in the case of NLS rendering ($\alpha = 0.05$). Especially the difference in SRT between the two distances with NLS (Figure 2.6, middle panel) was found to be similar as in Room 019 (left panel), whereas the pattern looks clearly different for HOA (right panel). This is reflected in a non-significant interaction between 'Room' and 'Distance' [F(1,79) = 0.1441, p = 0.7053] with NLS, whereas the same interaction was significant with HOA [F(1,79) = 9.9380, p = 0.0023]. This suggests that NLS coding preserves more of the cues that contribute to speech intelligibility, despite the simple algorithm, especially with respect to distance.

Since a VSE system will probably mostly be used to compare perceptual outcome measures in different conditions, the benefit in SRT from the BF over the omnidirectional program was computed as $SRT_{\mathrm{Omni}} - SRT_{\mathrm{BF}}$ (cf., Figure 2.7). In Room 019, this benefit was, on average, 6.2 dB for a target distance of 2 m, while it dropped to about 4.5 dB for the 5-m distance. The values measured in the VSE were found to be slightly lower in all cases. With NLS, the values dropped to 4.6 dB at 2 m distance, and to 3.5 dB for the 5-m distance. With HOA, the average benefit was 4.3 dB for the 2-m distance and 2.9 dB for the 5-m distance. An ANOVA on these benefits again showed significant main effects of the factors 'Room' and 'Distance', indicating that the BF benefit is not equal, but smaller in the VSE than in the real room, and decreases with increased distance. However, a set of one-sample t-tests showed that the mean value underlying the measured benefits was larger than zero in all conditions, indicating that the BF yielded a clear advantage in speech intelligibility relative to the omnidirectional

processing in all tested conditions.

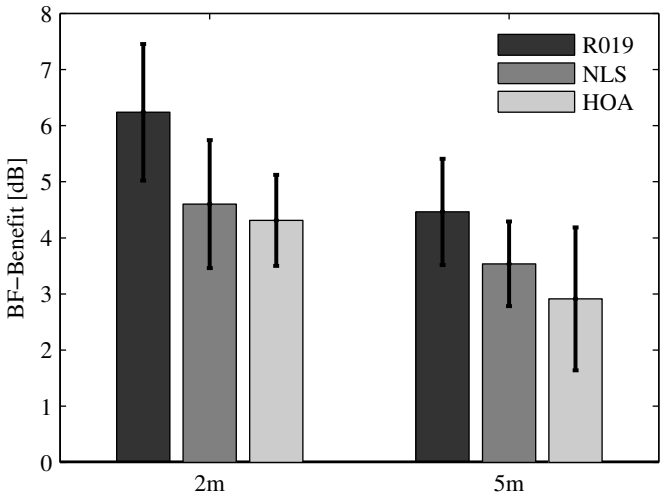


Figure 2.7: Benefit from the BF algorithm over the omnidirectional microphone pattern for all room conditions and the two target source distances. Higher values indicate better performance, the error bars indicate ± one standard deviation.

2.3.3 Subjective impression

In each run, the listeners were also asked to sketch their subjective impression of localization and extent of the sound sources in a schematic drawing of the listening situation with a listener and a circle indicating the radius of the loudspeakers. In the real room, the result tended to change from a very clear and focused image in the unaided case (see Figure 2.8a for an example) to a spatially much less defined image with HAs in the omnidirectional setting (Figure 2.8c). This impression may have resulted from the loss of the directional-dependent pinna cues due to the microphone position above the ear. Switching to the BF program, many of the listeners again reported a change in the spatial impression. Often, the sound sources were described as being closer around the head and sometimes the target speech was perceived inside the head, i.e., internalized (Figure 2.8e). Some listeners also reported hearing the noise source inside the head, while the speech was located outside. In the VSE, the virtual sound sources were often perceived as being wider and less well-defined than in the classroom (Figure 2.8b). Especially the three noise sources were often fused into

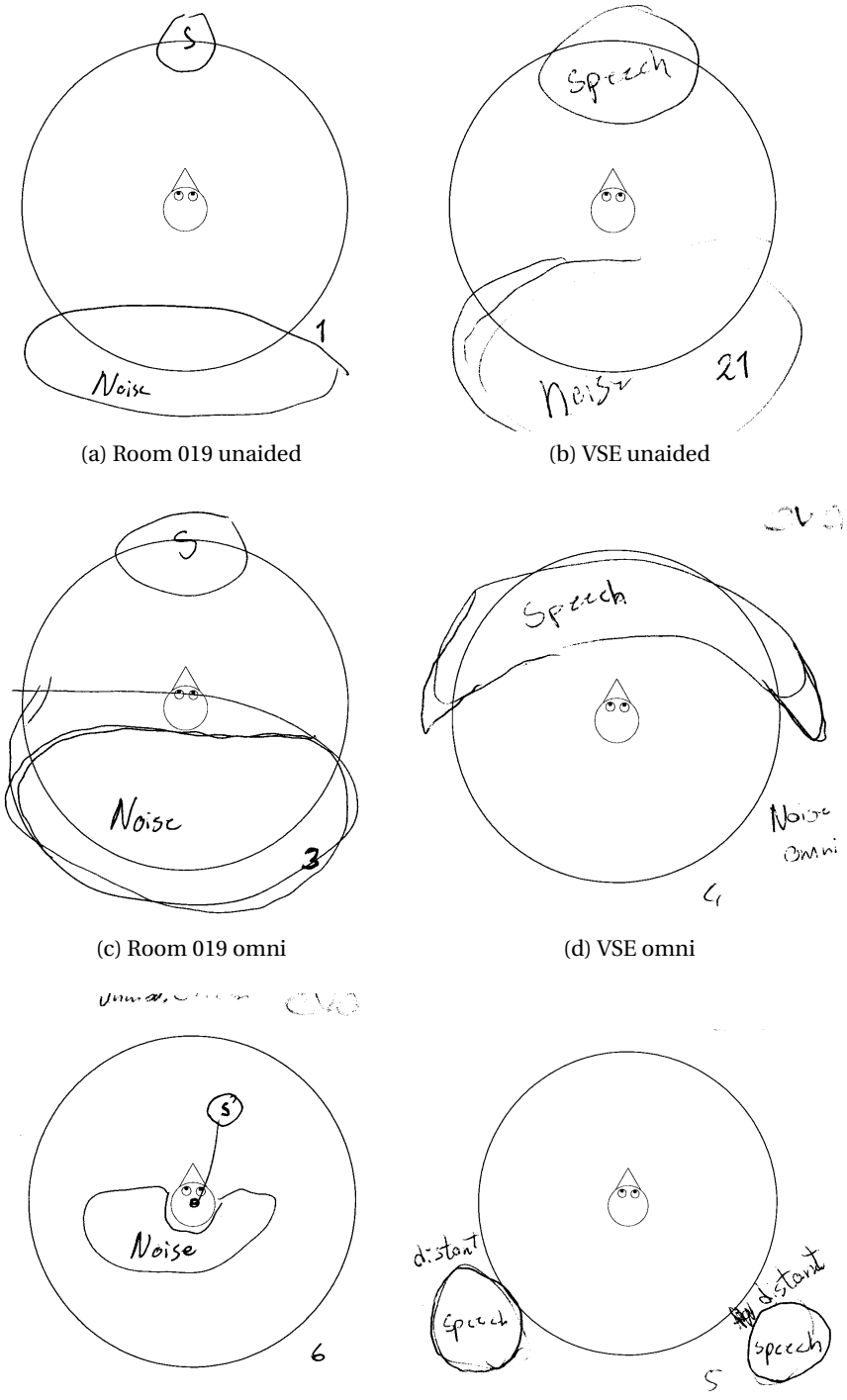
a single percept or the listeners reported that the noise was ‘somewhere behind’ them; some listeners described the speech as sounding more reverberant. The noise sources were perceived even wider when the HAs were used with the omnidirectional program. In this setting, many listeners perceived the noise as coming from all around the room. The speech source was often described as being much broader than in the classroom (Figure 2.8d). With the BF program, the descriptions became more diverse. Some listeners again reported the target speech to be closer to them or even inside their head, in some cases the sound image split and was indicated at different places (Figure 2.8f). The noise sources were often perceived at two separate locations, either close to the ears or at loudspeaker distance at the sides of the array. Even though there was a lot of variability in the subjective impression, it was clear that all conditions with hearing aids tended to distort the spatial perception of direction, source width, and distance. Interestingly, some listeners had the impression that they performed much worse in the BF than in the Omni conditions, even though their SRTs were actually consistently better.

Finally, some listeners reported that the transition from understanding the whole sentence to not understanding anything seemed less gradual in the VSE than in the classroom, which is reflected in the generally smaller variability in the data obtained in the VSE compared to the real room. This might indicate that the underlying psychometric function is actually steeper in the VSE than in the real room, which would imply that the sensitivity of the speech test is actually higher inside the VSE.

2.4 Discussion

2.4.1 Physical evaluation

The results from the physical measurements should provide some insights regarding the different limiting factors in the auralization chain: the ODEON simulation, the auralization system with the LoRA toolbox and the loudspeaker array, and the playback room. A room acoustic computer model can only provide a rough approximation of the actual sound field in a room. Inside such a model, the room geometry needs to be simplified and usually assumptions need to be made regarding the materials in the room and their acoustical properties.



(a) Room 019 unaided

(b) VSE unaided

(c) Room 019 omni

(d) VSE omni

(e) Room 019 cardioid

(f) VSE cardioid

Figure 2.8: Subjective evaluation of listening test conditions. The scans show the descriptions of one listener in Room 019 (left) and the VSE (right) for the Unaided condition (upper), the Omni program (middle) and the BF (bottom), respectively. In conditions (d) and (f) the listener indicated that the noise was perceived as coming from all directions.

Typically, room acoustic simulation programs are evaluated in terms of their prediction of room acoustic parameters (e.g., Bork, 2005). Here, the measured room acoustic parameters agreed well between the ODEON simulation and the real room. The values for T_{30} and C_{50} measured in the VSE agreed very well with the ODEON results, indicating that the temporal energy decay in the playback room closely follows the model and that the playback room is sufficiently dampened. Lower values for the $IACC$, however, indicated that there are differences in the spatial characteristics of the sound field between the real room and the VSE and that the sound field reproduced inside the loudspeaker array is more diffuse than the one in the classroom. This might, at least partly, account for the larger perceived spaciousness and reverberance. Another indication of a more diffuse sound field in the VSE is the reduced directivity obtained with the BF algorithm in the HAs. The main source of the increased diffuseness is probably the finite number of loudspeakers, which imposes the limitation of a spatial quantization with the NLS method and the requirement to truncate the HOA series after the 4^{th} order which, in turn, limits the spatial resolution of the system. However, the usual room acoustic parameters might not be sufficient to describe the performance of the room acoustic models and the input data for the auralization system might also be a limiting factor for the authenticity of the VSE.

2.4.2 Listening experiments

In general, the VSEs could reproduce the trends in the SRT variations found in the real room very well, even though the SRTs were generally shifted towards slightly higher levels, indicating poorer speech intelligibility in the VSE. This finding is not surprising, because each step in the generation of the VSE, i.e., the ODEON simulation, the LoRA toolbox, and the loudspeaker array and playback room, imposes some limitations on the overall result. Most geometrical room acoustic simulation methods are only appropriate when the dimensions of the room are long compared to the wavelength (Vorländer and Summers, 2008) and therefore not very reliable at frequencies below the Schroeder frequency (Schroeder and Kuttruff, 1962). Another aspect that potentially limits the performance of the auralization system is the rendering method. If HOA is used, the number of the loudspeakers limits the Ambisonics order which, in turn, limits the localization accuracy. It also implies an upper frequency limit for correct sound field reproduction. In the system under test, this frequency

limit is at about 2.2 kHz if a sweet spot of 20 cm diameter is considered (Favrot and Buchholz, 2010). Above this frequency, the magnitude of the sound is still correct, but the phase relations might be incorrect. If the NLS technique is used instead, these limitations do not apply. However, in this case, the sound source positions are limited to the angles at which loudspeakers are available and the reflections are subject to spatial discretization, which might also blur the perceived localization of the sound source. If the localization accuracy is reduced compared to the real room, it might become more difficult to segregate the target speech from the noise leading to a higher SRT. If the playback room is not sufficiently close to anechoic, the natural reverberation will increase the reverberation in the VSE and will add a sense of increased spaciousness. In the system under test, however, this was not considered an issue due to the very short reverberation time.

Another result from this study was that the SRTs measured with HOA tended to be higher than the ones obtained with NLS. This finding is consistent with the results of an earlier study (Favrot and Buchholz, 2009b) that found higher intelligibility scores with NLS than with 4^{th} order HOA which, in turn, were higher than the ones measured with 1st order Ambisonics. Differences between the SRTs measured with the two tested HA programs, however, could clearly be observed in all VSE conditions and they were similar to the ones measured in the classroom. This is an important finding since it demonstrates that the results measured in the realistic VSE seem to be a good indicator of real-world performance. Also for other differential measures, e.g., the comparison of the listening performance in several simulated rooms with different acoustical properties (Minnaar et al., 2011), the VSE seems to be well-suited.

Regarding the reports of the subjective impression of the perceived position and the extent of the sound sources, visual cues might have contributed to the result that the sound sources were usually perceived as wider in the VSE than in the classroom. The listeners were surrounded by 29 loudspeakers in the VSE, whereas there were only four single loudspeakers in the classroom. The role of potential visual cues in the evaluation cannot be clarified in the present study. However, in all experimental conditions, the sources in the VSE were simulated at angles at which there were loudspeakers in the array, which might have helped to consolidate the auditory image.

2.4.3 Perspectives

The auralizations in this study were based on room simulations. This approach has the major advantage that it makes the auralization method very flexible. Existing models can easily be adapted to new listening situations with, e.g., additional sound sources. Furthermore, it is possible to auralize rooms that do not physically exist (yet) or acoustic situations that do not occur in real rooms, but allow for the study of basic aspects of spatial hearing, e.g., the influence of single reflections on speech intelligibility (Arweiler and Buchholz, 2011). One limitation, however, is that while the method works well for static scenes, it is quite cumbersome to implement moving sound sources. Furthermore, the inherent limitations of ray-tracing based room acoustic models do not allow accurate reproduction of low-frequency effects, like room modes, and only roughly represent the acoustic properties of a room. Also fast fluctuations in the reverberant tail of the room impulse response are difficult to capture with the present system.

Some limitations can be overcome when the auralization is based on array microphone recordings instead of room simulations. A recent study (Minnaar et al., 2013) used multiple VSEs in a loudspeaker array similar to the one used in the present study that were recorded with a spherical 32-microphone array and rendered using a direct inversion method. This method was shown to lead to a very convincing auralization of complex scenes, even with moving sources. However, this happens at the cost of reduced flexibility because the scene cannot be changed once recorded. A spherical HOA microphone array with 52 1/4-inch microphones in a rigid sphere with a diameter of 10 cm has been developed and is currently being tested (Marschall et al., 2012). With this technique, array recordings of real acoustic scenes can be combined with simulation techniques to place target or interfering sources in a virtual scene. This could be done either by recording the background scene directly and by measuring impulse responses at the same position without background noise (which might not always be possible), or by combining the background recordings with target sources based on a room simulation.

2.5 Summary and conclusion

In this study, speech intelligibility in noise was used as a measure to assess the authenticity of a VSE based on a carefully calibrated room acoustic model of an existing classroom. The VSE was compared to the real room by means of T_{30}, C_{50}, and $IACC$. It was found that the average values for T_{30} and C_{50} measured in the VSE were very close to the values simulated in ODEON. The slight differences between the parameters measured in the classroom and the VSE were most likely caused by the setup of the room model in ODEON rather than by the LoRA processing or the reproduction room. However, the $IACC$ was found to be lower in the VSE than in the real room. The HA directivity patterns showed a reduced level of detail in the classroom compared to the anechoic chamber and a further reduction in detail in the VSE as a consequence of the slightly more diffuse sound field in the VSE compared to the real room.

In the listening experiments, the SRTs were generally found to be slightly higher in the VSE than in the classroom. It was shown that the SRTs in the VSE in the conditions with HAs improved when a static BF was used instead of an omnidirectional microphone, even though the improvement was slightly smaller than in the real room. Furthermore, the dependence of the SRT on the target source distance was found to be very similar in the VSE and in the classroom, when the NLS rendering method was used. The NLS method thus seems to preserve more of the crucial acoustical features of a real room than HOA.

Even though the SRTs differed between real room and simulation, all differential results translated well to the real world. Since the evaluation of new HA signal processing features typically considers such differential measures, the VSE system may represent a valuable tool for such testing where end users can be involved early in the HA development process. For the time being, NLS should be preferred over HOA for experiments in which the reduced spatial resolution of NLS compared to HOA is not too critical, like speech intelligibility experiments, because it seems to preserve more of the underlying cues.

Acknowledgements

The authors wish to thank the editor and the two anonymous reviewers for their constructive feedback and Sylvain Favrot and Pauli Minnaar for their valuable contribution to this study. This work was supported by a research consortium with Oticon, Widex and GN ReSound. Parts of this work were presented at the AIA-DAGA Conference on Acoustics in Merano, Italy, 18-21 March 2013.

3

Spatial perception and speech intelligibility with hearing aids[a]

Abstract

Cubick and Dau (2016) showed that speech reception thresholds (SRTs) in noise, obtained with normal-hearing listeners, were significantly higher with hearing aids (HAs) than without. Some of the listeners reported a change in their spatial perception of the sounds due to the HA processing, with auditory images often being broader and closer to the listener or even internalized. The current study investigated whether the worse speech intelligibility with HAs might be explained by a distorted perception of the acoustic scene and as a result a reduced ability to spatially separate the target speech from the interferers. SRTs were measured in normal-hearing listeners with or without "ideal" HAs (broadband, with linear, flat gain) in the presence of three interfering talkers or speech-shaped noises. The interferers were presented either at +/- 90° and 180° azimuth or collocated with the target sentences at 0°. Furthermore, listeners were asked to sketch their spatial perception of the acoustic scene. Consistent with the previous study, SRTs increased with HAs when the interferers were spatially separated. The spatial release from masking was lower with HAs than without. The speech perception data can be accounted for by a binaural speech intelligibility model. Even though the sketches showed a change of spatial perception with HAs, no direct link between the spatial perception and speech intelligibility could be shown.

[a] This chapter is based on Cubick, Buchholz, Best, Lavandier, and Dau (2017).

3.1 Introduction

In terms of speech intelligibility, hearing-aid (HA) users usually benefit most from their HAs in low-noise acoustic scenarios with a single talker. In more challenging acoustic situations, such as a social gathering in a crowded room, they typically have difficulties to follow a conversation (Bronkhorst, 2000), whereas normal-hearing listeners perform well almost effortlessly. Cherry (1953) introduced the term "cocktail-party" to refer to such situations, where a listener attempts to understand a target speaker among various competing interferers. It has been demonstrated that spatial auditory cues are utilized by the auditory system to facilitate good intelligibility in these situations, such that interferers cause less masking when they are spatially separated from the target talker in terms of their azimuthal position (Hawley et al., 1999; Plomp, 1976) or distance (Westermann and Buchholz, 2015b). In the case of spatially separated sources, speech intelligibility can be improved compared to collocated sources due to "better-ear" listening, where the sound at one ear, at a given moment, may provide an improved target-to-masker ratio, and/or due to the benefit of "true" binaural processing, often named binaural unmasking, which could be interpreted as a de-noising operation in the central auditory system, e.g. via an equalization-cancellation process (Durlach, 1972) that improves the "internal" target-to-masker level ratio. Both strategies, better-ear listening and binaural unmasking, have been considered in various speech intelligibility modelling approaches (e.g., Beutelmann and Brand, 2006; Beutelmann et al., 2010; Lavandier and Culling, 2010; Wan et al., 2010; Rennies et al., 2011; Lavandier et al., 2012; Wan et al., 2014; Chabot-Leclerc et al., 2016) and are thought to reduce the effects of energetic masking (EM) of the target sound by the interferer(s).

However, some effects of typical cocktail-party scenarios on speech intelligibility cannot be accounted for in terms of EM. Instead, the term informational masking (IM) has been introduced to cover interference effects caused by competing talkers that affect a listener's ability to understand a target talker even in the case of sufficient target energy (for a review, see Kidd et al., 2008). IM can refer to both difficulties in segregating speech mixtures (i.e., determining which parts belong to the target speech) and difficulties in terms of attending to a specific source in the sound mixture (i.e., overcoming confusion or distraction; Shinn-Cunningham, 2008). Spatial information regarding the target and the

interferers in a speech mixture can strongly affect the amount of IM such that sound sources that are perceived as spatially separate objects facilitate selective attention to one or the other source (e.g., Freyman et al., 1999). Spatial separation can be particularly effective when there is little other information available to separate the competing sounds (e.g., when the competing voices are of the same gender and/or have approximately the same sound pressure level). In fact, the magnitude of the "spatial release from IM" can even be larger than the "spatial release from EM" (e.g., Kidd et al., 2005). Moreover, it appears that any cue that supports the perception of spatial separation of the target and the interferer(s) is sufficient to provide a release from IM. Such a release has been reported for interaural time differences (ITDs) and interaural level differences (ILDs) alone (e.g., Glyde et al., 2013), monaural spectral cues associated with a separation in distance and elevation (e.g., Brungart and Simpson, 2002; Martin et al., 2012; Westermann and Buchholz, 2015b; Westermann and Buchholz, 2017a) and for illusory separation (e.g., Freyman et al., 1999).

Because of the importance of spatial information in relation to EM and IM, any degradation of the spatial cues caused by a hearing loss and/or HA signal processing could potentially impair speech intelligibility in a cocktail-party environment. In recent years, a number of studies have explored the possibility that hearing loss impedes spatial perception, e.g., in terms of localization ability (Noble et al., 1994; Lorenzi et al., 1999; Best et al., 2010; Best et al., 2011) or ITD discrimination performance (e.g., Durlach et al., 1981; Strelcyk and Dau, 2009; Spencer et al., 2016. Furthermore, several studies have suggested that HAs disrupt the auditory cues involved in spatial perception (Van den Bogaert et al., 2006; Wiggins and Seeber, 2012; Akeroyd and Whitmer, 2016; Cubick and Dau, 2016; Hassager et al., 2017). For example, Hassager et al. (2017) showed that the localization accuracy in a moderately reverberant room was substantially degraded as a consequence of fast-acting dynamic-range compression in the left-ear and right-ear HAs, independent of whether the compression was linked across aids or not. The distortions were attributed to the stronger amplification of the low-level portions of the (speech) signals that were dominated by early reflections and reverberation, relative to the higher-level direct sound components. As a result, increased diffusiveness of the perceived sound and broader, sometimes internalized ("inside the head"), sound images as well as sound image splits of a single speech source were observed both in normal-hearing

and hearing-impaired listeners. However, the effects of these distortions on speech intelligibility were not investigated in that study. Cubick and Dau (2016) measured speech reception thresholds (SRTs) in normal-hearing listeners using omnidirectional regular production HAs with linear (i.e., level-independent) amplification. They found about 4 dB higher SRTs, i.e., degraded speech intelligibility, in the conditions with HA amplification compared to the conditions without amplification, in a setting with spatially distributed loudspeakers inside a classroom. The study did not provide a fully conclusive explanation for the elevated SRTs. However, some of the listeners in the study reported a largely degraded spatial perception of the acoustic scene in the conditions with HA processing; the auditory images associated with the sound sources were often broader with HAs and sometimes perceived to be closer to the head than to the actual source position. These findings suggested that the elevated SRTs might, at least partly, reflect the reduced ability of the listeners to perceptually separate the target and interfering sounds due to the disrupted cues on which localization is based.

Inspired by Cubick and Dau (2016), the current study investigated the potential effect of degraded spatial cues when listening through HAs on SRTs in spatially separated masking conditions. To do so, a very basic amplification scheme was used, that included linear gain and no sophisticated signal processing that might cause additional distortions. Thus, ideally, the only distortion of the incoming sound would be caused by the position of the microphones above the ears, which leads to modified spectral cues compared to natural listening. It was tested whether elevated SRTs as in Cubick and Dau (2016) would also be found with such "optimized" HAs. Furthermore, it was investigated to what extent HA processing affects the amount of IM (versus EM) in a complex acoustic setting with several interferers. SRTs were measured in a room with a target speaker in front of the listener and three interferers. The interferers were either competing talkers (potentially causing a large amount of IM) or noises (producing little if any IM), that were either spatially distributed around the listener (at +/- 90 and 180°) or collocated with the target source. In the extreme case, if the HAs would completely remove all spatial information, no spatial release from masking (SRM) would be expected for either mixture. On the other hand, if using HAs would distort the spatial information enough to reduce the listeners' ability to spatially separate the target and the interferer signals, then

this would reduce the spatial release from IM and the impact would primarily be seen in the case of speech interferers. To characterize the influence of HA processing on the spatial perception of the acoustic scenes in the horizontal plane, the same listeners were also asked to draw sketches to indicate the position and spatial distribution of the sound images they perceived using a method inspired by earlier studies (Plenge, 1972; Blauert and Lindemann, 1986; Cubick and Dau, 2016).

3.2 Methods

3.2.1 Listeners

Ten native Australian-English speaking listeners participated in the experiment. Most listeners were either students from Macquarie University or employed at the National Acoustic Laboratories. The average age of the listeners was 31 years. All listeners were required to have pure tone audiometric thresholds within 20 dB HL at audiometric frequencies between 125 Hz and 6 kHz. If a listener did not have a recent audiogram, an audiogram was measured before the experiment. All listeners received written information about the experiment and gave informed consent prior to testing. The experiments were approved by the Australian Hearing Human Research Ethics Committee. Listeners who were not employed at the National Acoustic Laboratories received a small gratuity in compensation for their travel expenses.

3.2.2 Stimuli and apparatus

Stimuli

For the target sentences in the SRT measurements, a speech corpus based on the Bamford–Kowal–Bench (BKB) sentence material (Bench et al., 1979) was used. This open-set corpus consists of 1280 short, meaningful sentences with a simple syntactical structure, which are divided in 80 lists of 16 sentences each. The sentences are spoken by a female Australian-English talker. In the speech-on-speech conditions, recordings of three female monologues were used as maskers (spoken by three female talkers different from the target). Even though all four talkers were female, the timbre of their voices was quite different. For speech-in-noise conditions, three instances of stationary speech-shaped noise

(SSN) were generated that matched the individual long-term magnitude spectra of the three interfering talkers. To do so, a 2048-tap finite impulse response filter was derived from the difference between the spectrum of a white Gaussian noise sample and the estimated spectrum of the masking speech. Convolving this difference filter with the white noise yielded the SSN.

Hearing aids

The HAs used in the experiment were based on the premise that the highest possible sound quality achievable with common HA hardware should be provided, such that, ideally, the only influence on the ear signal compared to the unaided condition would be the change in the pinna cues. A real-time HA processing platform was used that was developed in-house and that was run on a separate computer. The system used the microphones and receivers of standard behind-the-ear HA shells (Phonak Ambra). The microphone signals were amplified by a custom-made preamplifier and then fed into the computer via an RME Fireface UC audio interface. After the real-time processing, the output signal was sent to a calibrated limiter that interrupted the signal if it exceeded 85 dB (A). From here, the signal reached the HA receiver, which was coupled to the listeners' ears via tubes with foam plugs. The only HA processing used in the experiment was the application of a linear, frequency-independent ("flat") gain on the omnidirectional microphone signal of the two front microphones of the HAs. The gain was adjusted in the software of the real-time platform to provide an approximately constant insertion gain of 10 dB across all frequencies between 63 Hz and 10 kHz, evaluated on a 2cc coupler in a Siemens Unity 2 HA measurement box. The resulting gain settings were kept the same for all participants. In all conditions with HAs, the playback level of the loudspeakers was reduced by 10 dB to keep the sound pressure level at the listener's ears approximately constant between conditions with and without HAs.

3.2.3 Experimental procedure

Speech intelligibility

The experiment was conducted in a sound-treated listening room with a reverberation time T_{30} of about 200 ms. The listeners were seated in the centre of a ring of 16 Genelec 8020 loudspeakers with a radius of 1.3 m. The stimuli were played from a computer running Matlab and delivered through an RME

Fireface UFX audio interface and two RME ADI 8 DS 8-channel digital/analog converters. During the experiment, only four of the 16 loudspeakers were used for playback. The target sentences were always presented from the front (0°) 1 s after a 200-ms long 1 kHz tone pulse. The three maskers (speech or SSN) were presented continuously either from three loudspeakers at ±90 and 180° or from the same loudspeaker as the target sentences.

The target speech and the interferers were calibrated using an omnidirectional measurement microphone (Brüel & Kjær 4134) at the listening position. The masker level was kept constant at 65 dB (A) throughout the experiment, whereas the level of the target sentences was adapted using the 1-up-1-down procedure described in Keidser et al. (2013). Each threshold was determined using 16-32 sentences. Each run lasted until either the standard error for the threshold estimate was below 0.8 dB or the maximum number of 32 sentences was reached. The experimenter was seated inside the test room, but outside the loudspeaker ring and scored the correctly understood morphemes on a laptop that remote-controlled the PC used for stimulus generation.

Spatial perception

Similar to the procedure in Cubick and Dau (2016), the listeners were asked in each run to draw the perceived position (both in angle and distance) and the extent of the target and masker sounds into a sketch of the listening setup with a schematic head in the middle indicating the listener's position and a circle indicating the radius of the loudspeaker ring. The listeners were given time to make the drawings in the beginning of each run, after the presentation of the first sentence. Some listeners updated their drawings during the run after hearing more samples of the target sentences.

Listening effort

At the end of each run, the listeners were also asked to rate the listening effort on a 13-point scale ranging from 0 (no effort) over 2 (very little effort), 4 (little effort), 6 (moderate effort), 8 (considerable effort), 10 (much effort), to 12 (extreme effort), based on Luts et al. (2010). They were instructed to rate specifically how effortful it was to separate the target speech from the interferers.

3.2.4 Conditions

Overall, eight conditions were tested. The three interferers were either speech or SSN, they were either spatially collocated with the target speech or separated, and the listeners either wore HAs (aided) or not (unaided). All listeners were tested twice in each of the resulting eight combinations. The experimental conditions were counterbalanced across subjects based on a Latin Square Design with the only restriction that the four aided and the four unaided conditions were always tested in consecutive runs. This was done to avoid listeners taking off and inserting the HAs more often than necessary, and to avoid effects caused by potential variability of HA positioning. The testing took part either in one session with a total duration of about two hours including breaks or in two separate sessions of about 1 hour 15 min each, depending on the listener's preference.

3.2.5 Stimulus analysis

To allow for the analysis of the ear signals as they occurred in the experiment, binaural room impulse responses were measured at the listening position with a Brüel & Kjær 4128 head-and-torso simulator (HATS) with and without HAs for all loudspeakers used in the experiment. The impulse responses were measured with two repetitions of a 6-s logarithmic sine sweep (Müller and Massarani, 2001) and truncated to a length of 300 ms for the analysis. To compensate for level differences between the left and the right ear of the HATS, the first 3.85 ms of the impulse responses of both ears (corresponding to the direct sound from the front loudspeaker before the first room reflection) were filtered with the long-term magnitude spectrum of the target speech. The RMS values of the resulting filtered direct sound signals were compared and the signals were adjusted to have the same RMS. The resulting correction factor between left- and right-ear signals was subsequently applied to all recorded signals. The target sentences and interferer signals were convolved with the adjusted impulse responses and the contributions from the different sources were summed to approximate the ear signals that occurred during the experiment. It should be noted though that the calibration of the experiment was done relative to an omnidirectional microphone in the centre of the circle, and that the signal-to-noise ratio (SNR) was defined accordingly. Thus, the presence of the HATS inside the sound field may change the effective SNR at the ears compared to the value measured

during calibration. However, similar changes are to be expected due to the head of the listeners during the experiment.

3.2.6 Modelling

In order to better understand the influence of the HAs in the present experiment, a model was used to quantify the amount of EM in the tested conditions. An updated version of the model proposed by Collin and Lavandier (2013) was used to predict binaural speech intelligibility in the presence of multiple non-stationary noises. The model is based on the model of Lavandier and Culling (2010). It combines the effects of better-ear listening and binaural unmasking and is based on two inputs, the ear signals generated by the target, and the ear signals generated by the sum of all interferers. Based on these inputs, the model computes the better-ear target-to-interferer ratio and the binaural unmasking advantage in frequency bands, and finally produces the (broadband) effective target-to-masker ratio in the corresponding condition (Jelfs et al., 2011; Lavandier et al., 2012), referred to as the "binaural ratio" in the following. Binaural ratios are inversely proportional to SRTs, such that high binaural ratios correspond to low SRTs. The predicted differences in terms of (inverted) binaural ratios were directly compared to corresponding SRT differences, without any fitting of the model to the data. The predictions in Collin and Lavandier (2013) are based on a short-term version of the model, similar to Beutelmann et al. (2010) and Rhebergen and Versfeld (2005). A ceiling parameter corresponding to the maximum better-ear ratio allowed by frequency band and time frame was introduced, to avoid the target-to-masker ratio tending to infinity in interferer pauses. The binaural unmasking advantage is set to zero if the interferer power is zero at one of the ear in the considered band and frame.

The predictions presented here were computed using two minutes of the masker signal in each of the eight tested conditions. The target (either unaided or aided) was represented by averaging 144 target sentences, whereby the first 680 ms were omitted and all sentences were truncated to the duration of the shortest sentence. The root-mean-square power of the averaged signal was then equalized to that of the corresponding (unaided/aided) collocated (speech or noise) maskers. The masker signals and target sentences used for the modelling were obtained as described in Sec. 3.2.5. The model used 24-ms half-overlapping Hann windows as time frames (having an effective duration

of 12 ms; Beutelmann et al., 2010), a gammatone filterbank (Patterson et al., 1987) with two filters per equivalent rectangular bandwidth (ERB; Moore and Glasberg, 1983), and a 20-dB ceiling parameter.

3.3 Results

3.3.1 Speech intelligibility

Fig. 3.1 (a) shows the mean SRTs and standard deviation across participants for the unaided (squares) and the aided conditions (circles) for both the separated (open symbols) and the collocated case (black filled symbols). The results for the speech interferers are shown on the left, the results obtained with SSN on the right. The lowest SRT of -12 dB was observed in the unaided condition with separated speech interferers. With HAs, the thresholds increased for this configuration by 2.5 dB. The average unaided threshold with separated SSN interferers was -9.8 dB, and thus 2.2 dB higher than with the speech interferers. With HAs, the SRT obtained with separated SSN increased by 2 dB to -7.8 dB (i.e., 1.7 dB above those obtained with speech interferers).

The thresholds for the collocated conditions were in all cases higher than for the corresponding condition with separated maskers. The SRM (Fig. 3.1 (b)) was calculated as the difference between the individual separated and collocated SRTs. The highest SRM (8 dB unaided, 6.5 dB aided) was found in the conditions with the speech interferers (left). In the case of SSN (right), the SRM was much lower (2.4 dB unaided, 0.8 dB aided). A linear mixed effects model was fitted to the SRT data with the three factors 'Masker', 'Distribution', and 'HA condition'. The full model with all interaction terms was then simplified by removing the non-significant three-factor interaction. The subsequent ANOVA showed that all three main effects 'HA condition' [$F(1,144) = 138.84$, $p < .0001$], 'Masker' [$F(1,144) = 7.7513$, $p = 0.0061$], and 'Distribution' [$F(1,144) = 522.74$, $p < .0001$], and the two-factor interactions between 'HA condition' and 'Distribution' [$F(1,144) = 15.27$, $p = 0.0001$] and 'Masker' and 'Distribution' [$F(1,144) = 191.92$, $p < .0001$] were significant. Only the interaction between 'HA condition' and 'Masker' was not significant [$F(1,144) = 1.51$, $p = 0.2213$].

Similarly, a linear mixed effects model was fitted to the SRM data with factors 'HA condition' and 'Masker'. Here, the ANOVA showed significant main effects

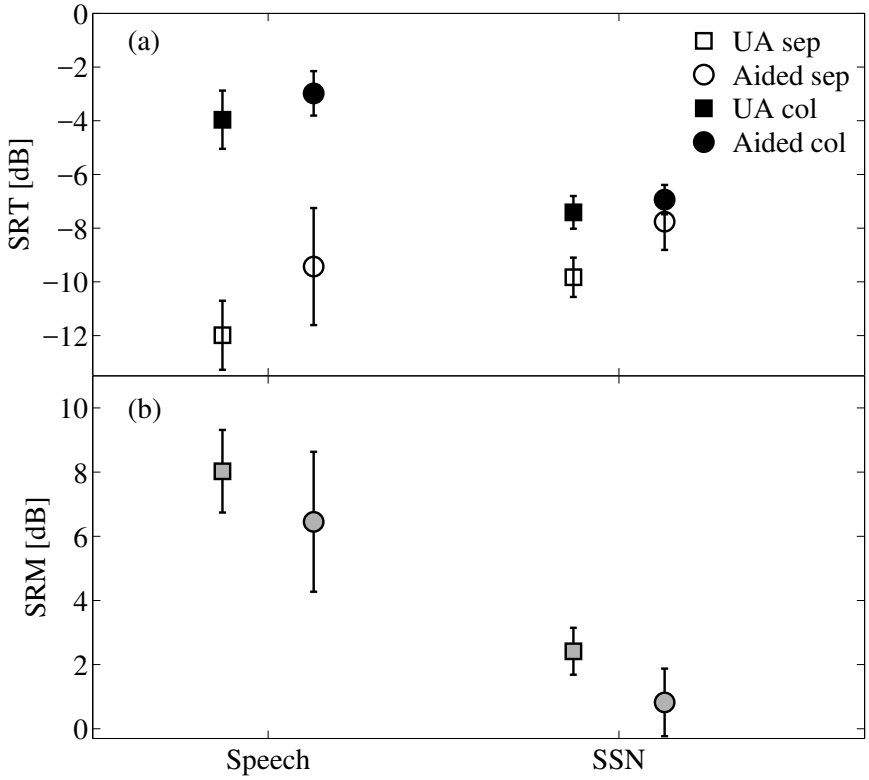


Figure 3.1: (a) Average speech reception thresholds and standard deviation for ten normal-hearing listeners for unaided (UA, squares) and aided (circles) conditions using speech interferers (left) or SSNs (right), for collocated (col, black filled symbols) and separated maskers (sep, open symbols). (b) Average spatial release from masking (grey filled symbols) and standard deviation across listeners for the two masker types and HA conditions.

for both 'HA condition' [$F(1, 67) = 9.63$, $p = 0.0028$] and 'Masker' [$F(1, 67) = 122.00$, $p < .0001$], but no significant interaction [$F(1, 67) = 0.0006$, $p = 0.9804$], i.e., the SRM in the SSN conditions was significantly lower than in the speech conditions, and HAs reduced the amount of SRM similarly in the SSN and the speech conditions.

3.3.2 Spatial perception

Fig. 3.2 shows the digitized data from the position sketches collected from all listeners for the four conditions with speech interferers. Pixels representing the target sound are shown in blue, pixels belonging to the interferers are shown in red. The outer circle indicates the ring of loudspeakers, the inner

circle represents the listener's head. The squares on the outer circle indicate the loudspeakers that were actually playing in the corresponding condition. All images were superimposed; therefore, areas of higher saturation represent areas that were marked as belonging to the auditory image by more listeners. The left column represents the unaided conditions; the right column shows the data from the aided conditions. In the unaided separated case (top left panel), all listeners drew clearly separated images for the target signal and the three distracting talkers. Only one listener sketched the target sound image as being close to and inside the head in both repetitions of the experiment. Compared to the unaided condition, the corresponding sketches for the aided separated condition (top right) indicate a much larger variability in the data. In many cases, not only was the image position more variable across listeners, but the images were also often broader and differed in their perceived distance. Several listeners indicated that they had perceived the target sound and/or the interferers inside their head, or to be spread indistinguishably in the whole room. In the collocated conditions (bottom panel), most listeners indicated the target and the interferer sound images to be somewhere between their head and the front loudspeaker in the unaided condition (left panel). Again, with HAs, the data showed more variability where, e.g., the interfering sounds were perceived from different directions and resulted in broader auditory images and sometimes internalized percepts.

Fig. 3.3 shows the corresponding results for the conditions with SSN. One effect that cannot be seen from the figure is that, unlike in the conditions with speech interferers, all listeners indicated only one or two interfering sources in all SSN conditions. Apparently, the spectral differences between the noise maskers were not sufficient to perceive them as separate auditory objects, and the three noise maskers were fused into one or two objects instead. In the unaided case (top left), the target speech again yielded sharply focussed and fairly narrow auditory images between the listener and the loudspeaker at 0°, as seen by the narrow blue "wedge". All listeners perceived the target speech externalized in this condition. In the individual sketches (not shown), the three noise sources were fused into either two wide auditory images to the left and right of the listener, or into a single auditory image behind the listener or perceived all around the room. In the aided separated condition (top right), the sound sources often changed their position compared to the unaided case.

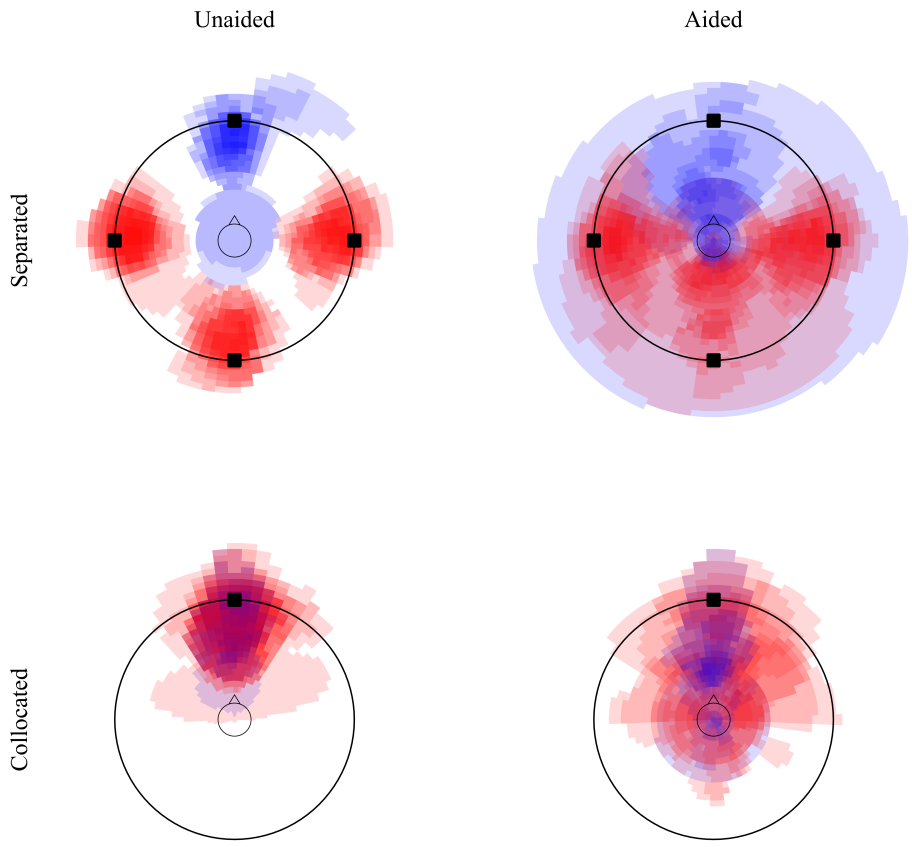


Figure 3.2: Superimposed images of the perceived positions of the sound sources for target speech (blue) and interfering talkers (red) for the unaided conditions (left column), aided conditions (right column), and the separated condition (top row) and the collocated conditions (bottom row). The two circles indicate the listener's head (inner) and the loudspeaker ring (outer) as shown in the sketch template provided to the listeners during the experiment. The black squares indicate the positions of the loudspeakers through which the stimuli were presented.

Some listeners perceived the target speech inside their heads or from behind them. Also the position of the noise sound images often moved. In some cases, the images were indicated closer to the listener or all over the room. In the unaided collocated condition (bottom left panel), the sketches show a larger spread than in the corresponding condition with interfering talkers (Fig. 3.2, bottom left panel), but in the majority of the cases, the auditory images of both target and masker were perceived in the front. In the aided collocated condition (right bottom panel), there seems to be a tendency that the noise maskers created a larger auditory image than the target speech, and that the noise sources were perceived far away and broad, whereas the auditory image of the target speech tended to be closer to the listener and more compact. Interestingly, there were some cases, especially in the aided collocated conditions, where target and masker seemed to be perceived more separated than in the corresponding unaided condition[a].

3.3.3 Listening effort

Fig. 3.4 shows the listening effort ratings of the participants. Again, a linear mixed effects model was fitted to the individual data, averaged across the two repetitions. An ANOVA revealed that the main effect 'HA condition' and all of its interactions were not significant. The aided conditions thus did not necessarily require a higher listening effort than the unaided conditions. The average effort rating was significantly higher for the speech masker than the SSN [$F(1,147)$: 68.33, $p < 0.0001$] and for the collocated compared to the separated conditions [$F(1,147)$: 30.04, $p < 0.0001$]. Also the interaction between Masker and Distribution [$F(1,147)$: 6.79, $p = 0.0101$] was found to be significant. Interestingly, the effort ratings were only weakly (positively) correlated with the SRTs, but there was a tendency that listeners reported higher listening effort in conditions with higher SRTs, i.e., worse speech intelligibility.

[a] For a quantitative analysis, first attempts have been made to derive a measure similar to the d' known from signal detection theory (Wickens, 2002), which potentially captures the degree of perceived separation between target and maskers. Correlation of this value with the SRTs might help explain large individual differences between listeners.

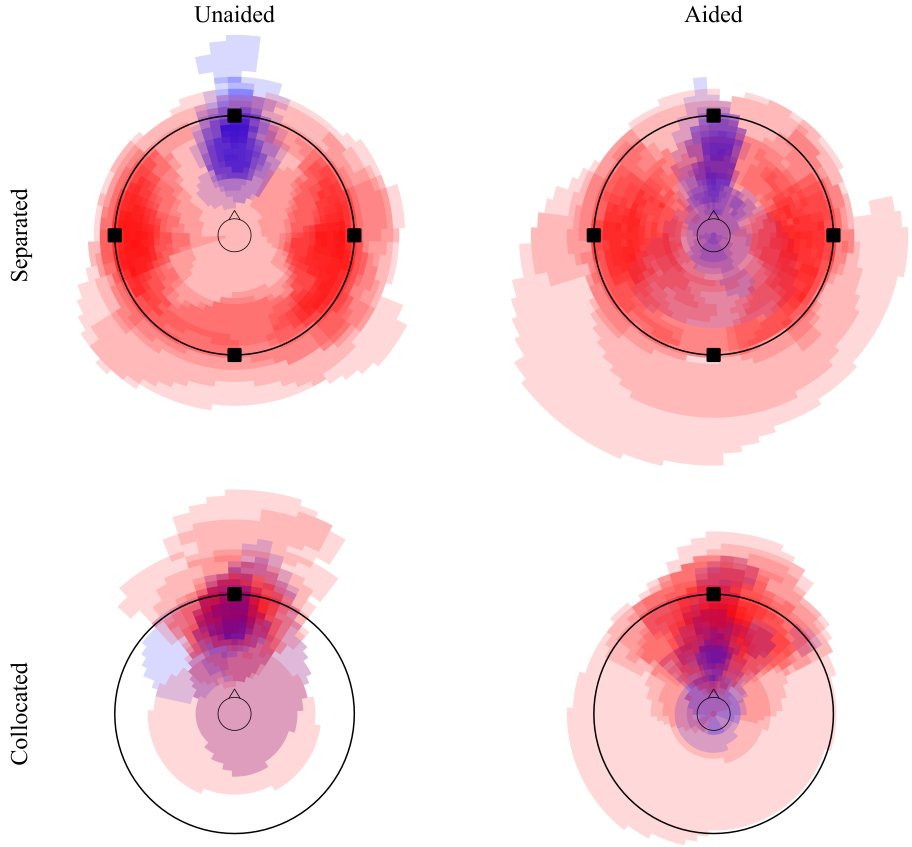


Figure 3.3: Superimposed images of the perceived positions of the sound sources for target speech (blue) and SSN (red) in the unaided condition (left column) and the aided condition (right column) and the separated (top row) and collocated condition (bottom row).

3.3.4 Stimulus analysis

Figure 5 shows the long-term magnitude spectra of all concatenated target sentences and of the interferer signals at the ears of the HATS. The left column represents the unaided condition as measured through the ears of the HATS, the right column shows the aided condition, measured as the acoustic output of the HAs placed on the HATS' ears and coupled to its ear canals with foam tips. The top row shows the spectra of the target sentences, the middle row shows the frequency-dependent SNR in the collocated conditions, i.e., the difference (in dB) between the target sentence spectrum and the interferer spectrum at a

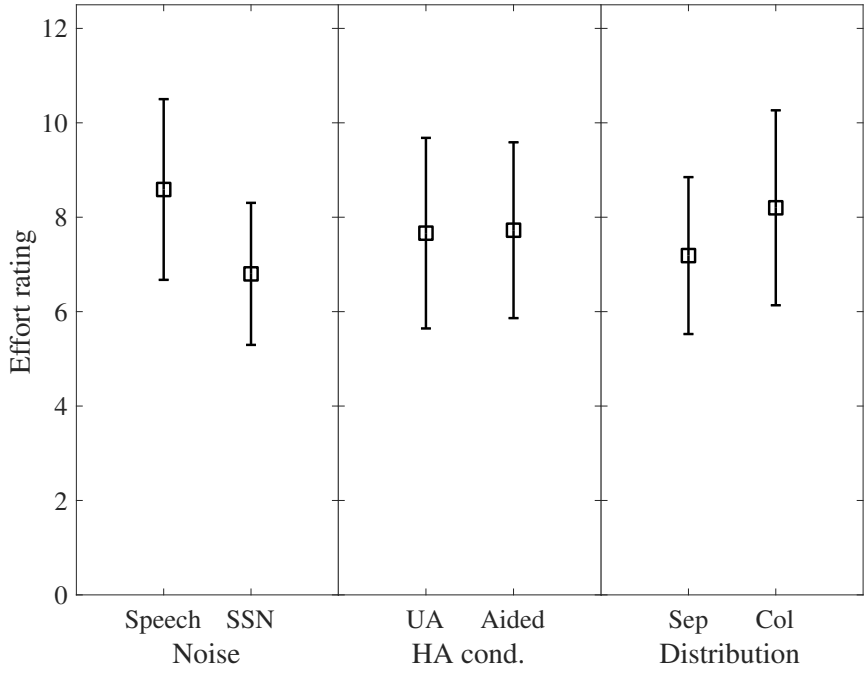


Figure 3.4: Effort rating for the three factors tested in the experiment. Only the average ratings for Noise and Distribution and their interaction were significantly different.

nominal SNR of 0 dB. The bottom row shows the level difference between the ear spectra of the collocated condition and the separated condition. Dashed lines represent the left ear, solid lines the right ear. All interferer spectra are only shown for the condition with three female interfering talkers, because the magnitude spectra of the SSNs are identical.

Considering the unaided speech spectrum (top left panel), the spectra for the left and the right ear are nearly identical, due to the level equalization of the ear signals by their direct sound. The small remaining deviations between the left and the right ear may be attributed to the characteristics of the room or slight asymmetries in the setup. In comparison, the aided speech spectra (top right panel) show larger deviations between left and right ear, especially in the highest frequencies. These deviations resulted from differences in the sensitivity of the HA microphones, such that the left HA could not be adjusted to a low enough gain to achieve a completely flat frequency response on the 2cc coupler. Another effect worth noting is that the aided target speech spectra do not show the broad resonance peak at the mid frequencies that can be

seen in the unaided frequency response, a typical feature of a 0° head-related transfer function and an indication that HAs change the signal spectrum at the ear. While the frequency-dependent SNR between the target sentences and the speech interferers on the HATS' ears (middle left) is generally close to zero, some deviations can be seen, in particular at the lowest and the highest frequencies. These deviations are caused by the differences in the long-term spectra between the different voices used as target and interferers and were expected. With HAs (middle right), the SNR function shows an additional overall level offset. The difference spectra between the separated and the collocated interferers (bottom row) fluctuate around zero at low and mid frequencies, whereas higher fluctuations occur at high frequencies (unaided). In the aided case (bottom right), the level difference between the two signals is close to zero, but there is a slight negative offset, indicating a slightly higher level of the separated maskers.

3.3.5 Modelling

As observed previously, using omnidirectional HAs led to an increase in SRT in all tested conditions. Fig. 3.6 presents this "HA disadvantage" calculated for each condition (collocated and spatially separated, speech and SSN interferers) as the difference between the SRTs in the aided and unaided conditions. The solid lines indicate the disadvantages predicted by the model. The average and maximum prediction errors (absolute difference between measured and predicted disadvantages) across conditions were 0.6 and 1.1 dB, respectively. The deleterious effect of the HAs in the tested conditions is predicted by the model, suggesting that this effect is associated with EM, since the model cannot account for IM.

3.4 Discussion

The lowest SRTs in this study were found in the unaided condition with separated speech interferers. In the corresponding condition with SSN interferers, the average SRT was 2.2 dB higher. This was expected for an interferer consisting of three monologues, because the individual speech interferer signals are characterized by highly fluctuating envelopes that give ample opportunity for 'listening in the dips', which is generally found to improve intelligibility (e.g., Festen and Plomp, 1990). In contrast, the SSN maskers exhibit a relatively constant

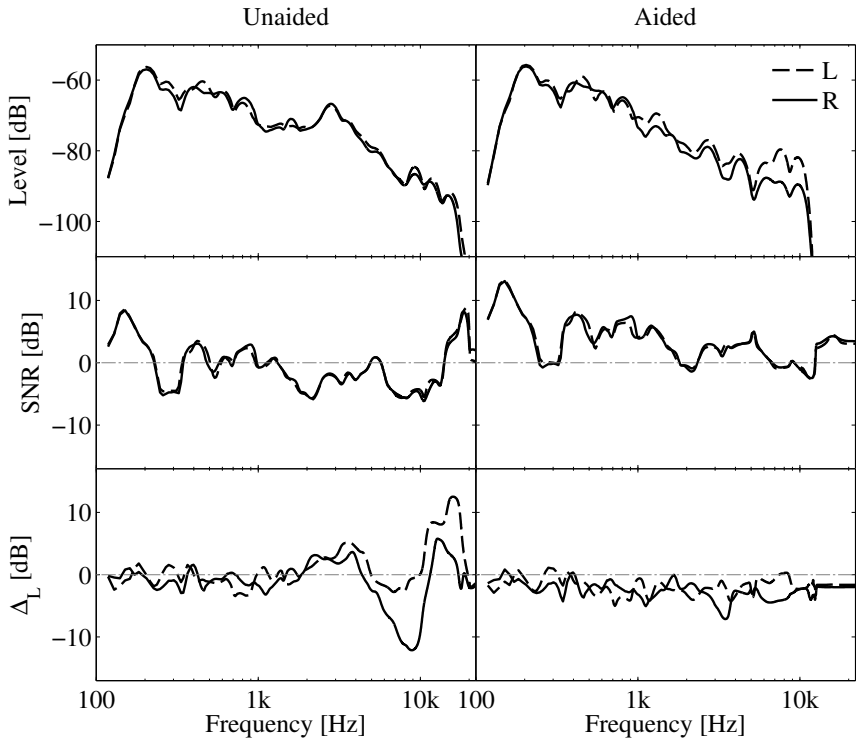


Figure 3.5: Magnitude frequency spectra of the target signal (top row), frequency-dependent SNR between the target and the collocated interferer signal at a nominal SNR of 0 dB (middle row), and difference between the collocated and separated interferer spectra at the ear (bottom row) for the unaided (left column) and the aided (right column) condition. The spectra were smoothed with a 1/3-octave wide moving average filter. Spectra for the SSNs are not shown, because their long-term average magnitude spectrum is identical to that of the speech interferers.

envelope with less low-frequency modulations that offer fewer opportunities for dip listening. SSN thus represents a more effective masker in the separated condition.

With HAs and separated interferers, speech intelligibility was generally worse than without HAs, independent of the type of interferer, in accordance with the findings of Cubick and Dau (2016). However, the SRT increase found in the conditions with separated interferers in this experiment was only 2.5 dB for speech interferers and 2 dB with SSN, compared to the 4 dB (with SSN) reported by Cubick and Dau (2016). This difference might be due to the fact that in this study a PC-based real-time processing platform was used, allowing

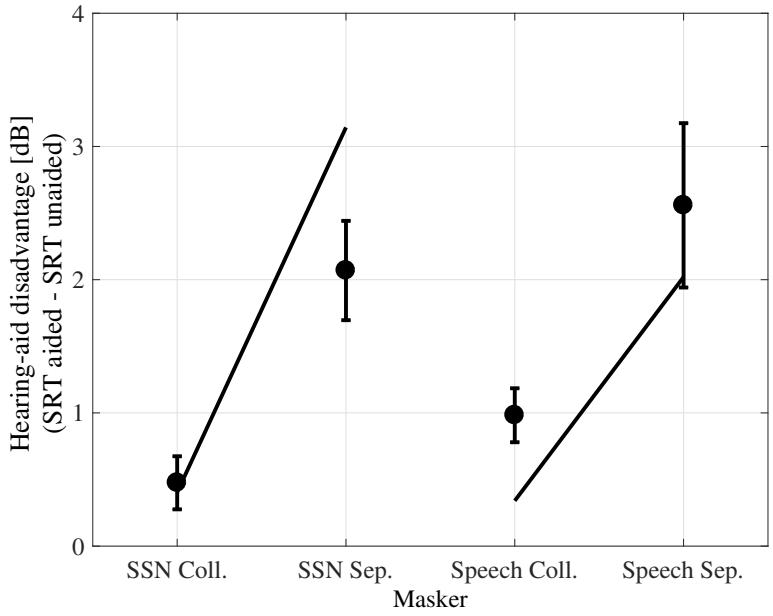


Figure 3.6: Hearing-aid disadvantage in dB evaluated as the difference between the SRTs in the aided and unaided conditions for the SSN and speech interferers in the collocated and separated conditions. The circles represent the measured average disadvantage across listeners, the error bars indicate the standard error. The lines indicate the disadvantage predicted by the model.

for a wider bandwidth, less noise, and a better overall sound quality than the regular production HAs used in Cubick and Dau (2016). Another difference between the two studies is that the room in Cubick and Dau (2016) was more reverberant ($T_{30} = 0.5$ s compared to 0.2 s) than the room considered in the present study. This might have made intelligibility with HAs in Cubick and Dau (2016) worse, because the longer reverberation time leads to a greater amount of diffuse sound energy inside the room, which is considered to be detrimental for speech intelligibility (e.g., Plomp, 1976). This increase in reflected energy might be particularly detrimental in aided listening since the natural directivity of the pinna is lost, which would otherwise attenuate sounds from the back to some extent and thus emphasize the direct sound. With HAs, the omnidirectional microphones on the side of the head result in a higher sensitivity towards lateral angles. Another difference between the two studies is that the loudspeakers in Cubick and Dau (2016) were placed at ±112.5°, in contrast to ±90° used in the current study.

The average SRTs in the collocated conditions were consistently higher than those in the separated conditions. The resulting SRM was much larger in the unaided conditions with speech interferers (8 dB) than in the unaided conditions with SSN (2.4 dB). This is somewhere in between the extremely high SRM found with highly unrealistic speech corpora like the coordinate response method (CRM, Bolia et al., 2000) and the lower values observed with more realistic speech materials and environments (e.g., Westermann and Buchholz, 2015a). It has also been found that SRM was larger for interferers that cause some degree of IM than those that only cause EM (e.g., Kidd et al., 2005). This effect has been referred to as spatial release from IM. The resulting spatial release from IM in the present study was estimated to be 5.6 dB, irrespective of the HA condition. In other words, the HAs reduced the SRM by 1.6 dB, irrespective of the type of interferer. This suggests that, while a release from IM was observed in this study, it did not change with the use of HAs. The reduction of SRM by the HAs could therefore be fully attributed to EM, not IM. The results from the modelling are consistent with this conclusion (see below).

Interestingly, the collocated SRTs showed hardly any difference between the unaided and the aided conditions. Intuitively, this makes sense, since the loss of the pinna cues caused by the HAs should have a more detrimental influence on settings in which spatial cues are crucial, whereas speech intelligibility in the collocated condition is mostly dominated by monaural cues because the signals at the two ears are nearly identical. There was hardly any difference in SRT between the aided separated condition and the aided collocated condition with SSN. This implies that also in the aided separated condition with SSN, speech intelligibility was dominated mainly by monaural cues, such as the SNR at each ear, since the stationary SSN did not allow for better-ear glimpsing and the perceived position of the sources probably does not matter for a purely energetic interferer signal (Freyman et al., 1999).

Speech intelligibility was decreased more strongly by HAs when the maskers were spatially separated than when they were collocated. This is reflected in the lower SRM found with HAs (cf. Fig. 3.1). The same effect was observed in the predicted SRMs, calculated as the difference between the spatially-separated and the collocated binaural ratios. In the model, the better-ear and binaural unmasking components are computed independently, hence their relative contribution

to SRM can be evaluated. The predicted binaural unmasking advantage was neither influenced by the HA, nor by the type of interferer. It accounted for about 1.5 dB of the overall SRM. The predicted better-ear advantage was very small for the unaided SSN condition, reflecting that there is no long-term better ear effect with a masker on either side of the listener, and not much glimpsing available within three unmodulated SSNs. The better-ear advantage was -2 dB in the aided SSN condition. This might be explained by the fact that, at high frequencies, the listener's head acts as a small baffle for the omnidirectional HA microphones, thereby effectively amplifying the high frequencies for sounds from the sides, such that the effective SNR at the ears is worse in the case of spatially separated maskers than when the maskers are collocated with the target speech. A long-term version of the model, which considers the whole duration of the signals instead of short-term predictions, provided very similar better-ear predictions, indicating that the better-ear disadvantage for SSNs is a long-term SNR effect rather than associated with short-term glimpsing.

The predicted better-ear benefit was larger for speech maskers than for SSNs, i.e., the model predicts some better-ear glimpsing with the speech maskers, both in the unaided condition (1.3 dB) and in the aided condition (2.3 dB). In addition to this better-ear glimpsing benefit, the model predicts the SRTs to be about 6 dB lower for the collocated speech compared to the collocated SSN, whereas the stationary model predicts similar SRTs, such that "monaural glimpsing" was quite strong even with three speech maskers involved. The model was also used to predict the spatial release from IM in the unaided and aided conditions involving the speech maskers. Since the model can only predict the effect of EM, not IM, this IM release was estimated as the difference between the measured and predicted SRMs for the speech maskers. It should be noted that EM prediction errors were thus incorporated in this IM release estimation. The predicted spatial release from IM was 5.1 dB in the unaided condition and 5.2 dB in the aided condition, supporting our interpretation that spatial release from IM was probably not influenced by the HAs in the present study.

Listening effort ratings showed that conditions with speech maskers were generally perceived as more effortful than conditions with noise maskers, which is consistent with the estimated higher IM with speech maskers in speech con-

ditions. In addition, conditions with collocated maskers required more effort than conditions with separated maskers. The effort ratings for conditions with HAs were not significantly different from conditions without HAs. This might, in parts, be due to the experimental procedure, because effort ratings are probably prone to sequence effects and all aided/unaided conditions were tested in separate blocks. It might also again be related to IM, which was not affected by the use of hearing aids. Interestingly, the listening effort ratings were only weakly correlated with SRTs.

Figs. 3.2 and 3.3 demonstrated that the spatial perception of the acoustic scene was distorted in conditions with HAs. However, it would be valuable to quantify this influence and to correlate it to the speech intelligibility results, especially in the conditions with separated maskers. Only normal-hearing listeners were tested in the present study. An important question is whether similar effects can be seen in hearing-impaired listeners. While there is evidence that HAs disturb binaural cues, the effect of these distortions on hearing-impaired listeners, who typically show reduced frequency resolution and deficits in temporal processing, is not easy to predict in the conditions tested in this study. Bronkhorst and Plomp (1992) and Marrone et al. (2008) found that hearing-impaired listeners benefit less from spatial separation of target speech and maskers, especially in more reverberant conditions. Eventually, a similar experiment should be run with real HAs instead of the best-case HAs used in this study. Modern HAs with their highly non-linear and adaptive processing have been shown to affect binaural cues and spatial perception (e.g., Keidser et al., 2006; Van den Bogaert et al., 2006; Van den Bogaert et al., 2008; Brown et al., 2016). The spatial perception may thus be even more distorted with regular production HAs.

Several listeners reported hearing both the target speech and the 180°-speech interferer to be very close to or even inside their head. This reminds of the findings of Plomp (1976), where interferers at 180° were more effective than at lateral angles, especially in rooms with short reverberation times. Given that the perceived separation is known to be crucial for spatial release from IM, it is surprising that the distortions induced by the HAs (and captured in the sketches) did not have a larger effect on the SRM for speech maskers than that for SSN maskers. However, because the separations were so large (90°), it is

possible that the broader images were still sufficiently distinct from one another to support segregation. It might be interesting to investigate the effect of HA processing on speech intelligibility in conditions with interferers close to the target speaker in terms of azimuth angle, where IM effects have been reported (e.g., Westermann and Buchholz, 2017b).

3.5 Summary and conclusions

In this study, SRTs of normal-hearing listeners were found to be worse in aided than in unaided conditions, and SRTs measured with spatially separated SSN interferers were higher than with interfering talkers. Substantial SRM was found for the speech interferers, whereas a much smaller SRM was observed with SSN. HAs reduced SRM to the same degree with speech and with SSN. It is therefore concluded that the reduction of SRM by hearing aids can be described entirely by effects of EM, supported by predictions of a binaural speech intelligibility model.

Acknowledgements

This work was carried out in connection to the Centre for Applied Hearing Research (CAHR), supported by Oticon, Widex, and GN ReSound, and the Technical University of Denmark.

4

Effects of stimulus bandwidth and playback room on distance perception[a]

Abstract

This study investigated effects of playback room and stimulus bandwidth on auditory distance perception. Two experiments were conducted in which listeners rated the distance of headphone-presented speech stimuli that were generated using individual binaural room impulse responses (BRIRs) measured for nine different distances. Experiment 1 was carried out in the same room where the BRIRs had been recorded, whereas experiment 2 was performed in a listening booth using the same acoustic stimuli. In both experiments, one broadband (0.05 to 15kHz) and two low-pass filtered versions (with cut-off frequencies at 2 and 6 kHz) of the speech were considered, inspired by earlier studies of bandwidth effects on distance and externalization perception. It was found that the results obtained in the recording room (experiment 1) differed from those in the listening booth, showing a less compressive relationship. The variability in the distance ratings was found to be larger in the listening booth conditions (experiment 2) potentially caused by the mismatch between the acoustic properties of the binaural signals and the acoustics of the listening booth as well as by the lack of a visual distance scale in the listening booth. No influence of the stimulus bandwidth on distance perception was found in either experiment. Overall, the results suggest that the playback room crucially affects distance perception, even in in the case of headphone stimulation.

[a] This chapter is based on Cubick and Dau (2017).

4.1 Introduction

Correct localization of a sound source in a natural listening environment requires both a proper sensation of the source direction and distance from the observer. While most studies have investigated the directional aspect of sound source localization, distance perception has received less attention. Reviews of research on distance perception and the underlying acoustic cues can be found in, e.g., Coleman (1963), Zahorik et al. (2005), and Kolarik et al. (2015). One major finding has been that the perceived distance of the actual (or simulated) sound source can be described by a power function with an exponent below one, i.e., the sound source distance is generally overestimated at close distances and progressively underestimated at large distances. According to Zahorik et al. (2005), the three primary physical cues for distance perception are the intensity, the direct-to-reverberant sound energy ratio and the spectral content of the sound. Sounds are commonly perceived to be farther away when they are lower in intensity, exhibit a lower direct-to-reverberant energy ratio, and contain less energy at high frequencies. However, not only the properties of the acoustic signals entering the ears affect distance perception. Calcagno et al. (2012) investigated the influence of vision on distance perception. Listeners were asked to rate the distance of sounds that were presented from a loudspeaker in the dark, i.e., with no visual information about the sound source available. It was found that the distance was only underestimated when the listeners were not provided any visual reference scale for judging the actual distance. As soon as this scale was supplied (using pairs of LEDs), the listeners actually tended to overestimate the sound source distance, particularly at medium distances between 2 and 5 m.

Two recent studies investigated the influence of the frequency content of the stimuli on sound externalization, i.e., the perception of sounds to be outside the listener's head (in contrast to sound internalization where the sound is perceived to be "inside" the listener's head). Boyd et al. (2012) reported that hearing-impaired listeners with a high-frequency hearing loss provided, on average, lower externalization ratings than normal-hearing listeners in an experiment where individual binaural room impulse responses (BRIRs) were used to auralize the speech stimuli at a certain distance. Furthermore, it was found that the average externalization ratings of normal-hearing listeners dropped to the level found for the hearing-impaired listeners when the stimuli were

lowpass-filtered at 6.5 kHz to simulate a typical hearing-aid bandwidth. A slight reduction of the externalization rating compared to their broadband speech baseline condition was also observed in Catic et al. (2013) for normal-hearing listeners in a similar experiment using stimuli that were lowpass-filtered at 4 kHz. These findings suggested that hearing-impaired listeners generally perceive sounds to be less externalized than normal-hearing listeners, potentially due to the reduced audibility at high frequencies. Also for externalization, the perception of the stimuli can be affected by non-acoustical effects. Several studies have shown that externalization can be reduced when binaural stimuli are delivered via headphones in rooms that differ from the room where the BRIRs have been measured (Werner and Siegel, 2012; Udesen et al., 2015; Gil-Carvajal et al., 2016). If similar effects exist for distance perception as well, experiments in a listening booth should yield different results than corresponding experiments in the room that matches the presented stimuli.

In the present study, inspired by the two externalization studies of Boyd et al. (2012) and Catic et al. (2013), it was investigated whether the perceived distance of sounds is influenced by the high-frequency content of the stimuli in a similar way as externalization and to what extent the playback room affects the results. A binaural technique similar to that in Zahorik (2002a) was employed to auralize the stimuli via headphones. Two experiments were carried out. First, the listening experiment was conducted in the same room where the binaural room impulse responses had been measured. Second, a subset of the same listeners repeated the experiment in a double-walled listening booth to test the influence of the playback room on auditory distance perception.

4.2 Experiment 1: Distance perception in a workshop room

4.2.1 Methods

Listeners

Ten normal-hearing listeners (average age: 29, one female) participated in the study. All listeners had some prior experience with listening experiments. The listeners were informed about the purpose of the experiment and the procedure and gave informed consent prior to testing. The experiments were approved by the Danish Science-Ethics Committee (reference H-3-2013-004).

BRIR measurements

Individual binaural room impulse responses (BRIRs) were measured at nine log-spaced distances (0.43, 0.61, 0.86, 1.22, 1.72, 2.44, 3.45, 4.88 and 6.9 m), as in Zahorik (2002a). The BRIRs were obtained at an azimuth angle of 25°, i.e., close to the 30° angle at which Lounsbury and Butler (1979) achieved the best results in a distance discrimination experiment. The listeners were blindfolded before being guided into the experiment room, a workshop with the dimensions 12.65 x 6.75 x 3.10 m with an acoustic ceiling and an average reverberation time, T_{30}, of about 0.6 s (for a photograph of the room, see Fig. 4.1). During the BRIR measurements, the listeners were seated in a listening chair and provided a small headrest to help keeping the position of the head fixed. The BRIRs were obtained with DPA 4060 lapel microphones positioned at the entrance of the open ear canal with wire hooks that were adapted to each individual ear. The BRIRs were measured using six repetitions of a 5-s logarithmic sine sweep and a deconvolution method (Müller and Massarani, 2001). Directly after the measurement of the loudspeaker responses, the listeners put on a pair of Sennheiser HD800 headphones, without moving the microphones, and the headphone impulse responses were measured with 10 repetitions of a 2-s logarithmic sine sweep. Inverse filters for headphone equalization were derived from the measured headphone responses using a least mean squares time domain inversion method.

Stimuli

For each experimental run, a random sentence from the Danish hearing-in-noise test (HINT) speech corpus (Nielsen and Dau, 2011) was convolved with the BRIRs for all nine distances and with the inverted headphone impulse responses. The resulting auralized signals were band-limited with 6th-order Butterworth filters with cut-off frequencies of 50 Hz and 15 kHz. In addition to these broadband signals, two lowpass-filtered versions of the stimuli were generated with cut-off frequencies at 2 kHz (to simulate a typical high-frequency hearing loss) and 6 kHz (to simulate the limited bandwidth of a hearing aid), respectively. These filters were implemented as 32-tap Hamming-window based FIR filters.

Figure 4.1: Photograph of the listening test setup in a workshop room with dimensions 12.7 x 6.8 x 3.1 m and a reverberation time T_{30} of about 0.7 s. Stimuli were presented via Sennheiser HD800 headphones, and visual distance markers were provided at 2, 4, 6, and 8 m.

Experimental procedure

During the experiment, the listeners were seated at the same position in the same room where the BRIRs were recorded. They were asked to rate the perceived distance of the stimuli on an absolute scale in metres. To facilitate the estimation of the perceived distance, visual markers were provided in the room at distances of 2, 4, 6, and 8 m (see Fig. 4.1). A small computer monitor was placed on a small table in front of the listener. The responses were given via a MUSHRA-like (ITU-R BS.1534-2, 2014) graphical user interface in MATLAB. In the interface, nine playback buttons allowed to play back sound samples that were auralized at the nine different distances for which BRIRs had been measured. The distance rating was given via a corresponding slider with a scale that corresponded to the markers in the room. The stimuli for the different distances were assigned randomly to the sliders, but the bandwidth of the stimuli was kept constant within each individual run. The listeners could listen to the stimuli as often as desired. They were asked to provide a rating in absolute terms according to the scale represented by the markers in the room, which

was also indicated in the user interface. Any stimulus that was perceived inside the head was to be rated at a distance of 0 m. The listeners were asked to rate the perceived distance of the auditory image, i.e., the percept generated by the physical sound source, rather than estimating the distance of the sound source (Blauert, 1997). To familiarize the listeners with the task, all three bandwidth conditions were tested once (for all distances) During the main experiment, all conditions were presented four times. The overall experimental session lasted about one hour per listener.

4.2.2 Results and discussion

Fig. 4.2 shows the average results and standard deviations of the perceived distance ratings for all listeners in the conditions with the 2-kHz low-pass-filtered stimuli (left-pointing triangles), 6-kHz low-pass-filtered stimuli (up-pointing triangles), and broadband stimuli (right-pointing triangles). The light-grey, dash-dotted line shows the veridical values and the dark-grey, dashed line indicates the average value of the distance estimates from Zahorik (2002a).

The average distance ratings increased monotonically with increasing auralized distance in all stimulus conditions. For the two closest auralized distances, the auditory image was, on average, perceived to be closer to the listener than the auralized distance. For medium distances between about one and five metres, the average perceived distance was judged fairly close to the veridical values (light grey, dash-dotted line in Fig. 4.2), but slightly overestimated, whereas the sounds were perceived to be slightly closer to the listener than the actual loudspeaker position in the BRIR measurement for the farthest distance. These findings are in contrast to the results from earlier studies, especially at close distances, where usually an overestimation of the perceived distance has been reported (see Zahorik et al., 2005 for a review). In contrast to the findings of reduced externalization for low-pass filtered stimuli reported in Boyd et al. (2012) and to the findings of increased distance for low-pass filtered signals in Levy and Butler (1978), Butler et al. (1980), and Little et al. (1992), no systematic difference between the different stimulus bandwidths was found in the average distance ratings in the present study.

A linear mixed-effects model was fitted to the data with ‘listener’ as a random effect and ‘distance’ and ‘condition’ as fixed effects. An ANOVA revealed

a significant main effect of 'auralized distance' [F(8,990) = 388.689, p <.0001], but no significant main effect of 'condition' [F(2,990) = 0.088, p = 0.9159] as well as no significant interaction between the two factors [F(16,990) = 1.4804, p = 0.0993]. Therefore, the hypothesized effect of the stimulus bandwidth on distance perception was not found in this experiment.

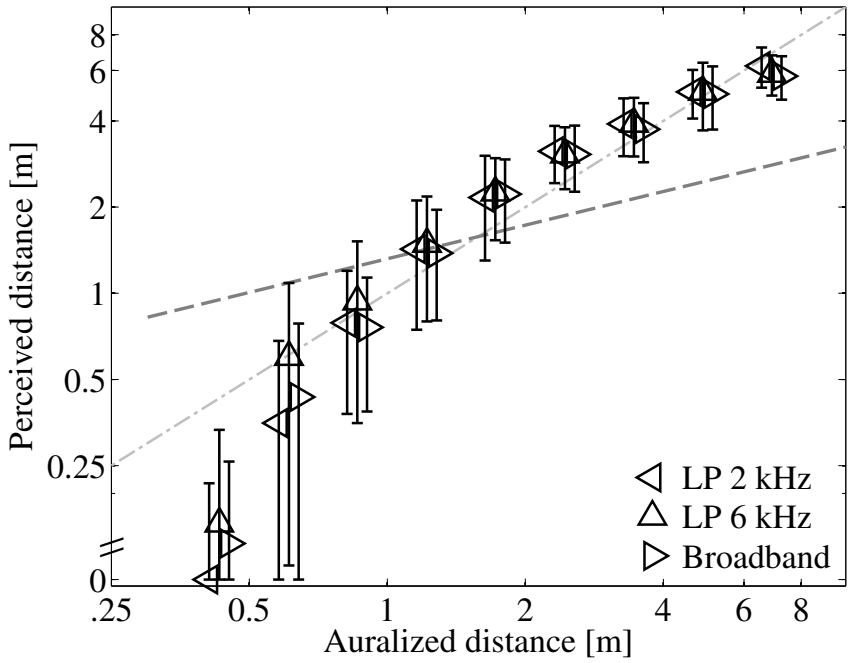


Figure 4.2: Average distance ratings of 10 listeners measured in the workshop room. The light-grey, dash-dotted line indicates the veridical values. The symbols have been slightly jittered along the abscissa to increase readability. The dark-grey, dashed line indicates the fitted values reported in Zahorik, 2002a.

The average distance ratings shown in Fig. 4.2 strongly differ from the average data presented in Zahorik (2002a) (grey dashed line). Even though the BRIRs were measured in a very similar way in Zahorik (2002a) and the present study, the two investigations differed in one important aspect. In the experiment described in Zahorik (2002a), the BRIRs were measured in an auditorium whereas the actual distance estimation experiment was conducted in a listening booth. Thus, the visual and auditory impression of the listening booth did not match the auditory impression of the stimuli presented through the headphones. Furthermore, in the listening booth, no visual reference was provided

for the distance estimation task whereas a visual scale was provided inside the workshop in the present study. Another difference between the current study and the one described in Zahorik (2002a) is the response method. In Zahorik (2002a), a direct scaling method was used, where the listeners entered the distance numerically after each presentation. Thus, the listeners in Zahorik (2002a) judged the stimuli independently whereas the modified MUSHRA interface in the current study allowed the listeners to compare the stimuli, which may have caused the lower variability of the estimates.

4.3 Experiment 2

4.3.1 Rationale

The distance perception results from Experiment 1 differed substantially from the results from other studies. To investigate to what extent such differences could be accounted for by the mismatch between the room in which the experiment was conducted and the one in which the BRIRs were recorded, additional experiments were carried out. Seven of the originally ten listeners from Experiment 1 were available to participate in the same experiment again, using the same stimuli, but this time presented in a double-walled, sound-insulated listening booth.

4.3.2 Methods

The same individual BRIRs as in experiment 1 were used. Also the hardware and user interface were the same as in experiment 1. The only difference was that, this time, the experiments were performed in a double-walled, insulated listening booth instead of the workshop room. Again, the participants were instructed to listen to all stimuli and judge the distance at which they perceived the auditory image on an absolute scale in metres. All stimuli that the listener perceived inside the head should be rated as zero. No visual reference scale was provided. In addition to the conditions from experiment 1, a diotic condition was considered in which the broadband signal for the right ear was presented to both ears. The diotic stimulus was expected to be internalized and, hence, produce distance ratings of zero because they do not contain any binaural information (Catic et al., 2013). This condition was added to test whether the listeners otherwise indeed externalized the stimuli in the experiment. Each

condition was repeated six times, resulting in 24 experimental runs with 9 stimuli each.

4.3.3 Results and discussion

Fig. 4.3 shows the listeners' average distance ratings and standard deviations obtained in the listening booth. The left-pointing triangles indicate the results obtained with the 2-kHz low-pass-filtered stimuli, the upwards-pointing triangles represent the condition with the 6-kHz low-pass-filtered stimuli, and the right-pointing triangles show the results for the broadband condition. The light-grey, dash-dotted line indicates the veridical values and the dark-grey, dashed line represents the fitted function reported in the study of Zahorik (2002a). As in experiment 1, the average distance ratings increased monotonically with increasing auralized distance. Compared to the results obtained in the workshop room (cf. Fig. 4.2), the closest auralized distances were, on average, rated farther away and closer to the veridical values. The farthest auralized distances were perceived slightly closer to the listener than in experiment 1. The resulting average distance function obtained in the listening booth is thus shallower than that obtained in the workshop, indicating a more "compressed" distance perception. This result thus shows a trend towards the data from Zahorik (2002a) (dashed grey line) but still reflects a steeper function than the one found in that study. As in experiment 1, no systematic difference was found between the different bandwidth conditions.

To compare the outcomes of the two experiments, a linear mixed effects model was fitted to the data of the seven listeners who had participated in both experiments with the factors 'auralized distance' and 'condition' and the additional factor 'room' as fixed factors and 'listener' as random factor. Similar to the results from experiment 1, the ANOVA showed a significant main effect of 'distance' [$F(8,1846) = 111.86813$, $p < .0001$], no significant main effect of 'condition' [$F(2,1846) = 0.622$, $p = 0.5371$], a significant main effect of 'room' [$F(1,1846) = 8.352$, $p = 0.0039$] and a highly significant interaction of 'auralized distance' and 'room' [$F(8,1846) = 17.507$, $p < .0001$]. The two remaining two-factor interactions were not significant. The effect of 'room' demonstrates that the distance ratings obtained in the workshop room and the booth are indeed different, even though the stimuli were acoustically identical. The significant interaction between 'auralized distance' and 'room' demonstrates that the ratings were

not only different in the two rooms, but also depended on the auralized distance.

Most listeners indicated that the experiment inside the listening booth (experiment 2) was much more challenging than the experiment carried out in the workshop room (experiment 1). This is consistent with the larger observed variability in the data from experiment 2 in this condition, both at the level of the individual listeners and across listeners. Apart from the very different rooms that caused mismatches of the visual and the auditory room-related information, the ability of listeners to accurately estimate distances without a visual reference may be reduced (Wettschureck et al., 1973; Calcagno et al., 2012) as in the case of the experiments in the listening booth.

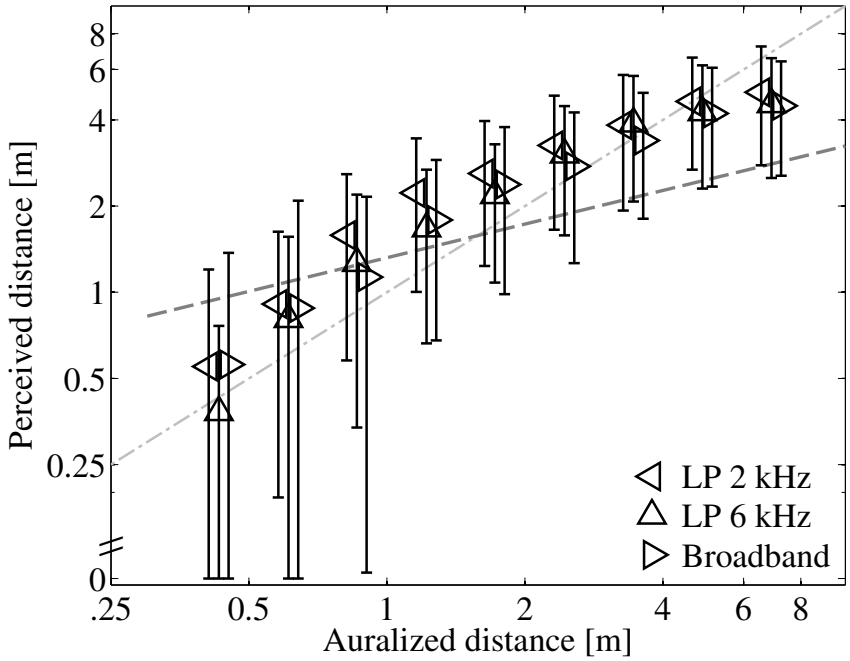


Figure 4.3: Average distance ratings measured in the listening booth. The light-grey, dash-dotted line indicates the veridical values, the dark-grey, dashed line indicates the fitted values reported in Zahorik (2002a).

Panels (a)-(g) of Fig. 4.4 show the average values and the standard deviations of the distance ratings from experiment 2 for the seven individual listeners. The symbols are the same as the ones chosen in Fig. 4.2 and 4.3. The across-listener average is represented in panel h (replot from Fig. 4.2). The results for the diotic control condition are indicated as crosses. The values obtained in the diotic

condition showed a large variability across the listeners. In fact, two groups of listeners showed fundamentally different patterns in their results. Listeners (a) and (e) clearly internalized the diotic stimuli whereas no effect of the diotic stimulus presentation on the distance ratings (relative to the binaural stimulation) was observed for the listeners (c), (d), and (f). The results for the listeners (b) and (g) were in between the two other groups. These listeners typically internalized the stimuli in the case of the four closest auralized distances but perceived the farther ones at a distance even if not quite as far as in the case of the binaural stimuli. Blauert (1997) argued that in many studies on distance perception no clear distinction was made between asking for an estimate of the distance from the listener to the source vs. the distance to the auditory image that is evoked by the source. Even though it was emphasized in the instruction for the present experiment that the listeners should focus on the perceived distance of the auditory image, it appears that some listeners might instead have estimated the source distance, since such an estimate would not necessarily require an externalized percept of the stimulus.

4.4 Overall discussion

One main question of this study was in which way high-frequency information in the stimuli influences the perceived distance of sounds. Neither in experiment 1 nor in experiment 2, any systematic dependency of the perceived distance on the high-frequency content of the stimuli was observed. This is in contrast to the observations of Levy and Butler (1978) and Butler et al. (1980) who found that trains of high-pass filtered noise bursts were judged to be closer to the head than broadband noises whereas low-pass filtered noises were perceived to be further out in space. Also Little et al. (1992) reported greater perceived distances for low-pass filtered noise than for the broadband noise condition. The present results also differ from the findings of the externalization studies of Boyd et al. (2012) and Catic et al. (2013) where reduced high-frequency content was reported to reduce the amount of externalization, i.e., to produce auditory images that are closer to the listener than in the case of broadband signals. The lack of a bandwidth effect in the present study might partly be due to the MUSHRA-based user interface employed in the experiment where all stimuli had the same bandwidth in any given experimental run. The comparative nature of the

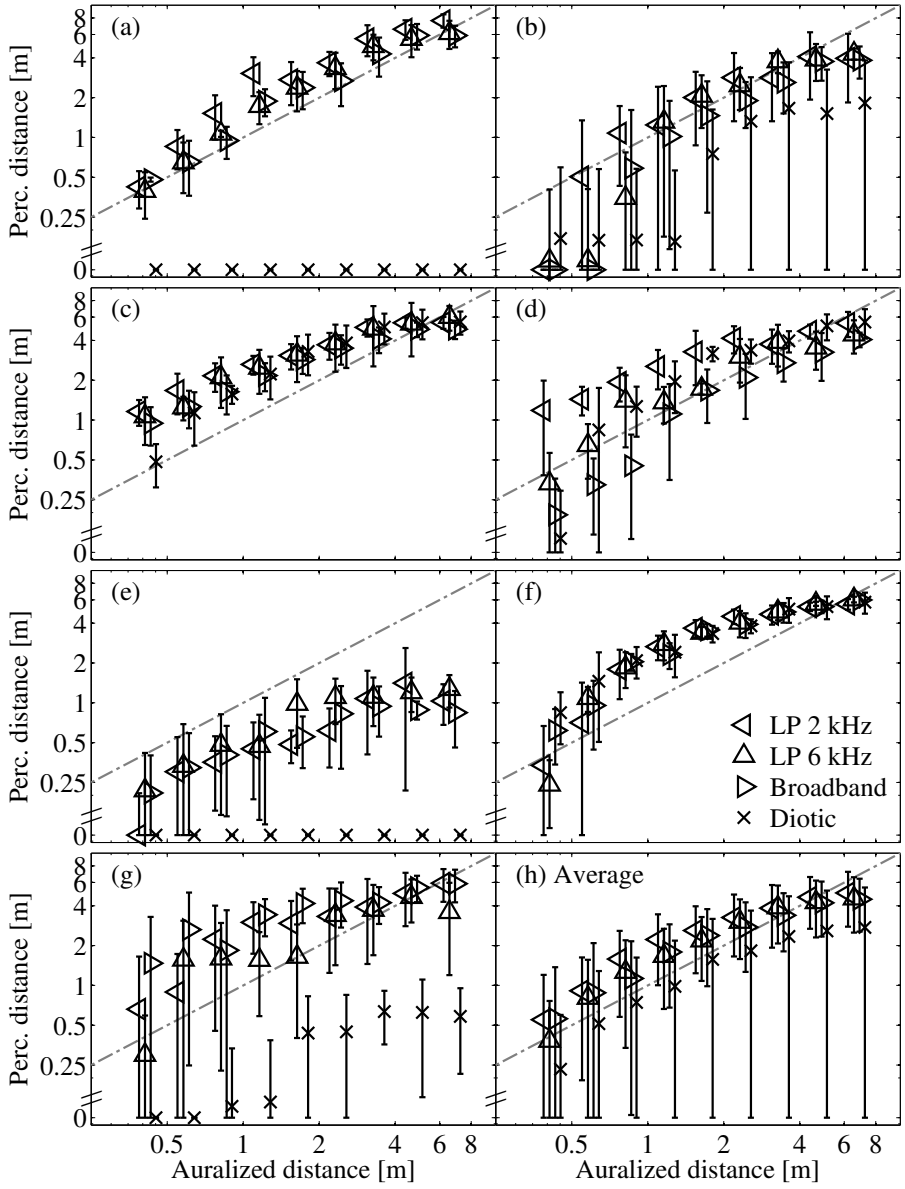


Figure 4.4: (a)-(g): Average and standard deviation of the individual perceived distance ratings obtained in the listening booth in experiment 2. The symbols indicate the 2-kHz lowpass-condition (left-pointing triangle), the 6-kHz lowpass-condition (upward-pointing triangle) the broadband condition (right-pointing triangles), and the diotic condition. (h): Group average, replot of Fig. 4.3 with the addition of the diotic condition (crosses).

procedure may imply that listeners rate the stimuli within a given (bandwidth) condition reliably whereas comparisons across conditions might be less reliable. In future experiments, one might consider randomizing the conditions across runs and/or adding a meaningful reference and anchor to each run to help stabilize the range of judgements and make the results comparable across conditions.

The distance functions presented here are in contrast to some of the findings from the literature on distance perception, where it often has been reported that the distance of a sound source is overestimated at close distances and underestimated at far distances (Zahorik, 2002a; Zahorik et al., 2005). In the present study, stimuli at close auralized distances were typically perceived to be closer to the listener than the auralized distance. Listeners frequently reported a distance of zero, i.e., an internalized auditory image, whereas no "zero" responses were reported in Zahorik (2002a). While the average power function fit from that study, obtained from 33 experiments across 10 studies, described the data from the present study reasonably well at far distances, it did not at all do so at close distances. However, a power function may not capture the observation in the data that sounds are perceived to be closer than the auralized distances, or even inside their heads at close auralized distances. Indications for this can be found in Anderson and Zahorik (2014) where a larger spread in the residuals indicated a less than ideal fit of the model to the experimental data for the closest source distances and where the coefficients of determination were not always high.

Experiment 2 investigated whether presenting binaural stimuli through headphones in a room that differs from the one where the stimuli were recorded results in changes in distance perception. The experiment was inspired by corresponding investigations regarding externalization perception (Werner and Siegel, 2012; Udesen et al., 2015; Gil-Carvajal et al., 2016). While the results still did not show an overestimation at close distances (as in Zahorik, 2002a), the average distance ratings were farther from the listener (than in Exp. 1) and very close to the veridical values. It should be noted though, that all listeners in the current study had performed experiment 1 before experiment 2, which might have biased the results. At least two listeners indicated that they used their experience from experiment 1 to help them rate the distances in experiment 2. Zahorik et al. (2005) reported that the standard deviation estimates of the individual data in Zahorik (2002a) were between 20 and 60% of the tested

distances. In the present experiment, the average intra-subject standard deviation was about 21% of the tested distance in the workshop (in the broadband condition), but about 55% in the listening booth. The larger standard deviation of the distance ratings in the booth also suggests that the listening task inside the booth was more difficult, which agrees with the subjective reports of the listeners.

The general shape of the distance rating functions might also partly reflect an artefact of rating log-spaced auralized distances on a linear scale. If listeners used a strategy to find the farthest stimulus and the closest stimulus and sorted the other stimuli to be distributed roughly equidistantly in between, the function would exhibit a similar curvature. However, in a distance perception experiment with a real sound source in a darkened room, Calcagno et al. (2012) found similar results as in the current study when visual cues were provided. In their experiments, the sound source distances were linearly distributed and the listeners made direct distance judgements in metres. Thus, the shape of the functions seems to reflect the perception of the listeners rather than being an artefact of log-spaced source distances and the response method.

In experiment 2, the comparison of distance ratings for diotic and binaural presentation showed that the listeners could be divided into two groups. One group consistently perceived diotic stimuli to be inside their heads and, hence, reported a distance of zero, whereas the presentation mode did not seem to have any effect in other listeners. The listeners who rated the diotic stimuli as internalized may have based their judgement on the perceived auditory image, while the others might have attempted to estimate the source distance (Blauert, 1997). Whereas the distance of a sound source can be estimated from a monaural signal (von Békésy, 1938; Lounsbury and Butler, 1979) and from monaural cues, like the sound pressure level and the D/R (e.g., Zahorik et al., 2005; Kolarik et al., 2015), previous research suggested that true binaural information is needed for robust auditory externalization (Boyd et al., 2012; Catic et al., 2013; Catic et al., 2015).

As an implication of the ambiguity between source distance estimation and distance perception, experiments with hearing-impaired listeners might prove difficult since some of the listeners might not usually perceive sounds outside

their heads. In fact, one hearing-impaired listener in a similar study on distance perception (Cubick et al., 2014) indicated that he does not usually perceive sounds outside his head in his everyday life, while several listeners in the same study indicated that their judgements were mostly based on the loudness of the stimuli, consistent with results from Akeroyd et al. (2007). It would therefore be valuable to further investigate distance perception as well as externalization in hearing-impaired listeners. A better understanding of their perception would allow for a more specific design of future hearing-aid processing schemes that might help restore a natural spatial perception of the listeners' surroundings.

4.5 Conclusion

Even after almost 150 years of research on distance perception, it still remains unclear how to "correctly" assess distance perception, given the substantial variability of the results across studies. The sensations of distance perception, distance estimation and externalization need to be defined and distinguished carefully. The distance to a sound source may be estimated even when the corresponding auditory image occurs inside the head, but the auditory image can only occur at a distance when the sound is perceived externalized. The results from the two experiments presented here demonstrated that testing distance perception inside a listening booth using binaural stimuli yields different results compared to testing in a more realistic environment, especially when a visual distance scale is available. The results from the present study may provide a valuable basis for investigating the auditory cues underlying the different sensations as well as the consequences of hearing loss and compensation strategies in hearing instruments on the listeners' spatial perception in a given environment.

4.6 Acknowledgements

This work was carried out in connection to the Centre for Applied Hearing Research (CAHR), supported by Oticon, Widex, and GN ReSound, and the Technical University of Denmark.

5

Comparison of binaural microphones for externalization of sounds[a]

Abstract

Ubiquitous availability of media content through portable devices like media players and smartphones has resulted in an immensely increased popularity of headphones in recent years. However, while conventional stereo recordings usually create a good sense of space when listened to through loudspeakers, the sounds tend to be perceived inside the head (internalized) when headphones are used for listening. A more natural perception in headphone listening with sounds being perceived outside the head (externalized) can be achieved when recordings are made with dummy head microphones or with microphones placed inside the ear canals of a person. In this study, binaural room impulse responses (BRIRs) were measured with several commercially available binaural microphones, both placed inside the listeners' ears (individual BRIR) and on a head and torso simulator (generic BRIR). The degree of externalization of speech and noise stimuli was tested in a listening experiment with a multi-stimulus test. No influence was found for the stimulus signal, but the externalization scores were found to be lower for 0° incidence. With all microphones, relatively high externalization scores were achieved, and for all but one microphone, individual BRIRs resulted in slightly better externalization than generic ones.

[a] This chapter is based on Cubick, Sánchez Rodríguez, Song, and MacDonald (2015).

5.1 Introduction

In recent years, headphones have gained a lot of popularity, mainly as a side-effect of mobile devices like laptops, media players, and smartphones becoming more and more omnipresent in our daily lives. This development has given new relevance to an old topic. It has long been known that sounds presented via headphones are often perceived inside the head, i.e., internalized rather than outside the head (externalized), like they usually are in everyday listening situations. References to some early studies that describe internalization or inside-the-head locatedness can be found in Blauert (1997).

A more spacious sound experience with externalized perception of the sound sources is usually desired to create a sense of immersion and reduce listening fatigue that can otherwise occur because of the 180° stereo panorama typically experienced in headphone listening when sounds are perceived internalized. In later years it was found that externalized perception of sounds can be achieved, if the signals at the two eardrums during headphone playback are identical to the signals in the corresponding natural listening situation and, specifically, if the frequency content and temporal relation of the signals at the two ears is correct (Laws, 1973; Wightman and Kistler, 1989a). One way to achieve this is to use a binaural recording technique, i.e., to record sounds directly at the ears of a listener. It was shown that the full spatial information is preserved if the recording is done at any depth in the ear canal or possibly even some millimeters outside of its entrance plane (Hammershøi and Møller, 1996). Recording at the blocked entrance of the ear canal is also valid. This can result in recordings that sound very realistic, especially for the same listener. Similarly, a recording technique can be applied, where the listener is replaced with a mannequin head (and sometimes torso) that is equipped with microphones inside the ears, often referred to as a dummy head microphone or a head and torso simulator (HATS).

If a human head and torso is inserted into a sound field, reflections at the head and in the cavities of the outer ear and diffraction of sound waves around the head will generate a filter that attenuates and amplifies certain frequencies. The coloration of the sound that finally arrives at the eardrum is highly dependent on the direction of the incident sound. Apart from recording directly at the ears of a listener or dummy head, the spatial information can therefore also be described by the head related impulse response (HRIR) in an anechoic sound

field or by a binaural room impulse response (BRIR), which also includes the acoustic properties of the room (Blauert, 1997; Møller, 1992; Vorländer and Summers, 2008), when constant direction of incidence is assumed. HRIRs or BRIRs measured on a dummy head are commonly referred to as generic. Convolution of anechoic sound signals with HRIRs or BRIRs generates a playback signal that often results in a surprisingly realistic acoustic impression of an acoustical scene. Today a number of microphones for binaural recordings are available on the market, ranging from accessories for portable recorders for recording of e.g. rock concerts or soundscapes to tools for sound quality evaluation and scientific work.

Most studies that evaluated the result of binaural recording techniques focussed on localization (e.g., Wightman and Kistler, 1989b; Møller et al., 1999; Minnaar et al., 2001) and they typically reported worse performance when stimuli were generated with non-individual BRIRs. For distance perception, Zahorik (2002b) reported that no difference could be found between conditions with individual BRIRs and non-individual BRIRs measured on another listener's head, and Werner and Siegel (2012) found no influence of using individual or generic BRIRs on externalization. Begault and Wenzel (1993) on the other hand found very high percentages of stimuli being perceived internalized for anechoic speech stimuli and non-individual HRIRs of a human head.

This study investigated the degree of externalization that could be achieved with five different commercially available binaural microphones and a dummy head using a virtual auditory space technique. In a listening experiment, eight normal-hearing listeners rated the perceived externalization of sounds presented via headphones for all microphones for four different source positions and two different types of stimuli in a multi-stimulus test paradigm.

There were four main research questions: 1) Does the stimulus material influence the perceived externalization? 2) Does the externalization percept depend on the incidence angle? And most importantly 3) Do the different microphones yield different externalization ratings? and 4) Does it make a difference whether individual or generic BRIRs are used?

5.2 Methods

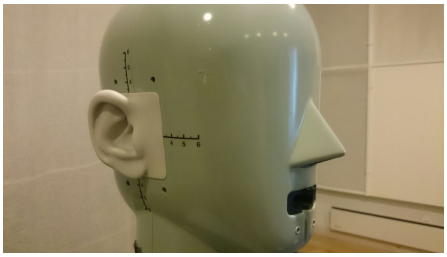
5.2.1 Microphones

Five different pairs of commercially available microphones were chosen for the comparison. In addition, the internal microphones of the HATS were used as a representative for dummy head recording techniques. An overview of the microphones and their background noise levels can be found in Table 5.1. All noise levels except for the HATS internal microphones was measured in an anechoic chamber at DTU, the values for the HATS were taken from the data sheet.

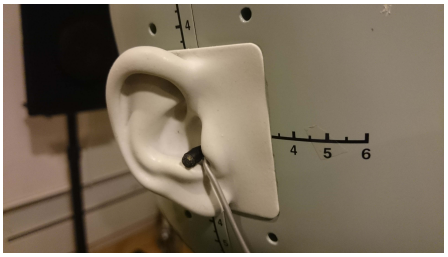
Table 5.1: Type, alias, and background noise level of the microphones used in this study. Note, that all given noise levels were measured except for the one of the HATS, which was taken from the data sheet.

Microphone	Alias	Noise level (L/R)
B&K HATS 4128- C-002	HATS	19.0/19.0 dB(A) 21.3/21.3 dB SPL
B&K 4101-A	4101	22.4/22.6 dB(A) 28.4/28.4 dB SPL
B&K 4965	4965	23.3/23.1 dB(A) 29.7/29.3 dB SPL
DPA 4060	DPA	22.5/22.7 dB(A) 35.3/36.4 dB SPL
Roland CS-10EM	Roland	27.4/27.3 dB(A) 30.4/30.2 dB SPL
Sound Professionals MS-TFB-2	SProf	25.0/25.3 dB(A) 31.9/31.9 dB SPL

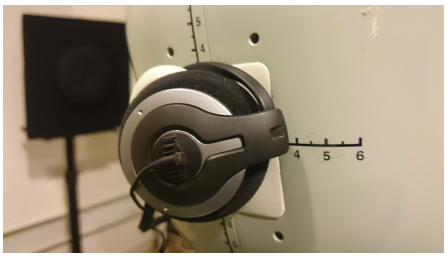
Figure 5.1 shows photographs of the binaural microphones under test mounted on a HATS. All microphones were used with the mounting solution provided by the manufacturer except for the DPA 4060 (Fig. 5.1d), which are originally clip microphones made for stage use. These microphones were positioned on the listeners' ears by means of a wire hook that was individually adjusted to place the microphone as close as possible to the entrance of the ear canal. Note that due to the differences in construction, the position with respect to the ear canal was quite different for the respective microphones.



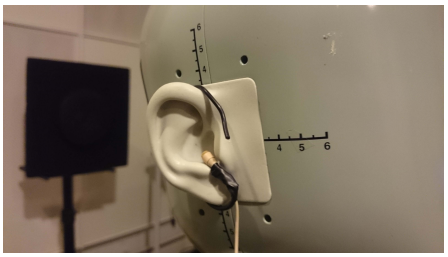
(a) B&K HATS 4128



(b) B&K 4101

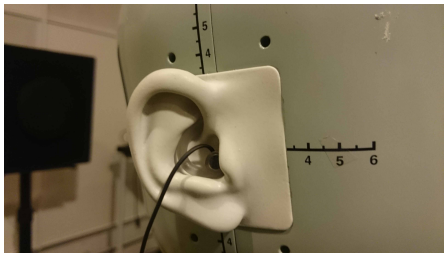


(c) B&K 4965



(d) DPA 4060

(e) Roland CS-10EM



(f) Sound Prof. MS-TFB-2

Figure 5.1: Binaural microphones used in this study mounted on a B&K HATS. Note the different positions of the microphones on the pinnae of the HATS. Especially the B&K 4965 microphones (c), but also the B&K 4101 (b) and the Roland microphones (e) are placed at a position clearly outside of the ear canal, which is less then optimal, because the transfer function from the microphone to the ear drum is not independent of direction (Hammershøi and Møller, 1996).

5.2.2 Listeners

The listening experiments were performed by eight normal-hearing listeners (aged 21-25, 2 female) with listening thresholds better than or equal to 20 dB HL on both ears at all of the audiometric frequencies from 125 Hz to 8 kHz. Five listeners were naïve, three listeners had participated in listening experiments before.

5.2.3 BRIR measurements

Individual BRIRs were measured for each listener for all five microphone pairs in an IEC listening room (IEC 60268-13, 1985) with an average reverberation time T_{30} of about 0.3 s and a volume of about 100 m^3. For each set of microphones, BRIRs were measured for four loudspeakers Dynaudio BM6P at azimuth angles of 0, 25, 60, and 90° and a distance of 2.5 m using six repetitions of a 5-s logarithmic sine sweep and a deconvolution method according to Müller and Massarani (2001). Furthermore, generic BRIRs were measured under the same conditions on the HATS for all five microphones and the internal microphones of the HATS. After the measurement of the BRIRs from the loudspeakers, a pair of Sennheiser HD 800 headphones was carefully placed on the head without moving the microphones and headphone impulse responses (HPIR) were measured to the respective microphones with ten repetitions of a 2-s logarithmic sine sweep. The inverse filters for the headphone equalization were derived from the measured impulse responses using a least means squares time domain inversion method. The listeners were instructed to keep the position of head as fixed as possible.

5.2.4 Stimuli

In the experiments, two different signals were used, sentences from the Danish HINT speech test corpus (Nielsen and Dau, 2011), and trains of pink noise bursts (5 bursts of 200 ms with a pause of 300 ms in between, 5-ms Hanning ramps at the beginning and end of each burst). For each experimental run, ten stimuli were generated by convolving the signal with the individual and the generic BRIRs for the five microphones. As a control condition, the signal was also convolved with the BRIR measured with the internal microphones of the HATS. The dry signal served both as a reference (played back via loudspeaker) and an anchor (played back diotically via headphones). To avoid loudness as a cue, the reference was adjusted to subjectively match the loudness of the other signals by two of the authors. In total, 13 stimuli were used within each experimental run. The 11 signals involving BRIRs were additionally filtered with the inverse filters derived from the measured HPIRs. All auralized signals were band-limited between 50 Hz and 15 kHz with 6^{th} order Butterworth filters.

5.2.5 Experimental procedure

During the experiment, the listeners were seated in the same room at the same position, where the BRIRs had been measured (see Figure 5.2 for a photograph of the setup.). They controlled the listening experiment via a graphical user interface in Matlab (cf. Figure 5.3). The procedure was a modified MUSHRA test (ITU-R BS.1534-2, 2014). Each of the stimuli described above was randomly assigned to one of the 13 buttons (A-M), which start the audio playback. The externalization rating for each stimulus was reported via the corresponding slider. Within each experimental run, the signal and the loudspeaker angle were kept constant. When speech stimuli were used, the same sentence was used for all stimuli within one experimental run. The angles of the loudspeakers were randomized across the experimental runs

Figure 5.2: Photograph of the experimental setup with a listener at the listening position inside the IEC listening room. The four loudspeakers were positioned at 0, 25, 60, and 90° at a distance of 2.5 m. The listeners controlled the experiment via a graphical user interface on a small screen using a wireless mouse.

The listeners were instructed to judge the degree of externalization on a scale from 0 to 100, where 0 means that the sound was perceived inside the head and 100 that the sound was perceived at the position of the loudspeaker. They were instructed to rate the hidden reference as 100 (if found). To help the judgement, a five-point scale similar to Boyd et al. (2012) and Catic et al. (2013) was supplied ranging from “Inside my head” (0), “Near my head” (25), “Close to me” (50), “Close to the loudspeaker” (75), and “At the loudspeaker” (100).

The listeners could listen to the stimuli as often as needed in order to make a

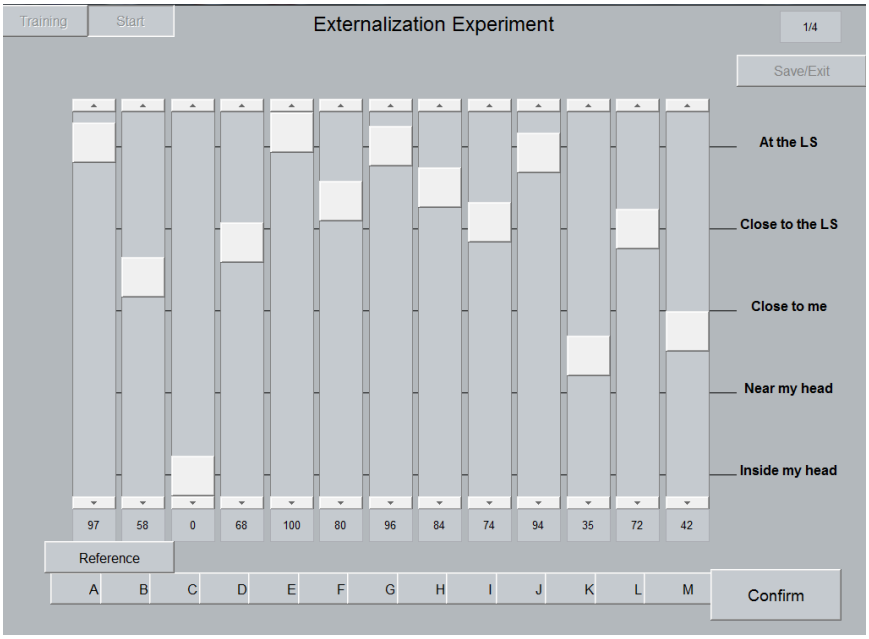


Figure 5.3: Graphical user interface for the listening experiment. The 13 buttons (A-M) allowed for playing back the stimuli (5 individual and 5 generic BRIRs for the microphones under test, the internal microphones of the HATS, the hidden reference, and the anchor in random order). The externalization rating was entered via the corresponding slider.

judgement. Once they rated all stimuli, hitting “Continue” started the next run. Before the experiment, the listeners performed two training runs with stimuli presented from 0° and 60°. The actual experiment consisted of eight runs (4 angles, 2 stimuli). The whole experimental session took about 40 minutes per listener.

5.2.6 Statistics

To test the results, a repeated measures Analysis of Variance (rANOVA) was carried out with “Angle”, “Microphone”, and “Stimulus” as within-subject factors. Post-hoc pairwise t-tests were carried out for all factors that showed a significant effect in the rANOVA.

5.3 Results

5.3.1 Influence of the stimulus signal

Fig. 5.4 shows the externalization rating averaged across all listeners, microphones and loudspeaker positions for the noise bursts (left) and the speech stimuli (right). To increase readability, the plot only shows the upper half of the response scale. The ratings for the reference and the anchor were excluded. The average rating for the noise signal was 66.3, the rating for the speech signal was 67.2 or slightly below "Close to the loudspeaker". The choice of the stimulus signal thus did not seem to have an influence on the perceived externalization, which was confirmed in the rANOVA, where the main factor "Stimulus" showed no significant effect [$F(1,7) = 0.069$, $p = 0.8$].

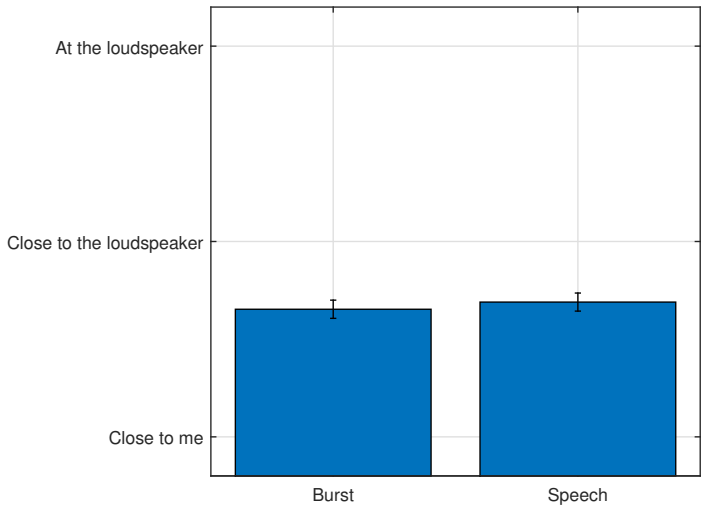


Figure 5.4: Average externalization rating for noise (left) and speech (right) stimuli. Error bars indicate ± one standard error. Note that only the upper half of the scale is shown.

5.3.2 Influence of the loudspeaker angle

Fig. 5.5 shows the average externalization rating of all listeners for all microphones over the four loudspeaker angles 0, 25, 60, and 90°. The ratings increase with angle from 61.6 at 0° over 65.8 at 25°, 69.7 at 60° to 70 at 90°. The rANOVA showed a significant effect of the factor "Angle" on the externalization rating

[F(3,21) = 3.228, p = 0.043], the post-hoc analysis revealed that the only significant differences are found between the rating for 0° and the ratings for 60° and 90°. This was expected, because front-back confusions and internalization were reported to be most common for directions close to the median plane (e.g., Begault and Wenzel, 1993), where the differences between the ear signals are small. A recent study, however, did not find a significant difference on externalization when presenting virtual stimuli from 0, 90, or 180° (Udesen et al., 2015).

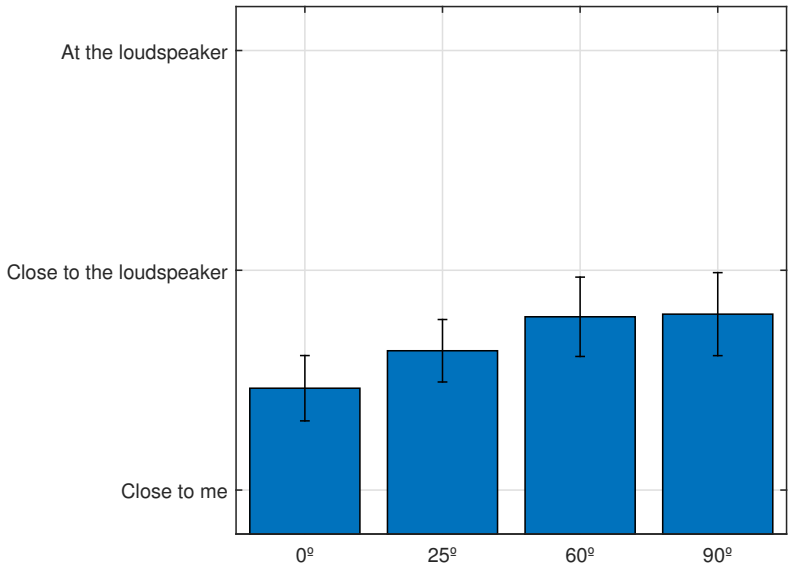


Figure 5.5: Average externalization rating for the four different loudspeaker angles. The error bars indicate ± one standard error. Again, only the upper half of the scale is shown.

5.3.3 Influence of the microphone type

Among the generic BRIRs, the highest externalization scores were obtained with the HATS internal microphones with an average value of 75.5, closely followed by the DPA microphones (74.5). Generic BRIRs measured with all other microphones resulted in significantly lower average externalization ratings, as confirmed by the post-hoc analysis.

All stimuli that were generated using individually recorded BRIRs were on average judged as fairly well externalized (with ratings of 68.5 for the B&K 4101,

		individual BRIR					generic BRIR					
	Mic	4101	4965	DPA	Roland	SProf	4101	4965	DPA	Roland	SProf	HATS
individual	4101						✓		✓			✓
	4965						✓	✓				✓
	DPA											
	Roland						✓					
	SProf						✓	✓	✓			✓
generic	4101	✓	✓		✓	✓			✓			✓
	4965		✓			✓			✓			✓
	DPA	✓				✓	✓	✓		✓	✓	
	Roland								✓			✓
	SProf								✓			✓
	HATS	✓	✓			✓	✓	✓		✓	✓	

Table 5.2: Results of the post-hoc analysis. The checkmarks indicate pairs for which a significant difference was found ($\alpha = 0.05$)

66.8 for the B&K 4965, 69.1 for the DPA 4060, 69.1 for the Roland, and 69.9 for the Sound Professionals). The ratings were thus just below the "Close to the LS" category. The post-hoc analysis showed that none of the microphones yielded significantly different externalization scores when the BRIRs were measured individually. For the individual BRIRs, there is therefore no statistical evidence that one of the microphones yields better results than the others. For a full overview over the results of the post-hoc analysis, see Table 5.2.

5.3.4 Individual vs. generic BRIRs

Only for the 4101 and the 4965 a significant difference in the externalization ratings was found between the individual and the generic BRIR for the same microphone. In both cases, the individual BRIR yielded higher ratings.

Another rANOVA was carried out to further analyze the effect of the individual versus generic BRIRs. The HATS was excluded from the calculation and the within- subject factor "Individual/Generic" was added. The results showed again that the angle has a significant effect, whereas the stimulus signal does not. The main effects of "Microphone" and "Individual/Generic" were not significant but they did show a trend. Furthermore, the interaction between "Microphone" and "Individual/Generic" was found to be significant [$F(4,28) = 3.305$, $p = 0.024$]. A look at the data reveals, that this interaction occurred because the DPA 4060 microphones yield higher externalization ratings for the generic than for the individual BRIRs, whereas for all other microphones

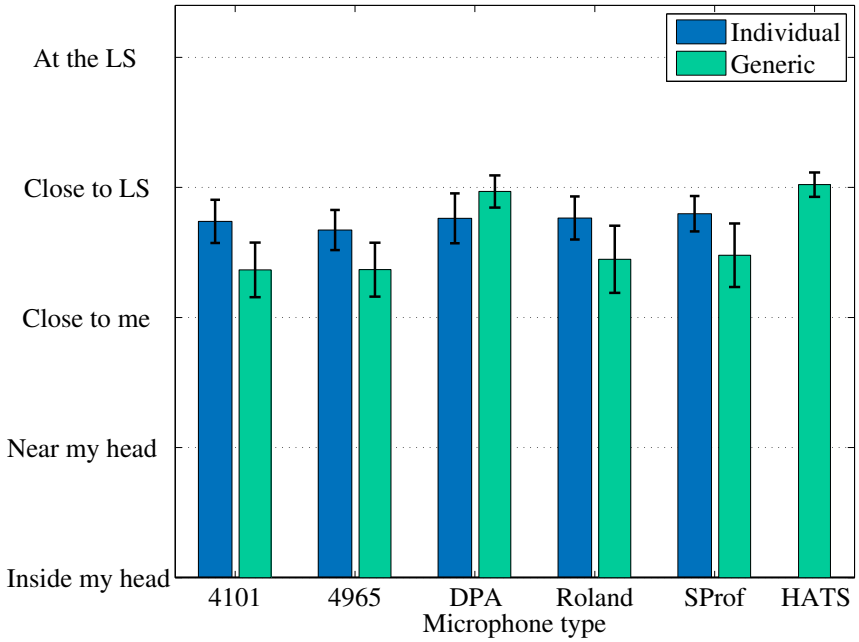


Figure 5.6: Average externalization rating for five different pairs of binaural microphones, each for individual and generic BRIRs.

the individual BRIR yielded higher externalization ratings (ca. 8% on average). When the DPA microphones were excluded from the statistical analysis, the main effect of "Individual/Generic" was significant [$F(1,7) = 8.151$, $p = 0.025$]. This could be explained by the fact that it was quite easy to accidentally move the DPA microphones during the measurement of the individual BRIRs due to the way they were attached to the ear, which could lead to less precise BRIR measurements and potentially incorrect equalization filters. It could be suspected that with a more optimal and stable placement of the microphones on the ears, the individual BRIRs would lead to higher externalization scores for these microphones as well.

5.4 Discussion

The externalization ratings found in this study were generally quite high with most of the ratings occurring somewhere between "Close to me" and "Close to the loudspeaker" (grand average: 66.8), indicating that the auralization technique used here works well. This corresponds well with the subjective impres-

sion, where most sources were clearly externalized and it was difficult to make out a clear difference between the stimuli. It seemed a bit surprising that no bigger difference was found between generic and individual BRIRs, even though the individual BRIRs yielded higher average externalization ratings for all but one microphone. What might have helped in the current study, was the fact that the experiments were performed in the same room as the BRIR measurements, since some recent work has pointed out that the auditory image is usually perceived most externalized when the playback room and the recording room are identical (Udesen et al., 2014; Gil Carvajal, 2015).

Note that the basic assumption of binaural technology has been violated in some of the measurements. The basic assumption is that the transfer path of the sound from a sound source to the eardrum can be divided into a directional-dependent and a directional-independent part and that the perception of an acoustic scene simulated via binaural technology will correspond to the one in the real scene, if the sound pressure is reproduced correctly at the eardrum or at a point at the ear, where the frequency response is independent of direction (Hammershøi and Møller, 1996). This is the case inside the ear canal, but not (far) outside it. Therefore, Equalizing the headphones relative to a microphone position outside the ear canal very likely introduces sound coloration, a disturbed localization, and might also cause a reduced externalization percept. Especially for the B&K 4965 Microphones, but also for the B&K 4101 and the Roland microphones, this was expected to be problematic, because the microphones are positioned rather far outside the ear canal. Interestingly, this “wrong” equalization did not seem to have a big impact on the externalization rating, since the ratings were not significantly different from the ones for the other microphones. It might, however, be one of the reasons why both the 4101 and the 4965 scored lower average externalization ratings than the DPA 4060, even though all three microphones are based on the same microphone capsules. One area where the “wrong” placement of the microphones might have an influence are attributes of sound quality. Especially in some conditions with noise stimuli, timbral differences between the microphones were quite obvious. In future investigations, this and other perceptual attributes like compactness or localization of the auditory image should be considered, because that would allow for a more complete understanding and clearer ranking of the binaural microphones.

Looking at the results it should also be considered that the microphones under test will most likely be used in very different ways in practice. Someone, who invested in a very expensive HATS, will most likely be aware of the necessity to equalize the headphones, the amateur who occasionally records a concert of a local rock band on a cheap portable recorder will most likely not and just listen to the recording as it is through whatever headphones available. If these different approaches had been considered in the listening experiments, some larger differences might have been found in the externalization ratings between the microphones.

Considering that most of the stimuli were perceived well externalized, using an omnidirectional room impulse response for the anchor signal might have resulted in a wider range of judgements, whereas the anechoic signal used in this study, being very different from the other stimuli, might have limited the range of responses that has been used by the listeners.

5.5 Conclusion

Five commercially available types of binaural microphones were evaluated with respect to the achieved amount of externalization. In a listening experiment with eight listeners, the average externalization scores were relatively high (just below “Close to the LS”). With the exception of the DPA 4060, individual BRIRs resulted in higher ratings than generic BRIRs. However, the differences were surprisingly small. This indicates that, if only externalization is considered, BRIRs measured on dummy heads might well be sufficient in many situations to generate a more natural sound experience with sources perceived well outside the head. This argument is supported by the fact that the stimuli that used BRIRs measured using the internal microphones of the HATS consistently yielded the highest average externalization scores. Using either speech or pulsed noise stimuli did not change the overall judgement. As found by others before, good externalization seems most difficult to achieve for frontal directions, which is reflected in the lower externalization scores measured for 0° incidence.

However, externalization scores are only one aspect in judging the performance of binaural microphones. As a next step, other outcome measures should be considered as well. It seems especially crucial that the microphones do not in-

troduce coloration, that they allow for natural localization, and that the auditory image is compact.

6

Spatial Hearing with Incongruent Visual or Auditory Room Cues[a]

Abstract

In day-to-day life, humans usually perceive the location of sound sources as outside their heads. This externalized auditory spatial perception can be reproduced through headphones by recreating the sound pressure generated by the source at the listener's eardrums. This requires the acoustical features of the recording environment and listener's anatomy to be recorded at the listener's ear canals. Although the resulting auditory images can be indistinguishable from real-world sources, their externalization may be less robust when the playback and recording environments differ. Here we tested whether a mismatch between playback and recording room reduces perceived distance, azimuthal direction, and compactness of the auditory image, and whether this is mostly due to incongruent auditory cues or to expectations generated from the visual impression of the room. Perceived distance ratings decreased significantly when collected in a more reverberant environment than the recording room, whereas azimuthal direction and compactness remained room independent. Moreover, modifying visual room-related cues had no effect on these three attributes, while incongruent auditory room-related cues between the recording and playback room did affect distance perception. Consequently, the external perception of virtual sounds depends on the degree of congruency between the acoustical features of the environment and the stimuli.

[a] This chapter is based on Carvajal, Cubick, Santurette, and Dau (2016).

6.1 Introduction

The impression of auditory space occurs on the basis of auditory cues provided by sound waves arriving at each ear, directly from the source, and after bouncing off the surfaces of the environment (Blauert, 1997; Erulkar, 1972). Time and intensity differences between the two ear signals determine, in most cases, the azimuthal localization of sounds (Strutt, 1907; Middlebrooks and Green, 1991), whereas the perception of elevation is mainly associated with the direction-dependent filtering effect of the outer ear (Roffler and Butler, 1968). Distance perception has been shown to rely mostly on intensity, the ratio between the energy of direct and reflected sound, and the frequency content of the signal (Mershon and King, 1975; Zahorik et al., 2005; Plack, 2013; Bronkhorst and Houtgast, 1999). In an acoustic environment listeners are exposed to physical stimuli (sound events) that lead to perceived auditory images (auditory events). However, the same sound event can yield different auditory events due to cognitive factors and cross-modal processing (Blauert, 2005). For instance, the spatial impression can be affected by multisensory interaction, particularly between vision and hearing. Several studies indicated that there is a combined perception that considers inputs from the two sensory modalities (McDonald et al., 2001; Alais and Burr, 2004; Stein and Stanford, 2008; Jack and Thurlow, 1973). This knowledge has been exploited in different applications, such as video gaming and multimedia reproduction in connection with virtual sound techniques that enable the generation of externalized sound images via headphones (Blauert, 2005; Vorländer and Summers, 2008), such that real-world sound sources are convincingly reproduced (Wightman and Kistler, 1989a; Kulkarni and Colburn, 1998).

Sound externalization refers to an out-of-head position for a given auditory event. Externalization can be defined as accurate when the auditory event is properly localized within a confined space in terms of distance and direction (Hartmann and Wittenberg, 1996). In contrast, internalization refers to an in-head auditory event position, with sound perceived between the ears or lateralized, without a projection of the auditory image in space (Plenge, 1974). This typically occurs during the reproduction of acoustic signals that have been simulated or recorded without considering the acoustic filtering due to diffraction from the pinna, head, and torso (Kim and Choi, 2005). Such filtering is

described by the head-related transfer function (HRTF), an accurate representation of which can enable listeners to perceive externalized sound images. Individualized HRTFs can be recorded from each listener. Alternatively, HRTFs can be synthesized or obtained from dummy heads, which results in decreased localization accuracy (Wenzel et al., 1993). When the recording environment is not anechoic, the HRTFs also contain information about the acoustical properties of the environment and the corresponding impulse responses are called binaural room impulse responses (BRIRs)(Blauert, 2005; Kleiner et al., 1993). When the playback room and the room in which the BRIRs are recorded differ, the listener may receive incongruent room-related cues from the headphone reproduction and the listening environment (Udesen et al., 2015). Here, we aimed to test whether such incongruent room cues affect externalization accuracy, defined as the degree of coincidence between the virtually reproduced sound event and the perceived auditory event in terms of distance, direction, and compactness (Hartmann and Wittenberg, 1996).

In our experiments, we investigated whether the externalized perception during sound reproduction breaks down in certain environments where the spatial acoustic information from the playback signal and the cues obtained about the room are incongruent. We asked eighteen naive listeners to rate three spatial attributes of real sound sources (distance, azimuthal direction, and compactness) independently in order to evaluate sound externalization of virtual stimuli delivered via headphones. Although previous studies have addressed externalization through headphones as an overall percept (Kim and Choi, 2005; Begault et al., 2001; Boyd et al., 2012; Catic et al., 2013), none investigated which specific cues arising from the room are most important for externalization, and how they might be affected by a change of the listening environment.

Three rooms were used for the experiments. A standard IEC listening room (IEC 60268-13, 1985) was the Reference room for the listening test, in which individual BRIRs were recorded. As room acoustic parameters, such as reverberation time, are generally related to the volume of the room (Kuttruff, 2009), a smaller and larger room were also used in the listening tests. However, these rooms were acoustically treated such that the smaller room (Reverberant Small) was much more reverberant and the larger room (Dry Large) was anechoic, i.e., not reverberant at all. Thus, the incongruence between the spatial cues of the

Reference and the other test rooms differed depending on whether the listener considered the room-related visual cues (i.e., difference in room volume) or auditory cues (i.e., difference in reverberation time).

For the listening experiment, the participants were divided into two groups and provided either auditory or visual awareness of the test rooms, while input from the other modality was limited as much as possible (Fig. 6.1). One group of listeners could see the rooms but did not receive any auditory stimuli except the processed speech sentences. The other group entered the rooms blindfolded and performed the ratings in the dark, but was provided room-related auditory cues from a loudspeaker emitting noise bursts every 5 s. All subjects then performed the same experiment with both visual and auditory room cues available. The target signals were anechoic speech sentences convolved with BRIRs obtained for each listener individually before testing. All BRIRs were recorded for seven source positions in the reference room, while the test subjects wore both earplugs and blindfolds to avoid a priori knowledge of the room. The listeners evaluated the three externalization attributes using subjective scales (Fig. 6.1b). Each stimulus was rated twice in each condition per room. Directional ratings were based on the selection of one out of twelve possible options arranged using a clock style notation. Compactness and distance perception were rated using a scale ranging from 0 to 5. For compactness, 0 corresponded to the most compact and 5 to the broadest perception of a sound. For distance, 0 indicated an auditory image perceived inside the head, 4 corresponded to a percept at the loudspeaker, and 5 beyond the loudspeaker position. During the experiments, four loudspeakers were visible at 0°, 30°, 90°, and 330° (XII, I, III, and XI o'clock, respectively, indicated by loudspeaker pictograms in Fig. 6.1b), while sounds were simulated for all seven recorded source positions (circled in red in Fig. 6.1b). An additional anechoic speech stimulus was presented diotically as a perceptual anchor, which was expected to be perceived inside the head due to the lack of spatial information in the signal (Hartmann and Wittenberg, 1996). The percepts from the simulation were compared to the percepts from a physical representation, which was achieved by delivering a randomly selected anechoic speech stimulus directly from the loudspeaker placed at III o'clock in the reference room. In the case of an ideal binaural simulation, this signal would be acoustically identical to the corresponding headphone signal and the two should be indistinguishable.

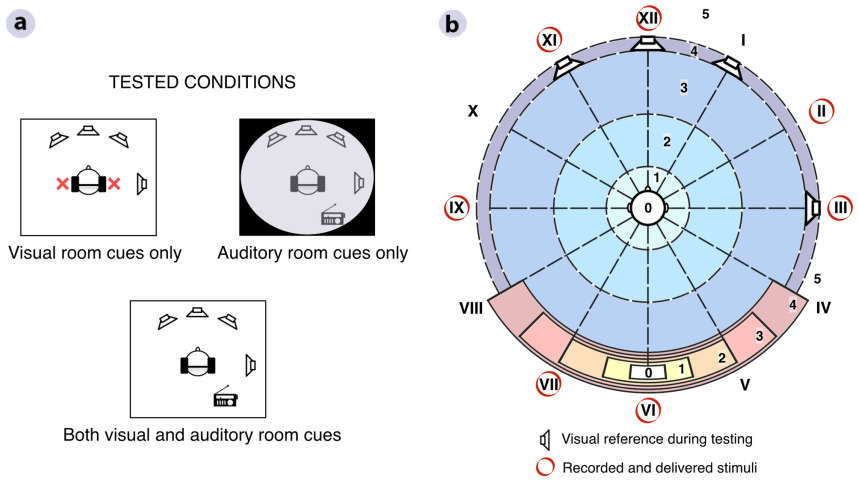


Figure 6.1: Experimental conditions and setup. a) Illustration of the three experimental conditions: visual room cues, auditory room cues, and both visual and auditory room cues. b) Loudspeaker setup and subjective rating scales used in the experiments. For azimuthal direction judgements, listeners could provide ratings from I to XII. For distance and compactness judgements listeners could provide ratings from 0 to 5.

6.2 Results

6.2.1 Effect of mismatch between playback and recording room

In order to present the results using a similar metric for the three attributes of interest, a criterion for "correct" externalization was used. A "correct" response was defined as an auditory event perceived coincident in space with the physical sound event that would be produced by a loudspeaker at the corresponding position used for the BRIR recording. In order to take the limitations imposed by the virtual sound reproduction into account, the criterion was based on a comparison of listener ratings for headphone vs loudspeaker presentation. The ratings for position III showed that the anechoic signal played back through a loudspeaker in the reference room and the corresponding headphone signal yielded very similar distance ratings (Fig. 6.A.1a; Loudspeaker: $M = 3.81$, $SD = 0.40$; Headphones: $M = 3.97$, $SD = 0.17$) and azimuthal direction (Fig. 6.A.1b; Loudspeaker: $M = 3.00$, $SD = 0.00$; Headphones: $M = 2.97$, $SD = 0.17$), whereas sounds delivered via headphones resulted in a wider range of compactness

ratings compared to sounds presented from the loudspeaker (Fig. 6.A.1c; Loudspeaker: M = 0.22, SD = 0.42; Headphones: M = 1.11, SD = 1.12). Based on these results, the criterion for "correct" responses was defined as a score of 4 for distance, 0 to 1 for compactness, and localization at the exact azimuthal direction.

Fig. 6.2 shows the percentages of correct responses for the three externalization parameters distance (Fig. 6.2a), azimuthal direction (Fig. 6.2b), and compactness (Fig. 6.2c). Ratings are shown using different colours for each playback room and are presented separately for the three conditions tested. The analysis for the condition in which both visual and auditory room cues were available (Fig. 6.2a, left) showed that sound externalization in terms of perceived distance is indeed room dependent. A linear mixed-effects-model analysis of variance (ANOVA) with Room, Cue, and Position as fixed factors and Listener as a random factor (Tab. 6.A.1) revealed a significant effect of Room [$F(2,1456) = 94.6$, $p < 0.001$]. Post hoc multiple comparisons using Tukey's honest significant difference test also revealed significant differences across all playback rooms in the condition with both visual and auditory cues (Tab. 6.A.3, $p < 0.001$). Listeners generally perceived sounds to be closer to their heads in both incongruent rooms than in the reference room. The reverberant room (Reverberant Small, blue bars) yielded the lowest externalization scores with approximately 18% of correct distance ratings. This confirms that a mismatch between recording and playback room adversely affects the externalization of binaural speech stimuli in terms of perceived distance.

Here, such a mismatch generally affected the distance ratings more for front and back positions than for lateral positions (Fig. 6.3), as reflected by a significant effect of Position [$F(2,1456) = 28.8$, $p < 0.001$] and a significant interaction between Room and Position [$F(12,1456) = 4.9$, $p < 0.001$] (Tab. 6.A.1). This was confirmed by significant differences for all post-hoc multiple comparisons between front-back (XII and VI) and lateral (III and IX) positions in the Reverberant Small and Dry Large rooms, while none of the differences between positions XII and VI (frontal) and between positions III and IX (lateral) were significant (Tab. 6.A.4). The post-hoc analysis also confirmed that the effect of room mismatch was more pronounced for the Reverberant Small room, where it was significant for all positions except position III, than for the Dry Large room, in

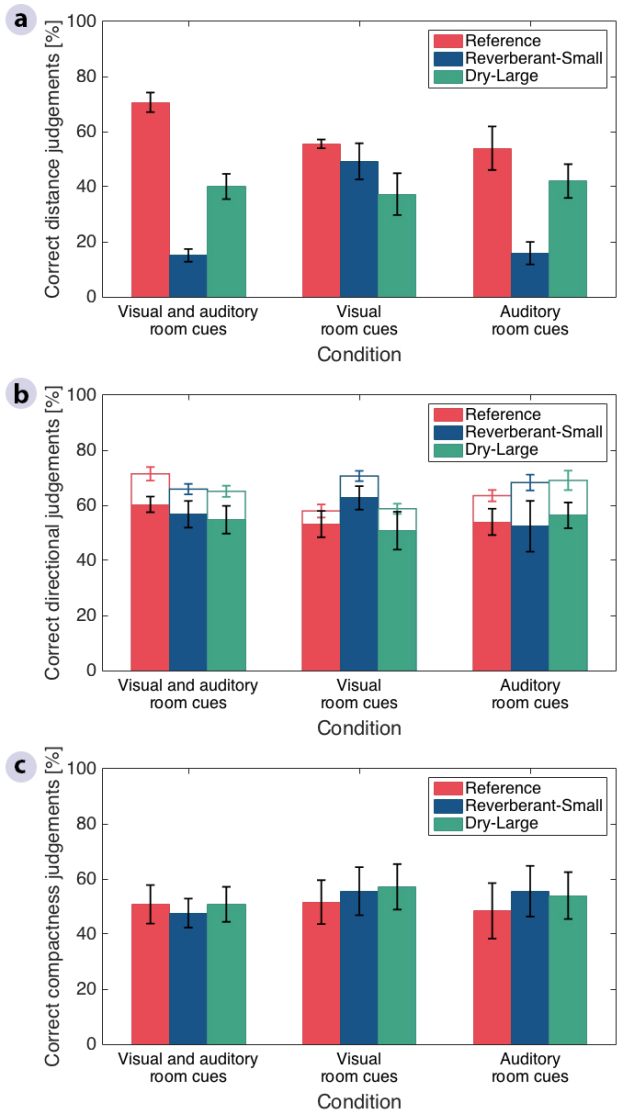


Figure 6.2: Total correct judgements of each externalization parameter in the reference room and the two rooms that were incongruent with the reference. a) Correct distance ratings. b) Correct azimuthal direction ratings. c) Correct compactness ratings. Percentages represent the across-listener mean calculated over the total number of correct judgements per listener across positions, while error bars show the standard error of the mean across listeners. In b) filled bars represent percentages of correct directional judgements, while the height of empty bars represent the same percentages when counting front-back confusions as correct. Front-back confusions were determined from the number judgements in a hemisphere that differed from that of the stimulus position over the total number of presentations.

which it reached significance only for positions XI and XII (Tab. 6.A.5). Finally, the distance ratings varied more across positions in the Dry Large and Reverberant Small rooms than in the Reference room (more significant between-Position differences in Tab. 6.A.4), and were overall lowest for sounds delivered from the VI o'clock position (Fig. 6.3).

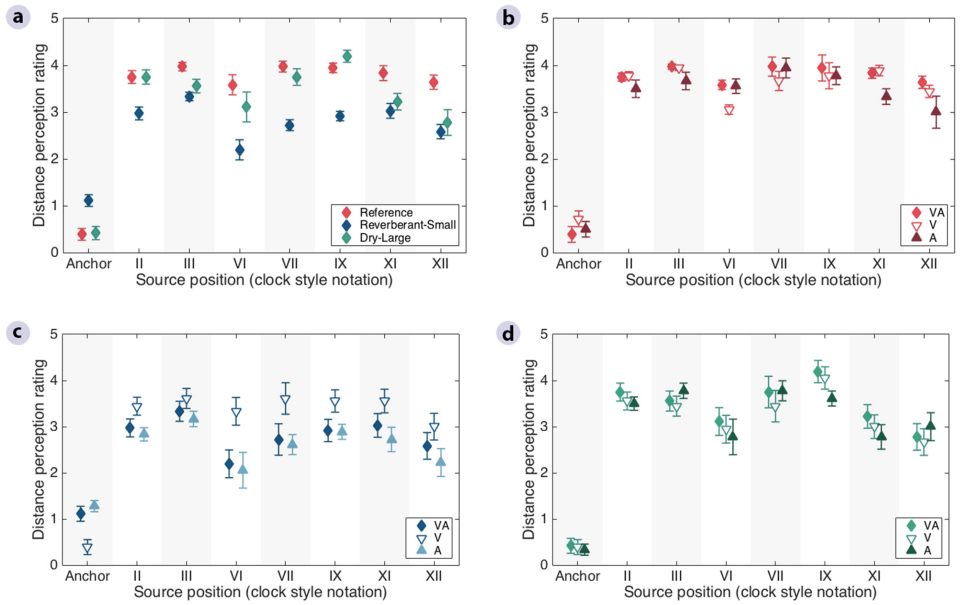


Figure 6.3: Average distance perception ratings in the reference room and the two rooms that were incongruent with the reference, as a function of source position. The ratings were obtained in three conditions, with both visual and auditory room cues (VA), and with either visual (V) or auditory room cues only (A). a) Condition with both visual and auditory room cues across the three rooms. b) Three tested conditions in the Reference room. c) Three tested conditions in the Reverberant-Small room. d) Three tested conditions in the Dry-Large room. The average of the two presentations per listener was used to calculate the across-listener mean, and the error bars show the standard error of the mean across listeners.

Unlike the distance ratings, the azimuthal direction (Fig. 6.2b, full bars) and compactness (Fig. 6.2c) judgements did not show a dependency on the listening environment. Instead, they varied with the position of the target stimuli (Figs. 6.4 and 6.5), with the overall percentage of correct judgements ranging from 40 to 60%. Linear mixed-effects-model ANOVAs with Room, Cue, and Position as fixed factors and Listener as a random factor (Tab. 6.A.6 and 6.A.9) confirmed that, for these two attributes, there was a significant effect of Posi-

tion [Direction: $F(6,700) = 40.6$, $p < 0.001$; Compactness: $F(6,1456) = 27.8$, $p < 0.001$] but no effect of Room [Direction: $F(2,700) = 0.4$, $p = 0.639$; Compactness: $F(2,1456) = 1.0$, $p = 0.357$]. Lateral positions (III and IX) were consistently rated more accurately than front (XII) and back (VI) positions for both direction and compactness (Figs. 6.4a and 6.5a). Post hoc multiple comparisons using Tukey's honest significant difference test showed that, when all room cues were available, 7 out of 8 front/back vs lateral comparisons were significant, while ratings for positions III vs IX and VI vs XII never differed significantly (Tab. 6.A.7 and 6.A.10).

For directional ratings (Fig. 6.4), the best performance was observed for speech signals delivered from the III and IX o'clock positions in all rooms, whereas the worst performance was observed for positions VI, VII, and XII o'clock. The height of the empty bars in Fig. 6.2b indicates what the percentage of correct directional judgements would be if front-back confusions were considered as correct responses, i.e., the difference between empty and full bars reflects the rate of front-back confusions for each condition. A linear mixed-effects model ANOVA performed on the rate of front-back confusions revealed no significant effect of Room [$F(2,700) = 0.6$, $p = 0.536$] but a significant effect of Position [$F(6,700) = 15.2$, $p < 0.001$] (Tab. 6.A.12), and subsequent post-hoc analysis showed significant differences only for pairwise comparisons involving either positions VI or XII o'clock (Tab. 6.A.13). This confirms that listeners tended to localize peripheral sounds more easily, whereas front and back positions were often confused. In addition, the comparison of variance of the direction judgements showed a significant difference across positions ($\chi^2(6) = 34.23$, $p < 0.001$), with the lowest variance observed for lateral positions III and IX o'clock (Tab. 6.A.15). However, no clear effect was found on whether or not the stimuli were colocated with visible loudspeakers (Tab. 6.A.16).

In terms of compactness ratings (Fig. 6.5a), the stimuli rated in all three environments showed a significantly larger reported source broadness for front and back than for lateral positions. The post-hoc analysis confirmed that, when both visual and auditory room cues were available, all front/back vs lateral comparisons were significantly different, while III vs IX and VI vs XII did not differ significantly (Tab. 6.A.7). Thus, virtual stimuli delivered from the front and back positions are not only problematic for sound localization, but also present a challenge for spatial segregation.

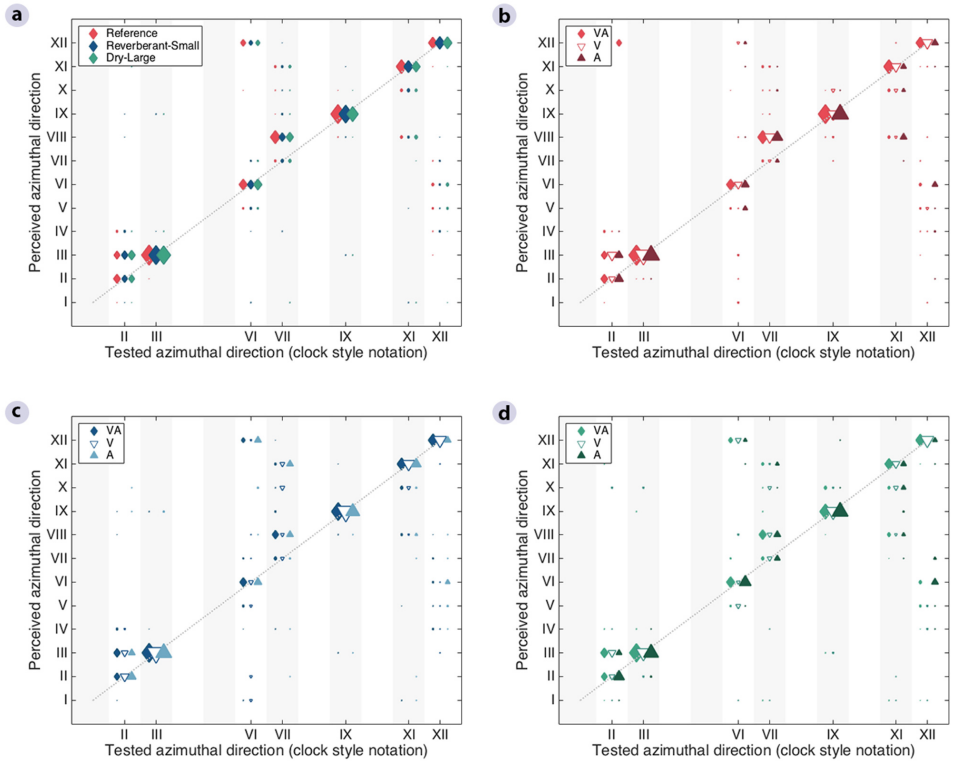


Figure 6.4: Azimuthal direction ratings in the reference room and the two rooms that were incongruent with the reference, as a function of source position. Marker size reflects the number of responses for each tested vs. perceived direction. The ratings were obtained in three conditions, with both visual and auditory room cues (VA), and with either visual (V) or auditory room cues only (A). a) Condition with both visual and auditory room cues across the three rooms. b) Three tested conditions in the Reference room. c) Three tested conditions in the Reverberant-Small room. d) Three tested conditions in the Dry-Large room.

6.2.2 Effect of auditory vs visual awareness of the room

The influence of the type of available room-related cues on sound externalization was studied by comparing the results for the condition where listeners received both visual and auditory room cues (Fig. 6.2, left panels) to the conditions where they received either visual (Fig. 6.2, middle panels) or auditory room cues (Fig. 6.2, right panels) only.

For azimuthal direction and compactness (Figs. 6.2b and 6.2c), no significant main effect was found between the tested conditions in the ANOVA [Direction: $F(2,700) = 0.4$, $p = 0.673$; Compactness: $F(2,1456) = 0.7$, $p = 0.478$] (Tab. 6.A.6

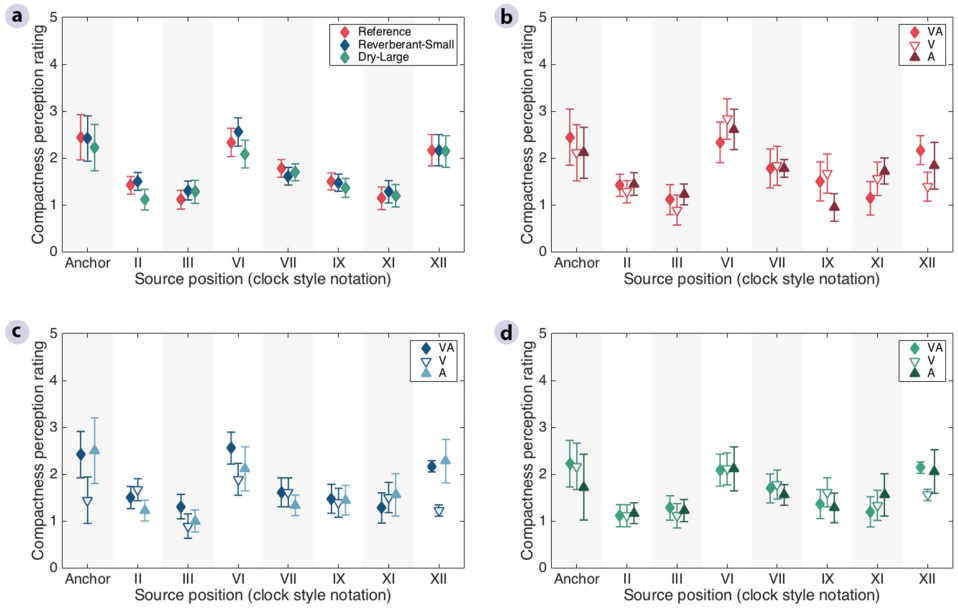


Figure 6.5: Average compactness perception ratings in the reference room and the two rooms that were incongruent with the reference, as a function of source position. The ratings were obtained in three conditions, with both visual and auditory room cues (VA), and with either visual (V) or auditory room cues only (A). a) Condition with both visual and auditory room cues across the three rooms. b) Three tested conditions in the Reference room. c) Three tested conditions in the Reverberant-Small room. d) Three tested conditions in the Dry-Large room. The average of the two presentations per listener was used to calculate the across-listener mean, and the error bars show the standard error of the mean across listeners.

and 6.A.9). There were significant interactions between Cue and Position for the two attributes [Direction: $F(12,700) = 2.0$, $p = 0.019$; Compactness: $F(12,1456) = 2.6$, $p = 0.002$]. However, the post-hoc analysis showed that none of the pairwise comparisons remained significant after correction for multiple comparisons for direction (Tab. 6.A.11), while only the comparison for all room cues vs visual room cues only in position XII remained significant for compactness (Tab. 6.A.8).

In contrast to the other two attributes, the pattern of distance judgements in the three rooms varied significantly between the three cue conditions (Fig. 6.2a). The ANOVA (Tab. 6.A.1) showed a significant effect of Cue [$F(2,1456) = 7.6$, $p = 0.001$], and a significant interaction between Room and Cue [$F(4,1456) = 16.6$, $p < 0.001$]. In the Reference room (Fig. 6.2a, red bars), the ratings obtained

for conditions with either visual or auditory cues (middle and right panels) resulted in slightly lower distance ratings than in the condition with both cues available (left panel), but these differences were not significant (Tab. 6.A.2). The distance scores obtained in the Dry-Large room (Fig. 6.2a, green bars) were essentially unaffected by the type of cue, reflected by insignificant differences in judgements across conditions (Tab. 6.A.2). In contrast, in the Reverberant-Small room (Fig. 6.2a, blue bars), listeners were significantly more accurate (by 35%) in the visual room cue only condition (middle panel) compared to the other two conditions (left and right panels), which were similar. The post-hoc analysis confirmed that the presence of incongruent auditory room cues led to significantly lower distance ratings in the Reverberant-Small room (Fig. 6.2a, blue bars) only ($p < 0.001$, Table S2). Moreover, the differences in distance ratings between the Reverberant-Small and the other two rooms were only significant when auditory room cues were available ($p < 0.001$, Table S3). Therefore, the listeners' distance judgements were reduced whenever they received auditory room cues from the playback room that did not match those from the recording room, while their judgements remained unaffected when they could see a room that differed from the one they heard through the headphone reproduction. This behaviour was consistent across all source positions (Fig. 6.3b, 6.3c, and 6.3d), and no interaction was found between Cue and Position in the ANOVA [$F(12,1456) = 0.6$, $p = 0.818$].

6.3 Summary and discussion

The above results indicate that a mismatch between the room in which the binaural headphone reproduction was played back and the room in which the BRIRs were recorded is detrimental to the externalization of the resulting auditory images in terms of their perceived distance. However, there was no evidence that such a room mismatch affects the perceived azimuthal direction or compactness of the auditory images.

The findings also suggest that the auditory modality has a higher impact on externalization in terms of perceived distance than the visual modality, when cues from the recording and playback room are incongruent. It should be noted that the perceptual judgements were obtained during an auditory-only task, which might explain why the observed effects only occurred when the infor-

mation was incongruent within the same (auditory) modality and not across modalities. Concerning the role of the visual modality, a clear distinction should be made between room-related visual cues, which did not affect externalization here, and source-related visual cues, which have been reported to influence the auditory spatial impression in experiments where audiovisual stimuli convey spatial discrepancies (McDonald et al., 2001; Alais and Burr, 2004). In the present study, a statistical comparison of the variance of the direction estimates showed that it varied significantly across positions (Tab. 6.A.15). Although this reflects the fact that more confusions occurred for some positions than others (Fig. 6.4), the significant pairwise comparisons did not systematically occur between positions with and without visible loudspeakers (Tab. 6.A.16).

Overall, our results demonstrate that the highest degree of externalization is obtained in the presence of both auditory and visual congruent information. In incongruent listening situations, the auditory information about the room becomes more critical for the perception of distance when the listening environment is more reverberant compared to the recording room, but not when the listening room is anechoic. Such a result might be explained by the fact that the difference in reverberation time between the Reference and the Reverberant-Small room (2.4 s) was much larger than that between the Reference and the Dry-Large room (0.4 s). Moreover, it may also be due to the anechoic nature of the Dry-Large room. In a reverberant room, the only natural scenario in which a listener could hear an acoustic signal with comparatively low reverberation is if the sound source is very close. This might explain the lower distance ratings in the condition with all cues available in the Reverberant-Small room. However, in the anechoic Dry-Large room, the noise signal carried practically no room information, which might be the reason why the auditory incongruence did not result in conflicting room information and thus did not affect distance ratings in the Dry-Large room. In that sense, an anechoic room is a very special environment, and the results might well be different in a “real” room with a short but non-zero reverberation time.

The outcomes of the present study are relevant in listening experiments that use binaural stimuli, especially when the acoustical features of the listening environments differ from the inherent acoustic properties of the target signals. Therefore, special care should be taken during the selection of tests rooms,

where matching acoustical features is more crucial than visual congruence. Considering this aspect may help reduce the bias of perceptual judgements in listening experiments using virtual headphone reproduction. In addition, our results suggest that the listening space of the user should be considered when designing virtual reality and multimedia reproduction systems.

6.4 Methods

6.4.1 Listeners and rooms

Eighteen naïve subjects participated in the experiment (20-29 years old). None had been in any of the test rooms before. The subjects reported normal or corrected-to-normal vision; three of them wore corrective lenses. All had normal hearing, which was verified with pure tone audiograms obtained for each subject before testing. All subjects provided informed consent prior to their participation in the experiments, which were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-3-2013-004) and carried out in accordance with the corresponding guidelines and relevant regulations on the use of human subjects for health-related scientific research. The experiments were conducted in three soundproof rooms, which were selected so that the acoustic features and dimensions contrast with each other. The reference room (Reference) had a reverberation time of 0.4 s and a volume of 99 m^3. The other two listening rooms (Reverberant Small and Dry Large) had reverberation times of 2.8 s and <0.01 s, and volumes of 43.2 m^3 and 330.4 m^3, respectively. The background noise level was below 19 dB(A) in all three rooms.

6.4.2 BRIR recordings

For the BRIR recordings, the listeners were instructed to keep as quiet as possible while they were in the Reference room. Blindfolds and earplugs reduced the available visual or auditory information about the room and the loudspeakers. The listeners were seated on a listening chair looking straight ahead. A headrest was provided to help them keep their heads still. Omnidirectional DPA 4060 lapel microphones were placed at the ear canal entrance, on top of the earplugs, and attached to the pinna with a wire hook. Seven azimuthal source positions were recorded at a distance of 1.5 m from the listener: 0°, 60°, 90°, 180°, 210°, 270°, and 330°, also referred to as positions XII, II, III, VI, VII, IX, and XI o'clock,

respectively (red circles in Fig. 6.1b). These were selected to provide front-back positions (XII and VI), within cone of confusion positions (II, VII, and XI), and lateral, outside cone of confusion positions (III and IX).

Six repetitions of a 5-s logarithmic sine sweep per position were reproduced through Dynaudio BM6P loudspeakers placed at eye level. The BRIRs were then obtained using a deconvolution method (Müller and Massarani, 2001). At the listener position, the sound pressure level measured with a B& K 2250 sound level meter was 65 dB(A). The recordings and playback were made through a portable M-audio Fast Track Ultra sound card at a sampling frequency of 48 kHz. Sennheiser HD 800 headphones were used in the listening experiment. To compensate for their effect in the transmission path, individual headphone impulse responses (HPIRs) were recorded. To do so, ten 2-s logarithmic sine sweeps were played back through the headphones positioned on the listeners while the microphones were still placed in the same position. As before, the resulting HPIRs were then transformed into the frequency domain using the fast Fourier Transform. A regularization parameter was used to remove the frequency content of the headphone responses below 50 Hz and above 18 kHz. The speech material, the BRIRs, and HPIRs were then convolved, and the resulting signals were stored to be used during the experiment. After this procedure was completed, the test subject was guided outside the room, where earplugs and blindfolds were removed.

6.4.3 Stimuli

The stimuli were male speech sentences with a duration of approximately 2 s each. In total, 24 different sentences were taken from the Danish version of the hearing in noise test (HINT; Nielsen and Dau, 2011). Eight different signals were used in each room, seven were convolved with the BRIRs, and one unprocessed signal was presented diotically through the headphones (Anchor). In the reference room one additional signal was reproduced from the loudspeaker positioned at III o'clock. This was done to inspect whether results were different between real (loudspeaker) and virtual (headphones) stimuli, and thus verify the accuracy of the binaural reproduction. The results obtained were used to define the criteria for correct ratings that were regarded as those with a score of 4 for distance, localized at the correct azimuthal direction, and within the range 0 to 1 for compactness. To ensure that the headphones did not unduly

attenuate the sounds from the room, real-ear insertion gains were measured with probe microphones for one listener. The result for a frontal loudspeaker position showed that the headphones caused some attenuation between about 1.5 and 5.3 kHz. The average attenuation across this range was -5.7 dB, the maximum attenuation occurred as a dip of 8.9 dB at 2.66 kHz. While this causes some minor coloration of the acoustic scene, it is still clearly possible to assess the characteristics of the room through the headphones.

6.4.4 Experimental procedure

The order of the rooms in the listening experiment was randomly determined for each listener. Each participant was assigned a group that defined the starting condition, either visual room cues only or auditory room cues only. Listeners from the first group entered the first room seeing the environment, but listening to loud music over headphones (approx. 85 dB SPL). The subjects were also instructed to avoid speaking and to keep as quiet as possible while they were in the room. Once in the listening position, the music was stopped and the listeners started the experiment. Participants from the second group entered the room wearing blindfolds but no earplugs. Once seated, the lights were turned off and they were allowed to uncover the eyes. The light provided by the user interface (iPad) was sufficient to see the loudspeaker positions but no further into the room. In addition, small dimmed lights were placed on top of each loudspeaker to ensure that listeners always had a clear visual reference for the position. To provide auditory room cues, white noise bursts of 500-ms duration were reproduced. The noise signal was played back through a Bose Soundlink Mobile speaker located behind the listener (at V o'clock). The distance from the test subject was 2 m and the sound pressure level at the listening position was 35 dB(A), which was well below the stimulus level and therefore assumed not to distract from the experimental task. Given the unfamiliar nature of the noise and its low level, listeners were able to segregate this signal from the target speech stimuli delivered over headphones. All listeners were also instructed to keep their head still and look at the front loudspeaker during stimulus presentation, but the head was not fixated, because the externalization percept seemed fairly robust with respect to small head movements. Once the starting condition was completed, the participants took a short break outside the room. Then, in the same room, the listeners performed the experiment in condition with both visual and auditory room cues (i.e., without any visual or auditory

restrictions). In this condition, the external noise source was also activated. The procedure was repeated in the other two rooms in random order. The listeners wore the headphones during the whole experiment and while entering the test rooms for the group with visual room cues only. They did not wear headphones while being guided from one room to the next.

The listening ratings were done through an iPad user interface implemented in MATLAB R2015a, where the subjects could push buttons to rate the different externalization parameters. A training session was conducted in the very first trial of the initial condition to familiarize the listeners with the task. The training comprised one complete experimental run for all three attributes and lasted about 10-15 min. No feedback was given on the ratings, but a short discussion was held to make sure that the attributes were understood correctly. In each condition the experiment was divided into two blocks. In the first block, listeners were first asked to judge the perceived direction of the stimuli by selecting that one of the twelve possible numbers on a clock style notation, which best represented the direction of the incoming sound (Fig. 6.1b). Once a direction was chosen and confirmed, the same stimulus was presented again, but this time a compactness rating was required. For this attribute, the test subjects were provided an interface containing concentric coloured areas with increasing broadness. Six options were available, where area 0 was the most compact perception, corresponding to the area occupied by the loudspeaker. Area 5, on the other hand, represented a compactness perception that exceeded 120°. Once the compactness rating was selected and confirmed, a new signal was delivered randomly at another position. The rest of the experimental block was completed by interleaving the ratings for the two parameters. In the second block, distance judgements were obtained for the same stimuli. Subjects were presented with a diagram containing six concentric zones with increasing diameter, where zone 0 represented perception inside the head, zone 4 a sound perceived as coming from the loudspeaker position, and zone 5 a stimulus perceived at a distance beyond the loudspeaker position. Zones 1, 2, and 3 represented the source being perceived at the ears, at a location closer to the listener than the loudspeaker, and at a location closer to the loudspeaker than the listener, respectively. Two ratings were obtained at each position tested for all the three parameters. A replay button enabled the subjects to repeat the stimuli as often as required.

Four loudspeakers placed at positions I, III, XI, and XII o'clock (loudspeaker pictograms in Fig. 6.1b) were visible during the test. These were labelled accordingly. The loudspeaker setup provided a visual reference during the experiment and served to study the potential influence of visual targets on auditory perception, especially for virtual stimulus positions adjacent to loudspeakers.

6.4.5 Statistical Analysis

A significance level of 0.05 was used for all analyses. The statistical analyses for distance and compactness data were performed on the raw listener ratings (summarized in Figs. 6.3 and 6.5), before transformation to the percent correct ratings shown in Fig. 6.2. Although the underlying subjective attributes for distance and compactness could be assumed to vary on continuous scales, the collected data was ordinal in nature due to the use of discrete rating scales, for which the assumption of equally-distant scale points was not necessarily valid. In order to test whether a parametric linear mixed-effects model could be robust enough to the non-continuous nature of the data and whether this would increase the risk of Type I errors, one thousand data sets were simulated by keeping Listener, Room, Cue, and Position the same as in the original data set but taking each listener's distance and compactness values to be a random sample from their original distance and compactness values with replacement. The resulting alpha-values were very close to 0.05, indicating that the Type I error rate was not unduly increased and that a linear mixed-effects model could be assumed to be robust enough to the non-continuous dependent variables.

A linear mixed-effects model with Room, Cue, and Position as fixed factors, and Listener as a random factor was fitted to the data. Visual inspection of the residuals showed no major deviation from normality or homoscedasticity. The statistical analysis for distance and compactness ratings was thus based on an ANOVA (Tabs. 6.A.1 and 6.A.6). A reduced model, from which insignificant three-way interactions were removed, was used. A post-hoc analysis using Tukey's honest significant difference test was carried out to study multiple pairwise comparisons when both main and interaction effects were significant (Tab. 6.A.2 to 6.A.5, 6.A.7 and 6.A.8).

For azimuthal direction, mixed-effects ANOVA models similar to that used for the analysis of distance and compactness judgements was used, except that

the analyses were performed on the rate of correct judgements (Tab. 6.A.9) and the rate of front-back confusions (Tab. 6.A.12) presented in Fig. 6.2b. Front-back confusions were calculated by dividing the number of judgements in a hemisphere that differed from that of the stimulus position over the number of presentations. Post-hoc pairwise comparisons were again studied using Tukey's honest significant difference test (Tab. 6.A.10, 6.A.11, 6.A.13, and 6.A.14).

The variance of directional ratings pooled across conditions for each position and subject was compared across positions using a Friedman test (Tab. 6.A.15) and pairwise comparisons were studied with Bonferroni-corrected Wilcoxon sign-rank tests (Tab. 6.A.16).

Acknowledgements

We thank Elisabeth Wreford Andersen (Technical University of Denmark), Frank Eriksson (Copenhagen University), and Mark Seeto (National Acoustic Laboratories, Sydney, Australia) for their help with the statistical analysis of the data, as well as Ewen N. MacDonald for proofreading the manuscript. Research funded by the Technical University of Denmark and a research consortium with Oticon, Widex, and GN ReSound.

6.A Supplementary information

6.A.1 Supplementary data: Comparison of loudspeaker and headphone presentation

6.A.2 Statistical analysis: Detailed results

The results of the statistical analyses are detailed in the following tables. Note that for distance and compactness perception all statistical tests were performed on the raw ratings data and not the percent correct outcomes presented in Fig. 2. For perceived azimuthal direction, the tests were performed on both the rate of correct judgements and the rate of front-back confusions. In the following tables, VA stands for "Visual and Auditory room cues", V for "Visual room cues only", and A for "Auditory room cues only". ***: $p < .0010$; **: $p < .0100$; *: $p < .0500$.

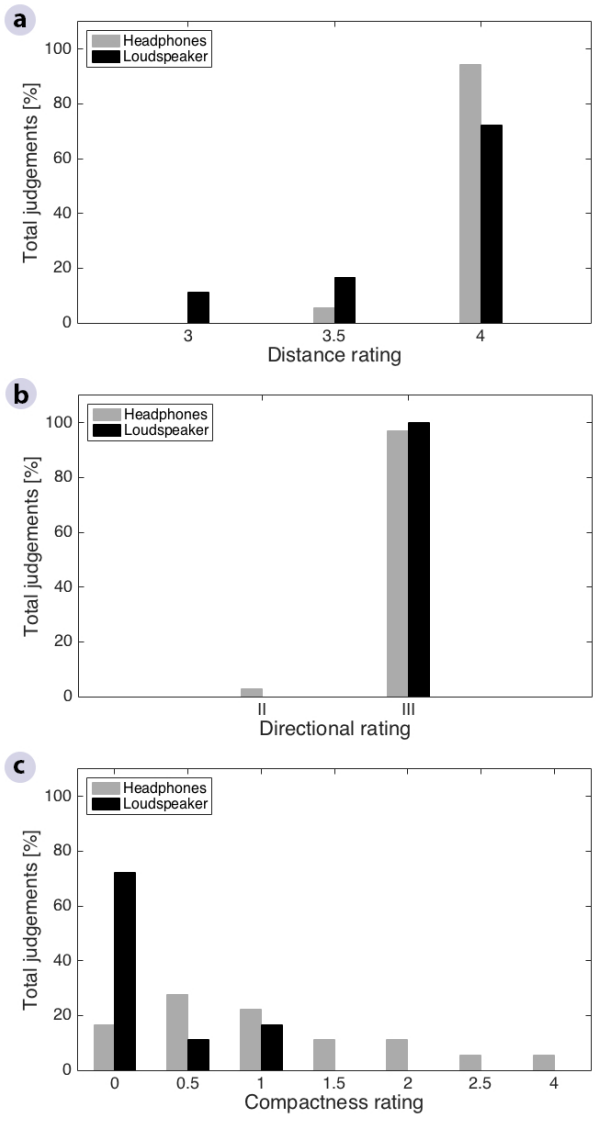


Figure 6.A.1: Distributions of listeners' judgements for each externalization parameter for anechoic stimuli presented through the loudspeaker (black bars) vs the corresponding headphone signal (grey bars). The ratings were obtained in the Reference room for stimuli delivered from position III with both visual and auditory room cues available. a) Distribution of distance judgements. Means and standard deviations: Headphones (M = 3.97, SD = 0.17); Loudspeaker (M = 3.81, SD = 0.40). b) Distribution of directional judgements. Means and standard deviations: Headphones (M = 2.97, SD = 0.17); Loudspeaker (M = 3.00, SD = 0.00). c) Distribution of compactness judgements. Means and standard deviations: Headphones (M = 1.11, SD = 1.11); Loudspeaker (M = 0.22, SD = 0.42). Distance and compactness judgements were averaged per subject over the two trials. Only ratings that had an occurrence above 0% are presented.

Table 6.A.1: Statistical analysis for perceived distance ratings. Results of a linear mixed-effects model ANOVA with Room, Cue, and Position as fixed factors and Listener as random factor. A full model showed a non-significant three-way interaction, such that a reduced model including two-way interactions only was used.

	numDF	denDF	F-value	p-value
Room	2	1456	94.6338	<.0001***
Cue	2	1456	7.6176	0.0005***
Position	6	1456	28.7569	<.0001***
Room:Cue	4	1456	16.6246	<.0001***
Room:Position	12	1456	4.8822	<.0001***
Cue:Position	12	1456	0.6296	0.8184

Table 6.A.2: Post-hoc analysis for perceived distance ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over positions for the Cue factor as a function of Room, using Tukey's honest significant difference test.

	Reference room	Reverberant-Small	Dry-Large
VA vs V	1.975 (0.5611)	-6.989 (<.0001***)	2.156 (0.4353)
VA vs A	3.005 (0.0670)	0.822 (0.9962)	1.737 (0.7233)
V vs A	0.822 (0.9962)	7.109 (<.0001***)	-0.334 (1.0000)

Table 6.A.3: Post-hoc analysis for perceived distance ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over positions for the Room factor as a function of Cue, using Tukey's honest significant difference test. Ref: Reference room; Rev: Reverberant-Small room; Dry: Dry-Large room.

	VA	V	A
Ref vs Rev	14.521 (<.0001***)	2.136 (0.4492)	9.282 (<.0001***)
Ref vs Dry	4.879 (<.0001***)	3.614 (0.0094**)	2.300 (0.3427)
Rev vs Dry	-9.642 (<.0001***)	1.479 (0.8656)	-6.982 (<.0001***)

Table 6.A.4: Post-hoc analysis for perceived distance ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over cue conditions for the Position factor as a function of Room, using Tukey's honest significant difference test.

	Reference Room	**Reverberant-Small**	**Dry-Large**
II vs III	-1.541 (0.9949)	-2.395 (0.6895)	0.379 (1.0000)
II vs VI	1.861 (0.9570)	4.636 (0.0007***)	4.956 (0.0002***)
II vs VII	-1.601 (0.9918)	0.960 (1.0000)	-0.427 (1.0000)
II vs IX	-1.257 (0.9997)	-0.083 (1.0000)	-2.857 (0.3377)
II vs XI	-0.095 (1.0000)	-0.095 (1.0000)	4.600 (0.0009***)
II vs XII	2.099 (0.8745)	3.592 (0.0474*)	6.473 (<.0001***)
III vs VI	3.403 (0.0858)	7.031 (<.0001***)	4.576 (0.0010**)
III vs VII	-0.059 (1.0000)	3.355 (0.0987)	-0.806 (1.0000)
III vs IX	0.285 (1.0000)	2.312 (0.7490)	-3.237 (0.1377)
III vs XI	1.446 (0.9977)	2.300 (0.7571)	4.221 (0.0046**)
III vs XII	3.640 (0.0405*)	5.987 (<.0001***)	6.094 (<.0001***)
VI vs VII	-3.462 (0.0717)	-3.675 (0.0359*)	-5.383 (<.0001***)
VI vs IX	-3.118 (0.1875)	-4.719 (0.0005***)	-7.813 (<.0001***)
VI vs XI	-1.956 (0.9311)	-4.731 (0.0005***)	-0.356 (1.0000)
VI vs XII	0.237 (1.0000)	-1.043 (1.0000)	1.518 (0.9958)
VII vs IX	-0.884 (1.000)	-1.043 (1.0000)	-2.430 (0.6627)
VII vs XI	1.506 (0.9962)	-1.055 (1.0000)	5.027 (0.0001***)
VII vs XII	3.699 (0.0331*)	2.632 (0.5046)	6.900 (0.0001***)
IX vs XI	1.162 (0.9999)	-0.012 (1.0000)	7.457 (0.0001***)
IX vs XII	3.355 (0.0987)	3.675 (0.0359*)	9.331 (0.0001***)
XI vs XII	2.193 (0.8243)	3.687 (0.0345*)	1.873 (0.9543)

Table 6.A.5: Post-hoc analysis for perceived distance ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over cue conditions for the Room factor as a function of Position, using Tukey's honest significant difference test. Ref: Reference room; Rev: Reverberant-Small room; Dry: Dry-Large room.

	II	**III**	**VI**
Ref vs. Rev	4.389 (0.0022**)	3.527 (0.0585)	7.193 (<.0001***)
Ref vs. Dry	0.370 (1.0000)	2.310 (0.7502)	3.496 (0.0645)
Rev vs. Dry	-4.020 (0.0102*)	-1.217 (0.9998)	-3.696 (0.0334*)

	VII	**IX**	**XI**
Ref vs. Rev	6.977 (<.0001***)	5.575 (<.0001***)	4.389 (0.0022**)
Ref vs. Dry	1.556 (0.9942)	-1.248 (0.9997)	5.113 (0.0001***)
Rev vs. Dry	-5.421 (<.0001***)	-6.823 (<.0001***)	0.724 (1.0000)

	XII
Ref vs. Rev	5.899 (<.0001***)
Ref vs. Dry	4.790 (0.0004***)
Rev vs. Dry	-1.109 (1.0000)

Table 6.A.6: Statistical analysis for perceived compactness ratings. Results of a linear mixed-effects model ANOVA with Room, Cue, and Position as fixed factors and Listener as random factor. A full model showed a non-significant three-way interaction, such that a reduced model including two-way interactions only was used.

	numDF	**denDF**	**F-value**	**p-value**
Room	2	1456	1.03099	0.3569
Cue	2	1456	0.73583	0.4780
Position	6	1456	27.83419	<.0001***
Room:Cue	4	1456	0.76010	0.5513
Room:Position	12	1456	0.98721	0.4587
Cue:Position	12	1456	2.58324	0.0021**

Table 6.A.7: Post-hoc analysis for perceived compactness ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Cue factor as a function of Position, using Tukey's honest significant difference test.

	VA	**V**	**A**
II vs III	0.783 (1.0000)	1.938 (0.9368)	0.646 (1.0000)
II vs VI	-6.917 (<.0001***)	-4.614 (0.0008**)	-4.983 (0.0001***)
II vs VII	-2.480 (0.6247)	-1.938 (0.9368)	-1.384 (0.9988)
II vs IX	-0.718 (1.0000)	-1.015 (1.0000)	0.277 (1.0000)
II vs XI	0.979 (1.0000)	-0.554 (1.0000)	-1.661 (0.9873)
II vs XII	-5.742 (<.0001***)	-0.185 (1.0000)	-3.876 (0.0176)
III vs VI	-7.700 (<.0001***)	-6.552 (<.0001***)	-5.629 (<.0001***)
III vs VII	-3.263 (0.1282)	-3.876 (0.0176)	-2.030 (0.9044)
III vs IX	-1.501 (0.9963)	-2.953 (0.2761)	-0.369 (1.0000)
III vs XI	0.196 (1.0000)	-2.492 (0.6154)	-2.307 (0.7523)
III vs XII	-6.526 (<.0001***)	-2.123 (0.8627)	-4.522 (0.0012**)
VI vs VII	4.437 (0.0018**)	2.676 (0.4701)	3.599 (0.0463)
VI vs IX	6.199 (<.0001***)	3.599 (0.0463*)	5.260 (<.0001***)
VI vs XI	7.896 (<.0001***)	4.061 (0.0087**)	3.322 (0.1085)
VI vs XII	1.175 (0.9999)	4.430 (0.0019**)	1.107 (1.0000)
VII vs IX	1.762 (0.9756)	0.923 (1.0000)	1.661 (0.9873)
VII vs XI	3.459 (0.0725)	1.384 (0.9988)	-0.277 (1.0000)
VII vs XII	-3.263 (0.1282)	1.753 (0.9768)	-2.492 (0.6154)
IX vs XI	1.697 (0.9839)	0.461 (1.0000)	-1.938 (0.9368)
IX vs XII	-5.025 (0.0001**)	0.831 (1.0000)	-4.153 (0.0060**)
XI vs XII	-6.721 (<.0001***)	0.369 (1.0000)	-2.215 (0.8117)

Table 6.A.8: Post-hoc analysis for perceived compactness ratings. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Position factor as a function of Cue, using Tukey's honest significant difference test.

	II	**III**	**VI**
VA vs V	-0.178 (1.0000)	1.387 (0.9987)	0.135 (1.0000)
VA vs A	0.492 (1.0000)	0.596 (1.0000)	0.387 (1.0000)
V vs A	0.558 (1.0000)	-0.659 (1.0000)	0.210 (1.0000)
	VII	**IX**	**XI**
VA vs V	-0.387 (1.0000)	-0.752 (1.0000)	-1.587 (0.9926)
VA vs A	0.909 (1.0000)	1.379 (0.9988)	-2.170 (0.8377)
V vs A	1.079 (1.0000)	1.775 (0.9736)	-0.485 (1.0000)
	XII		
VA vs V	4.205 (0.0049**)		
VA vs A	0.700 (1.0000)		
V vs A	-2.919 (0.2975)		

Table 6.A.9: Statistical analysis for the rate of correct directional judgements. Results of a linear mixed-effects model ANOVA with Room, Cue, and Position as fixed factors and Listener as random factor. A full model showed a non-significant three-way interaction, such that a reduced model including two-way interactions only was used.

	numDF	**denDF**	**F-value**	**p-value**
Room	2	700	0.44859	0.6387
Cue	2	700	0.39694	0.6725
Position	6	700	40.61668	<.0001***
Room:Cue	4	700	1.09885	0.3560
Room:Position	12	700	0.61351	0.8318
Cue:Position	12	700	2.03975	0.0189*

Table 6.A.10: Post-hoc analysis for the rate of correct directional judgements. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Cue factor as a function of Position, using Tukey's honest significant difference test.

	VA	**V**	**A**
II vs III	-7.359 (<.0001***)	-4.930(0.0002***)	-3.469 (0.0716)
II vs VI	-1.420 (0.9982)	0.913 (1.0000)	0.183 (1.0000)
II vs VII	3.098 (0.1986)	1.826 (0.9642)	3.104 (0.1960)
II vs IX	-6.068 (<.0001***)	-3.834 (0.0213*)	-3.286 (0.1218)
II vs XI	-2.711 (0.4443)	-1.461 (0.9974)	1.278 (0.9996)
II vs XII	-2.582 (0.5445)	-2.739 (0.4236)	1.826 (0.9642)
III vs VI	5.939 (<.0001***)	5.842 (<.0001***)	3.652 (0.0399*)
III vs VII	10.457 (<.0001***)	6.755 (<.0001***)	6.573 (<.0001***)
III vs IX	1.291 (0.9995)	1.095 (1.0000)	0.183 (1.0000)
III vs XI	4.648 (0.0008***)	3.469 (0.0716)	4.747 (0.0005***)
III vs XII	4.777 (0.0004***)	2.191 (0.8251)	5.295 (<.0001***)
VI vs VII	4.519 (0.0014**)	0.913 (1.0000)	2.921 (0.2974)
VI vs IX	-4.648 (0.0008***)	-4.747 (0.0005***)	-3.469 (0.0716)
VI vs XI	-1.291 (0.9995)	-2.374 (0.7049)	1.095 (1.0000)
VI vs XII	-1.162 (0.9999)	-3.652 (0.0399)	1.643 (0.9886)
VII vs IX	-9.166 (<.0001***)	-5.660 (<.0001***)	-6.390 (<.0001***)
VII vs XI	-5.810 (<.0001***)	-3.286 (0.1218)	-1.826 (0.9642)
VII vs XII	-5.680 (<.0001***)	-4.564 (0.0011**)	-1.278 (0.9996)
IX vs XI	3.357 (0.0999)	2.374 (0.7049)	4.564 (0.0011**)
IX vs XII	3.486 (0.0680)	1.095 (1.0000)	5.112 (0.0001***)
XI vs XII	0.129 (1.0000)	-1.278 (0.9996)	0.548 (1.000)

Table 6.A.11: : Post-hoc analysis for the rate of correct directional judgements. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Position factor as a function of Cue, using Tukey's honest significant difference test.

	II	**III**	**VI**
VA vs. V	-0.072 (1.0000)	0.239 (1.0000)	2.105 (0.8706)
VA vs. A	-1.587 (0.9925)	0.383 (1.0000)	-0.239 (1.0000)
V vs. A	1.271(0.9996)	-0.120 (1.0000)	1.966 (0.9272)
	VII	**IX**	**XI**
VA vs. V	-0.486 (1.0000)	0.447 (1.0000)	0.447 (1.0000)
VA vs. A	-0.550 (1.0000)	-0.447 (1.0000)	2.041 (0.8993)
V vs. A	0.054 (1.0000)	0.749 (1.0000)	-1.337 (0.9992)
	XII		
VA vs. V	-1.108 (1.0000)		
VA vs. A	2.559 (0.5623)		
V vs. A	-3.076 (0.2096)		

Table 6.A.12: Statistical analysis for the rate of front-back confusions. Results of a linear mixed-effects model ANOVA with Room, Cue, and Position as fixed factors and Listener as random factor. A full model showed a non-significant three-way interaction, such that a reduced model including two-way interactions only was used.

	numDF	**denDF**	**F-value**	**p-value**
Room	2	700	0.62513	0.5355
Cue	2	700	1.47168	0.2302
Position	6	700	15.20517	<.0001***
Room:Cue	4	700	0.67401	0.6101
Room:Position	12	700	0.58808	0.8528
Cue:Position	12	700	2.29771	0.0071**

Table 6.A.13: Post-hoc analysis for the rate of front-back confusions. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Cue factor as a function of Position, using Tukey's honest significant difference test.

	VA	**V**	**A**
II vs. III	2.327 (0.7383)	1.371 (0.9989)	0.274 (1.0000)
II vs. VI	-3.684 (0.0358*)	-0.548 (1.0000)	-3.565 (0.0531)
II vs. VII	-0.776 (1.0000)	-1.919 (0.9416)	-3.016 (0.2412)
II vs. IX	2.327 (0.7383)	1.371 (0.9989)	0.274 (1.0000)
II vs. XI	2.133 (0.8568)	1.371 (0.9989)	-0.548 (1.0000)
II vs. XII	-0.969 (1.0000)	0.823 (1.0000)	-4.661 (0.0007**)
III vs. VI	-6.011 (<.0001***)	-1.919 (0.9416)	-3.839 (0.0209*)
III vs. VII	-3.102 (0.1968)	-3.290 (0.1204)	-3.290 (0.1204)
III vs. IX	0.000 (1.0000)	0.000 (1.0000)	0.000 (1.0000)
III vs. XI	-0.194 (1.0000)	0.000 (1.0000)	-0.823 (1.0000)
III vs. XII	-3.296 (0.1185)	-0.548 (1.0000)	-4.936 (0.0002**)
VI vs. VII	2.908 (0.3055)	-1.371 (0.9989)	0.548 (1.0000)
VI vs. IX	6.011 (<.0001***)	1.919 (0.9416)	3.839 (0.0209*)
VI vs. XI	5.817 (<.0001***)	1.919 (0.9416)	3.016 (0.2412)
VI vs. XII	2.714 (0.4417)	1.371 (0.9989)	-1.097 (1.0000)
VII vs. IX	3.102 (0.1968)	3.290 (0.1204)	3.290 (0.1204)
VII vs. XI	2.908 (0.3055)	3.290 (0.1204)	2.468 (0.6339)
VII vs. XII	-0.194 (1.0000)	2.742 (0.4211)	-1.645 (0.9885)
IX vs. XI	-0.194 (1.0000)	0.000 (1.0000)	-0.823 (1.0000)
IX vs. XII	-3.296 (0.1185)	-0.548 (1.0000)	-4.936 (0.0002***)
XI vs. XII	-3.102 (0.1968)	-0.548 (1.0000)	-4.113 (0.0074**)

Table 6.A.14: Post-hoc analysis for the rate of correct localisation judgements. t-ratios and p-values (in brackets) of pairwise contrasts averaged over rooms for the Cue factor as a function of Position, using Tukey's honest significant difference test.

	II	**III**	**VI**
VA vs V	0.246 (1.0000)	-0.068 (1.0000)	2.602 (0.5286)
VA vs A	1.638 (0.9890)	0.068 (1.0000)	0.539 (1.0000)
V vs A	-1.187 (0.9999)	-0.116 (1.0000)	1.759 (0.9756)
	VII	**IX**	**XI**
VA vs V	-1.324 (0.9993)	-0.068 (1.0000)	0.089 (1.0000)
VA vs A	-1.189 (0.9999)	0.068 (1.0000)	-0.717 (1.0000)
V vs A	-0.116 (1.0000)	-0.116 (1.0000)	0.688 (1.0000)
	XII		
VA vs V	1.974 (0.9246)		
VA vs A	-2.916 (0.3005)		
V vs A	4.170 (0.0059**)		

Table 6.A.15: Statistical analysis of the variance of directional judgements as a function of position. Results of a Friedman test carried out over the variances calculated per subject per position.

	Mean rank
II	3.50
III	2.08
VI	5.39
VII	4.89
IX	2.97
XI	4.28
XII	4.89
N	18
Chi-square (χ^2)	34.230
df	6
Asymp. sig.	< 0.0001

Table 6.A.16: Post-hoc analysis on the variance of directional judgements as a function of position. Pairwise comparisons using Wilcoxon signed rank tests with Bonferroni corrections (significance indicated by * for $p < 0.0023$).

	p-value
II vs. III	0.002*
II vs. VI	0.007
II vs. VII	0.007
II vs. IX	0.177
II vs. XI	0.033
II vs. XII	0.011
III vs. VI	0.003
III vs. VII	0.002*
III vs. IX	0.233
III vs. XI	0.001*
III vs. XII	0.004
VI vs. VII	0.016
VI vs. IX	0.004
VI vs. XI	0.031
VI vs. XII	0.407
VII vs. IX	0.001*
VII vs. XI	0.744
VII vs. XII	0.112
IX vs. XI	0.009
IX vs. XII	0.006
XI vs. XII	0.109

7

Overall summary and perspectives

7.1 Summary of main results

This thesis investigated two main topics: (i) speech intelligibility with and without hearing aid processing in noisy environments with different spatial configurations of the target and the maskers (chapters 2 and 3), and (ii) aspects of spatial hearing, particularly distance and externalization perception (chapters chapters 4, 5, 6). Furthermore, Appendix A investigated speech intelligibility in mobile phones and the applicability of speech intelligibility prediction models in the corresponding stimulus conditions.

In chapter 2, a loudspeaker-based virtual sound environment (VSE) system consisting of 29 loudspeakers was evaluated both in terms of physical and perceptual measures. The comparison of room impulse responses measured in the VSE and the corresponding real room showed that the room acoustic parameters reverberation time and clarity were well preserved. The interaural cross-correlation coefficient, however, was lower in the VSE than in the real room, indicating a more diffuse sound field. For the perceptual evaluation, speech reception thresholds (SRTs) were measured in eight normal-hearing listeners both in the classroom and in the VSE inside the loudspeaker array. The listeners were tested with and without hearing aids, and with omnidirectional and directional microphone processing. The SRTs measured in the VSE were, on average, slightly higher than in the classroom. However, the differences between conditions that were found in the classroom were preserved quite well in the VSE. The speech intelligibility benefit from directional microphone processing in relation to an omnidirectional setting observed in the VSE was similar to the one obtained in the classroom, albeit slightly smaller. Finally, the directivity of the hearing aids was measured both in the real and the simulated room as well as in an anechoic chamber. In the VSE, the directivity pattern was less pronounced than in the classroom, again indicating a more diffuse sound

field in the VSE. Overall, these findings showed that, despite some inevitable deviations of the sound field in the VSE from that in the classroom, the outcomes of listening experiments in a loudspeaker-based VSE system can be very similar to those obtained in a real-world listening environment.

An interesting (and unexpected) finding of chapter 2 was that the SRTs of normal-hearing listeners were higher (i.e., speech intelligibility was worse) with omnidirectional hearing-aid processing than without hearing aids. The study presented in chapter 3 tested whether this outcome was due to a degraded spatial perception of the scene when wearing hearing aids, e.g., due to the loss of pinna cues caused by the microphone position above the ear. In the study, "ideal" hearing aids were used with linear amplification and a "flat", i.e., frequency independent gain, leaving only the microphone position above the ear as a degrading factor. Furthermore, the study addressed the question of whether hearing aids have an effect on spatial release from masking (SRM) and informational masking (IM), as well as listening effort. Speech intelligibility was measured with normal-hearing listeners with and without hearing aids in a setting with target speech coming from the frontal direction. The target was presented together with three interferers which were either collocated with the target speech or spatially distributed around the listener, and were either other speakers or stationary speech-shaped noises with the same long-term frequency content. The spatial perception of the listeners was tested by asking them to sketch the spatial position of the perceived auditory images evoked by the sound sources, as well as their width.

With separated interferers, SRTs were found to be generally higher with noise than with speech interferers. The collocated thresholds were higher than the thresholds for the separated conditions. Consistent with the findings from chapter 2, "aided thresholds" were higher than "unaided thresholds". A larger SRM was found for the speech interferers than for the noise interferers and the amount of SRM was larger for unaided listening than for listening with hearing aids. The difference in SRM between the unaided and the aided condition was the same for the speech and the noise interferers, indicating that the lower SRM with hearing aids is most likely an effect of energetic masking, not IM. With respect to listening effort, no significant difference was found between the aided and the unaided conditions, but the listening effort was rated signifi-

cantly higher in the conditions with speech interferers than with SSN and in the conditions with collocated than with separated interferers. The sketches clearly showed that using hearing aids caused a distortion of the spatial perception. However, no clear indication of a detrimental effect on speech intelligibility could be shown.

Chapter 4 presented two experiments on auditory distance perception. Previous studies on externalization suggested that externalization is reduced when binaural stimuli, presented over headphones, are lowpass-filtered (Boyd et al., 2012; Catic et al., 2013). In the first experiment, it was found that this was not the case for distance perception, since no effect of the stimulus bandwidth could be observed. Regarding distance perception, it was found that stimuli that were simulated at close distances were usually perceived to be closer than the veridical distance, whereas stimuli at medium distances were, on average, perceived to be farther away. At the farthest distance, the auditory event was again rated to be closer to the listener than the actual distance. These results were in contrast to those reported in Zahorik (2002b) where listeners usually overestimated the distance at close distances and progressively underestimated the distance at medium and far distances. However, while the stimuli used in chapter 4 and in Zahorik (2002b) were very similar, Zahorik's experiment was conducted in a listening booth, whereas the study described in chapter 4 was conducted in the same room where the BRIRs had been measured. Therefore, it was tested whether the playback room in which the experiment was conducted had an influence on the results. A follow-up experiment was conducted with a subset of the same listeners in a listening booth to evaluate the influence of the different playback rooms. The general shape of the distance functions was similar to that observed in experiment 1. However, the data showed a larger within-listener and across-listener variability, and the distance ratings obtained in the two rooms were significantly different.

Chapter 5 presented a study that compared five different commercially available binaural microphones and the built-in microphones of a head-and-torso simulator in terms of the externalization that can be achieved. Individual BRIRs were measured for eight normal-hearing listeners and four different loudspeaker positions. It was found that using either speech or trains of noise bursts did not influence the externalization scores and that lateral sources were

better externalized than frontal sources. The degree of externalization achieved with individual binaural room impulse responses was slightly higher with all but one of the microphones compared to the generic binaural room impulse responses measured on a dummy head. However, the highest average externalization rating was observed for the internal microphones of the dummy head. These findings were unexpected since earlier research regarding localization demonstrated that listener performance is usually worse with generic than with individual HRTFs (e.g., Wightman and Kistler, 1989b; Møller et al., 1999; Minnaar et al., 2001). The results suggested that the exact geometry of the head and the pinnae and the resulting influence on the spectrum of the ear signals are less crucial for externalization than for localization and that externalization might be dominated by the statistical properties of the sound field at the two ears (e.g., the ILD distributions).

Inspired by the results from chapter 4, chapter 6 investigated the influence of the playback room on the perception of binaural stimuli through headphones in greater detail. Individual BRIRs were measured for normal-hearing, blindfolded listeners in a reference room. The listeners were asked to indicate the direction, distance, and compactness of binaural stimuli in three different experiment rooms while they were provided either visual or auditory awareness of their environment. The results showed that only the perceived distance of the stimuli was affected by the room, and that a reduced perceived distance was only observed when listeners could hear the acoustics of the playback room, and when there was an acoustic mismatch between the headphone stimuli and the playback room. These findings support the conclusion that not the visual impression of the room but the auditory cues from the playback room are responsible for reduced externalization when recording and playback room are not identical.

Finally, Appendix A described a study on the intelligibility of speech transmitted through mobile phones. The stimuli were recorded through three different commercially available mobile phones. Speech intelligibility was measured at four fixed SNRs in six normal-hearing listeners and it was tested how well three different well-established speech intelligibility models were able to predict the measured outcomes. Transmission through a mobile phone generally degraded speech intelligibility compared to the unprocessed speech signal recorded at

the position of the mobile phone microphone. The SRT_{70}s that were measured for the three phones showed small but statistically significant differences of up to 2 dB. The ranking of the speech intelligibility performance across the three phones with traffic noise and pub noise were identical to the same ranking observed with SSN. Of the three tested speech intelligibility prediction models, the speech-based envelope power spectrum (sEPSM) model (Jørgensen et al., 2013) performed best in the SSN condition, whereas it failed to account for the differences in speech intelligibility across noise types. The short-time objective intelligibility (STOI) model (Taal et al., 2011) yielded the best predictions across all conditions. The extended speech intelligibility index (ESII) model (Rhebergen et al., 2006) failed to account for the differences between the three phones. Overall, the study showed that even with this black-box approach, where the experimenters had no access to the internal processing of the mobile phones, current speech intelligibility models were able to predict the outcome of the listening experiments quite well, unlike speech quality models (such as the perceptual evaluation of speech quality model, PESQ) that usually show a floor effect when used to predict speech intelligibility. This is plausible since speech quality testing is usually done at much higher SNRs of +10 or even +30 dB, where speech intelligibility should be close to 100%. Considering this difference in the considered SNR range, it seems that speech intelligibility models could be a valuable extension of the toolkit for mobile phone testing, because they allow for the prediction of a meaningful measure of mobile phone performance at low SNRs.

7.2 Perspectives

Based on the findings from chapter 2, new opportunities for research and development arise. For example, using VSEs could facilitate the evaluation of new signal processing strategies in hearing aids, where comparative tests with end users under controlled and repeatable conditions would become possible early in the development process. This could accelerate the development process of new features and thus help improve the situation of hearing-impaired listeners, whose most frequent complaint still is the lack of understanding speech in complex acoustic environments. However, to be able to infer real-world performance from results measured in a VSE requires further experiments that are specifically designed to validate VSE systems. Such experiments might, for

example, investigate how the limited spatial resolution of the simulated sound field inside the loudspeaker array affects the outcome of experiments for which spatial perception is crucial. Furthermore, VSEs could be combined with systems for the presentation of visual stimuli to study the integration of audiovisual information and human behaviour in virtual reality environments, instead of "only" considering real-world acoustic scenes. VSEs also allow for studies that are difficult, if not impossible, to conduct in real-world environments in a controlled and repeatable manner, such as the investigation of listener preference for different kinds of hearing-aid processing in different environments, like a living room, a church, a train station or a car. Finally, they even allow demonstrations of rooms that do not even physically exist (yet), which could prove a powerful tool in the room acoustic design of new architectural spaces like, e.g., open plan offices or concert venues.

The sketches of the "auditory images" in chapter 3 demonstrated that even the "ideal" hearing aids used in that study substantially deteriorated the spatial perception of an acoustic scene, even in normal-hearing listeners. Considering the potential effects of a hearing loss on spatial perception and the impact of hearing-aid signal processing (like frequency-dependent amplification, dynamic-range compression and adaptive noise-reduction algorithms), the difficulties of hearing-impaired listeners can be expected to be even more pronounced than those of normal-hearing listeners. Even though a direct connection between the spatial perception of an acoustic scene and speech intelligibility remains to be understood in detail, a more realistic perception of an acoustic scene should be helpful in the everyday life of a hearing-impaired listener. It seems crucial to preserve natural binaural cues as much as possible in hearing aid processing, or to consider strategies to enhance such cues.

It seems that no clear distinction has so far been made between the percepts of distance perception and externalization. Some authors treat externalization as a 'crude approximation of distance' (Durlach et al., 1992) or define every distance percept that was reported to be below a certain distance (4 inches, 10 cm, etc.) as internalized (Begault and Wenzel, 1993). Other authors require the auditory image to be outside the head, correctly localized, and compact (Hassager et al., 2016; Gil-Carvajal et al., 2016) to consider it as externalized, consistent with the description of natural sound sources in Hartmann and Wit-

tenberg (1996). Furthermore, externalization has sometimes been treated as a binary percept, where an auditory event is considered as externalized when it occurs outside the head without any requirement in terms of correct localization or compactness (Durlach et al., 1992). Blauert (1997) stated that many studies on distance perception did not make any distinction between asking for the distance to the sound source vs. asking for the distance to the auditory event that is evoked by the source. A simple example to demonstrate this point would be to listen to a monophonic orchestra recording through headphones. Even though the signals at both ears will be identical, it would be fairly easy to estimate the distance between the microphone position and the orchestra, but the listener will most likely not perceive the auditory event (far) outside their head. An interesting experiment to follow up would be to consider signals that emphasize the differences between distance and externalization perception. Signals can be created for which a source distance can be estimated quite well, even though the auditory event will be inside the listener's head, such as in the aforementioned example of the monophonic orchestra recording. There are also signals that are clearly perceived outside the listeners' head, i.e., externalized, and at a well-defined distance. This could be a binaural recording of the same orchestra. Another group of signals would be perceived internalized and it would be impossible to estimate a distance, e.g., an anechoic, monophonic speech recording, presented diotically. Finally, there might even be signals that evoke an auditory event, which is outside the head, but not localizable, such that it is impossible to estimate a distance.

Chapter 6 showed that the acoustic properties of the listening room in which a listener carries out the experiment is critical for the perception of externalization. Considering the recent increase in popularity of binaural technology both for entertainment (broadcast, gaming, virtual reality and augmented reality content) and for research purposes, investigating this topic further would be interesting to eventually realize the best possible user experience. Also here, the interaction between auditory and visual stimuli, e.g., in virtual reality systems, would be very interesting to study, and it might well be that the availability of plausible visual information that is congruent with the auditory stimuli would help consolidate the externalized perception of sound.

Bibliography

ANSI S3.5-1969 (1969). *American National Standard methods for the calculation of the Articulation Index*. Acoustical Society of America. New York, USA.

ANSI S3.5-1997 (1997). *Methods for the calculation of the speech intelligibility index*. Acoustical Society of America. New York.

Akeroyd, M. A. and W. M. Whitmer (2016). "Spatial Hearing and Hearing Aids". In: *Hearing Aids*. Springer, pp. 181–215.

Akeroyd, M. A., S. Gatehouse, and J. Blaschke (2007). "The detection of differences in the cues to distance by elderly hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 121, p. 1077.

Alais, D. and D. Burr (2004). "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration". In: *Current Biology* 14.3, pp. 257–262.

Amlani, A. M., J. L. Punch, and T. Y. Ching (2002). "Methods and applications of the audibility index in hearing aid selection and fitting". In: *Trends in amplification* 6.3, pp. 81–129.

Anderson, P. W. and P. Zahorik (2014). "Auditory/visual distance estimation: accuracy and variability". In: *Frontiers in psychology* 5.1097.

Arweiler, I. and J. Buchholz (2011). "The influence of spectral characteristics of early reflections on speech intelligibility". In: *Journal of the Acoustical Society of America* 130, pp. 996–1005.

Beerends, J. G., R. v. Buuren, J. v. Vugt, and J. Verhave (2009). "Objective speech intelligibility measurement on the basis of natural speech in combination with perceptual modeling". In: *Journal of the Audio Engineering Society* 57.5, pp. 299–308.

Begault, D. R. and E. M. Wenzel (1993). "Headphone localization of speech". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 35.2, pp. 361–376.

Begault, D. R., E. M. Wenzel, and M. R. Anderson (2001). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related

transfer functions on the spatial perception of a virtual speech source". In: *Journal of the Audio Engineering Society* 49.10, pp. 904–916.

Bench, J., Å. Kowal, and J. Bamford (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children". In: *British journal of audiology* 13.3, pp. 108–112.

Berkhout, A. J., D. de Vries, and P. Vogel (1993). "Acoustic control by wave field synthesis". In: *The Journal of the Acoustical Society of America* 93, pp. 2764–2778.

Best, V., S. Kalluri, S. McLachlan, S. Valentine, B. Edwards, and S. Carlile (2010). "A comparison of CIC and BTE hearing aids for three-dimensional localization of speech". In: *International Journal of Audiology* 49.10, pp. 723–732.

Best, V., S. Carlile, N. Kopčo, and A. van Schaik (2011). "Localization in speech mixtures by listeners with hearing loss". In: *The Journal of the Acoustical Society of America* 129.5, EL210–EL215.

Best, V., G. Keidser, J. M. Buchholz, and K. Freeston (2015). "An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment". In: *International Journal of Audiology* 54.10, pp. 1–9.

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

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

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

Blauert, J and W Lindemann (1986). "Spatial mapping of intracranial auditory events for various degrees of interaural coherence". In: *The Journal of the Acoustical Society of America* 79.3, pp. 806–813.

Blauert, J. (2005). *Communication acoustics*. Vol. 2. Springer.

Bolia, R. S., W. T. Nelson, M. A. Ericson, and B. D. Simpson (2000). "A speech corpus for multitalker communications research". In: *The Journal of the Acoustical Society of America* 107.2, pp. 1065–1066.

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

Boyd, A., W. Whitmer, J. Soraghan, and M. Akeroyd (2012). "Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers". In: *The Journal of the Acoustical Society of America* 131.3, EL268–EL274.

Bradley, J., H. Sato, and M. Picard (2003). "On the importance of early reflections for speech in rooms". In: *The Journal of the Acoustical Society of America* 113, p. 3233.

Bradley, J. S., R Reich, and S. Norcross (1999). "A just noticeable difference in C_{50} for speech". In: *Applied Acoustics* 58.2, pp. 99–108.

Brand, T. and B. Kollmeier (2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests". In: *The Journal of the Acoustical Society of America* 111, pp. 2801–2810.

Brimijoin, W. O., A. W. Boyd, and M. A. Akeroyd (2013). "The Contribution of Head Movement to the Externalization and Internalization of Sounds". In: *PLoS ONE* 8.12, e83068.

Bronkhorst, A. and R Plomp (1992). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing". In: *The Journal of the Acoustical Society of America* 92.6, pp. 3132–3139.

Bronkhorst, A. W. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". In: *Acta Acustica united with Acustica* 86.1, pp. 117–128.

Bronkhorst, A. W. and T. Houtgast (1999). "Auditory distance perception in rooms". In: *Nature* 397.6719, pp. 517–520.

Brown, A. D., F. A. Rodriguez, C. D. Portnuff, M. J. Goupell, and D. J. Tollin (2016). "Time-varying distortions of binaural information by bilateral hearing aids: effects of nonlinear frequency compression". In: *Trends in Hearing* 20, pp. 1–15.

Brungart, D. S. and B. D. Simpson (2002). "Within-ear and across-ear interference in a cocktail-party listening task". In: *The Journal of the Acoustical Society of America* 112.6, pp. 2985–2995.

Brungart, D. S., N. I. Durlach, and W. M. Rabinowitz (1999). "Auditory localization of nearby sources. II. Localization of a broadband source". In: *The Journal of the Acoustical Society of America* 106.4, pp. 1956–1968.

Buchholz, J., J. Blauert, and J. Mourjopoulos (2001). "Room Masking: Understanding and Modelling the Masking of Reflections in Rooms". In: *Audio Engineering Society Convention 110*.

Butler, R. A. and K. Belendiuk (1977). "Spectral cues utilized in the localization of sound in the median sagittal plane". In: *The Journal of the Acoustical Society of America* 61.5, pp. 1264–1269.

Butler, R. A., E. T. Levy, and W. D. Neff (1980). "Apparent distance of sounds recorded in echoic and anechoic chambers." In: *Journal of Experimental Psychology: Human Perception and Performance* 6.4, p. 745.

Calcagno, E. R., E. L. Abregu, and C. E. Manuel (2012). "The role of vision in auditory distance perception". In: *Perception* 41.2, p. 175.

Catic, J., S. Santurette, J. M. Buchholz, F. Gran, and T. Dau (2013). "The effect of interaural-level-difference fluctuations on the externalization of sound". In: *The Journal of the Acoustical Society of America* 134.2, pp. 1232–1241.

Catic, J., S. Santurette, and T. Dau (2015). "The role of reverberation-related binaural cues in the externalization of speech". In: *The Journal of the Acoustical Society of America* 138.2, pp. 1154–1167.

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

Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and with two ears". In: *The Journal of the acoustical society of America* 25.5, pp. 975–979.

Christensen, C. L. (2013). *ODEON Room Acoustics Software, Version 12, User Manual, ODEON A/S, Kgs. Lyngby, Denmark*.

Coleman, P. D. (1963). "An analysis of cues to auditory depth perception in free space." In: *Psychological Bulletin* 60.3, p. 302.

Collin, B. and M. Lavandier (2013). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers". In: *The Journal of the Acoustical Society of America* 134.2, pp. 1146–1159.

Cox, T. J., W. Davies, and Y. W. Lam (1993). "The sensitivity of listeners to early sound field changes in auditoria". In: *Acta Acustica united with Acustica* 79.1, pp. 27–41.

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

Cubick, J., S. Santurette, and T. Dau (2014). “Influence of high-frequency audibility on the perceived distance of sounds”. In: *7th Forum Acusticum*.

Cubick, J., C. Sánchez Rodríguez, W. Song, and E. N. MacDonald (2015). “Comparison of binaural microphones for externalization of sounds”. In: *Proceedings of International Conference on Spatial Audio 2015*.

Daniel, J., R. Nicol, and S. Moreau (2003). “Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging”. In: *Preprints-Audio Engineering Society*.

Dau, T., J. Verhey, and A. Kohlrausch (1999). “Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers”. In: *The Journal of the Acoustical Society of America* 106.5, pp. 2752–2760.

Dillon, H. (2001). *Hearing aids*. Thieme Medical Pub.

Dubbelboer, F. and T. Houtgast (2008). “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility”. In: *The Journal of the Acoustical Society of America* 124.6, pp. 3937–3946.

Durlach, N., C. Thompson, and H. Colburn (1981). “Binaural interaction in impaired listeners: A review of past research”. In: *Audiology* 20.3, pp. 181–211.

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

Durlach, N. I. (1972). “Binaural signal detection - Equalization and cancellation theory.” In: *Foundations of modern auditory theory. Volume 2*. Academic Press, Inc., pp. 371–462.

Durlach, N. I. et al. (1992). “On the externalization of auditory images”. In: *Presence: Teleoperators and Virtual Environments* 1.2, pp. 251–257.

EN ISO 3382-1 (2009). *Acoustics – Measurement of room acoustic parameters – part 1: Performance spaces*.

ETSI EG 202 396-1 (2008). *Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database*. European Telecommunications Standards Institute. Sophia Antiplis Cedex, France.

ETSI TR 102 251 (2003). *Speech processing, transmission and quality aspects (STQ); anonymous test report from 2nd speech quality test event 2002*. Eu-

ropean Telecommunications Standards Institute. Sophia Antiplis Cedex, France.

Erulkar, S. (1972). "Comparative aspects of spatial localization of sound." In: *Physiological Reviews* 52.1, pp. 237–360.

Favrot, S. and J. Buchholz (2009a). "Distance Perception in Loudspeaker-Based Room Auralization". In: *Audio Engineering Society Convention 127*.

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

Favrot, S. and J. M. Buchholz (2009b). "Validation of a loudspeaker-based room auralization system using speech intelligibility measures". In: *Audio Engineering Society Convention 126*. Audio Engineering Society.

Festen, J. M. and R. Plomp (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing". In: *The Journal of the Acoustical Society of America* 88.4, pp. 1725–1736.

Fluitt, K., T. Mermagen, and T. Letowski (2014). "Auditory Distance Estimation in an Open Space". In: *Soundscape semiotics – localisation and categorisation*. Ed. by H Glotin. Rijeka (Croatia): InTech.

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

Gerzon, M. A. (1973). "Periphony: With-height sound reproduction". In: *Journal of the Audio Engineering Society* 21.1, pp. 2–10.

Gierlich, H.-W. and F Kettler (2006). "Advanced speech quality testing of modern telecommunication equipment: An overview". In: *Signal processing* 86.6, pp. 1327–1340.

Gil-Carvajal, J. C., J. Cubick, S. Santurette, and T. Dau (2016). "Spatial Hearing with Incongruent Visual or Auditory Room Cues". In: *Scientific Reports* 6.37342.

Gil Carvajal, J. C. (2015). "The influence of visual cues on sound externalization". MA thesis. Technical University of Denmark.

Glyde, H., J. Buchholz, H. Dillon, V. Best, L. Hickson, and S. Cameron (2013). "The effect of better-ear glimpsing on spatial release from masking". In: *The Journal of the Acoustical Society of America* 134.4, pp. 2937–2945.

Green, D. and T. Birdsall (1964). "The effect of vocabulary size". In: *Signal Detection and Recognition by Human Observers*. Ed. by J. A. Swets. New York, USA: John Wiley & Sons, pp. 609–619.

Grimm, G., S. Ewert, and V. Hohmann (2015). “Evaluation of spatial audio reproduction schemes for application in hearing aid research”. In: *Acta Acustica united with Acustica* 101.4, pp. 842–854.

Hagerman, B. (1982). “Sentences for testing speech intelligibility in noise”. In: *Scandinavian Audiology* 11.2, pp. 79–87.

Hagerman, B. and Å. Olofsson (2004). “A method to measure the effect of noise reduction algorithms using simultaneous speech and noise”. In: *Acta Acustica united with Acustica* 90.2, pp. 356–361.

Hammershøi, D. and H. Møller (1996). “Sound transmission to and within the human ear canal”. In: *The Journal of the Acoustical Society of America* 100.1, pp. 408–427.

Hammershøi, D. and H. Møller (2005). “Binaural technique—Basic methods for recording, synthesis, and reproduction”. In: *Communication Acoustics*. Springer, pp. 223–254.

Hänsler, E. (1994). “The hands-free telephone problem: an annotated bibliography update”. In: *Annales des Télécommunications*. Vol. 49. 7-8. Springer, pp. 360–367.

Hartmann, W. and A. Wittenberg (1996). “On the externalization of sound images”. In: *The Journal of the Acoustical Society of America* 99, p. 3678.

Hassager, H. G., F. Gran, and T. Dau (2016). “The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment”. In: *The Journal of the Acoustical Society of America* 139.5, pp. 2992–3000.

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

Hawley, M. L., R. Y. Litovsky, and H. S. Colburn (1999). “Speech intelligibility and localization in a multi-source environment”. In: *The Journal of the Acoustical Society of America* 105.6, pp. 3436–3448.

Hawley, M. L., R. Y. Litovsky, and J. F. Culling (2004). “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer”. In: *The Journal of the Acoustical Society of America* 115.2, pp. 833–843.

Hodgson, M., N. York, W. Yang, and M. Bliss (2008). “Comparison of Predicted, Measured and Auralized Sound Fields with Respect to Speech Intelligibility in Classrooms Using CATT-Acoustic and ODEON”. In: *Acta Acustica united with Acustica* 94.6, pp. 883–890.

Hofman, P. M., J. G. Van Riswick, and A. J. Van Opstal (1998). "Relearning sound localization with new ears". In: *Nature neuroscience* 1.5, pp. 417–421.

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

Houtgast, T. and H. J. Steeneken (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria". In: *The Journal of the Acoustical Society of America* 77.3, pp. 1069–1077.

IEC 60268-13 (1985). *Sound system equipment - part 13: Listening tests on loudspeakers*. International Electrotechnical Commission. Geneva, Switzerland.

IEC 60268-16:2003 (2003). *Sound system equipment - part 16: Objective rating of speech intelligibility by speech transmission index*. International Electrotechnical Commission. Geneva, Switzerland.

ITU-R BS.1534-2 (2014). *Method for the subjective assessment of intermediate quality level of audio systems*. International Telecommunication Union - Radiocommunication Sector.

ITU-T P.800 (1996). *Methods for subjective determination of transmission quality - series P: Telephone transmission quality; methods for objective and subjective assessment of quality*. International Telecommunication Union. Geneva, Switzerland.

ITU-T P.830 (1996). *Subjective performance assessment of telephone-band and wideband digital codecs - telephone transmission quality; methods for objective and subjective assessment of quality*. International Telecommunication Union. Geneva, Switzerland.

ITU-T P.862 (2001). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*. International Telecommunication Union. Geneva, Switzerland.

Jack, C. E. and W. R. Thurlow (1973). "Effects of degree of visual association and angle of displacement on the 'ventriloquism' effect." In: *Perceptual and motor skills*.

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

Jørgensen, S. and T. Dau (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selec-

tive processing". In: *The Journal of the Acoustical Society of America* 130.3, pp. 1475–1487.

Jørgensen, S., S. D. Ewert, and T. Dau (2013). "A multi-resolution envelope-power based model for speech intelligibility". In: *The Journal of the Acoustical Society of America* 134.1, pp. 436–446.

Jørgensen, S., J. Cubick, and T. Dau (2015). "Speech Intelligibility Evaluation for Mobile Phones". In: *Acta Acustica united with Acustica* 101.5, pp. 1016–1025.

Keidser, G. et al. (2006). "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers". In: *International Journal of Audiology* 45.10, pp. 563–579.

Keidser, G., H. Dillon, J. Mejia, and C.-V. Nguyen (2013). "An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function". In: *International Journal of Audiology* 52.11, pp. 795–800.

Kidd, G., C. R. Mason, A. Brughera, and W. M. Hartmann (2005). "The role of reverberation in release from masking due to spatial separation of sources for speech identification". In: *Acta acustica united with acustica* 91.3, pp. 526–536.

Kidd, G., C. R. Mason, V. M. Richards, F. J. Gallun, and N. I. Durlach (2008). "Informational masking". In: *Auditory perception of sound sources*. Springer, pp. 143–189.

Kim, S.-M. and W. Choi (2005). "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach". In: *The Journal of the Acoustical Society of America* 117.6, pp. 3657–3665.

Kleiner, M (1981). "Speech intelligibility in real and simulated sound fields". In: *Acta Acustica united with Acustica* 47.2, pp. 55–71.

Kleiner, M., B.-I. Dalenbäck, and P. Svensson (1993). "Auralization-an overview". In: *Journal of the Audio Engineering Society* 41.11, pp. 861–875.

Kolarik, A. J., B. C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan (2015). "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss". In: *Attention, Perception, & Psychophysics*, pp. 1–23.

Kondo, K. (2011). *Estimation of speech intelligibility using perceptual speech quality scores*. Ed. by P. I. Ipsic. Rijeka, Croatia: InTech Open Access Publisher. Chap. 8, pp. 155–174.

Kopčo, N. and B. G. Shinn-Cunningham (2011). “Effect of stimulus spectrum on distance perception for nearby sources”. In: *The Journal of the Acoustical Society of America* 130.3, pp. 1530–1541.

Koski, T., V. Sivonen, and V. Pulkki (2013). “Measuring speech intelligibility in noisy environments reproduced with parametric spatial audio”. In: *Audio Engineering Society Convention 135*. Audio Engineering Society.

Kulkarni, A., H. Colburn, et al. (1998). “Role of spectral detail in sound-source localization”. In: *Nature* 396.6713, pp. 747–749.

Kuttruff, H. (2009). *Room acoustics*. Crc Press.

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

Lavandier, M., S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin (2012). “Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources”. In: *The Journal of the Acoustical Society of America* 131.1, pp. 218–231.

Laws, P (1973). “Entfernungshören und das Problem der Im-Kopf-Lokalisiertheit von Hörereignissen (Auditory Distance Perception and the Problem of “In-Head Localization” of Sound Images)”. In: *Acustica* 29.5, pp. 243–259.

Levy, E. T. and R. A. Butler (1978). “Stimulus factors which influence the perceived externalization of sound presented through headphones.” In: *Journal of Auditory Research* 18.1, pp. 41 –50.

Lindevald, I. and A. Benade (1986). “Two-ear correlation in the statistical sound fields of rooms”. In: *The Journal of the Acoustical Society of America* 80.2, pp. 661–664.

Litovsky, R., H. Colburn, W. Yost, and S. Guzman (1999). “The precedence effect”. In: *The Journal of the Acoustical Society of America* 106, pp. 1633–1654.

Little, A. D., D. H. Mershon, and P. H. Cox (1992). “Spectral content as a cue to perceived auditory distance”. In: *Perception* 21.3, pp. 405–416.

Liu, W. M., K. A. Jellyman, J. S. Mason, and N. W. Evans (2006). “Assessment of objective quality measures for speech intelligibility estimation”. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. IEEE, pp. 1225–1228.

Lőcsei, G., J. H. Pedersen, S. Laugesen, S. Santurette, T. Dau, and E. N. MacDonald (2016). “Temporal Fine-Structure Coding and Lateralized Speech

Perception in Normal-Hearing and Hearing-Impaired Listeners". In: *Trends in Hearing* 20, pp. 1–15.

Loomis, J. M., R. L. Klatzky, and R. G. Golledge (1999). "Auditory distance perception in real, virtual, and mixed environments". In: *Mixed reality: Merging real and virtual worlds*, pp. 201–214.

Lorenzi, C., S. Gatehouse, and C. Lever (1999). "Sound localization in noise in hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 105.6, pp. 3454–3463.

Lounsbury, B. and R. Butler (1979). "Estimation of distances of recorded sounds presented through headphones". In: *Scandinavian Audiology* 8.3, pp. 145–149.

Luts, H. et al. (2010). "Multicenter evaluation of signal enhancement algorithms for hearing aids". In: *The Journal of the Acoustical Society of America* 127.3, pp. 1491–1505.

Marrone, N., C. R. Mason, and G. Kidd Jr (2008). "Tuning in the spatial dimension: Evidence from a masked speech identification task". In: *The Journal of the Acoustical Society of America* 124.2, pp. 1146–1158.

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

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

Mattila, V (2003). "Measurement of the performance of noise suppression in mobile telecommunication networks". In: *Acta Acustica united with Acoustica* 89, S114–S115.

McDonald, J. J., W. A. Teder-Sälejärvi, and L. M. Ward (2001). "Multisensory integration and crossmodal attention effects in the human brain". In: *Science* 292.5523, pp. 1791–1791.

McLoughlin, I., Z. Ding, and E. C. Tan (2002). "Intelligibility evaluation of GSM coder for Mandarin speech using CDRT". In: *Speech Communication* 38.1, pp. 161–165.

Mershon, D. H. and L. E. King (1975). "Intensity and reverberation as factors in the auditory perception of egocentric distance". In: *Perception & Psychophysics* 18.6, pp. 409–415.

Middlebrooks, J. C. and D. M. Green (1991). “Sound localization by human listeners”. In: *Annual review of psychology* 42.1, pp. 135–159.

Mills, A. W. (1958). “On the minimum audible angle”. In: *The Journal of the Acoustical Society of America* 30.4, pp. 237–246.

Minnaar, P., S. K. Olesen, F. Christensen, and H. Møller (2001). “Localization with binaural recordings from artificial and human heads”. In: *Journal of the Audio Engineering Society* 49.5, pp. 323–336.

Minnaar, P., C. Breitsprecher, and M. Holmberg (2011). “Simulating Complex Listening Environments in the Laboratory for Testing Hearing Aids”. In: *Proc. Forum Acusticum*.

Minnaar, P., S. F. Albeck, C. S. Simonsen, B. Søndersted, S. A. D. Oakley, and J. Bennedbæk (2013). “Reproducing Real-Life Listening Situations in the Laboratory for Testing Hearing Aids”. In: *Audio Engineering Society Convention 135*. Audio Engineering Society.

Møller, H. (1992). “Fundamentals of binaural technology”. In: *Applied acoustics* 36.3, pp. 171–218.

Møller, H., D. Hammershøi, C. B. Johnson, and M. F. Sørensen (1999). “Evaluation of artificial heads in listening tests”. In: *Journal of the Audio Engineering Society* 47.3, pp. 83–100.

Moore, B. C. (2003). “An introduction to the psychology of hearing”. In:

Moore, B. C. and B. R. Glasberg (1983). “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns”. In: *The Journal of the Acoustical Society of America* 74.3, pp. 750–753.

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

Nielsen, J. B. and T. Dau (2011). “The Danish hearing in noise test”. In: *International Journal of Audiology* 50.3, pp. 202–208.

Noble, W., D. Byrne, and B. Lepage (1994). “Effects on sound localization of configuration and type of hearing impairment”. In: *The Journal of the Acoustical Society of America* 95.2, pp. 992–1005.

Ohl, B., S. Laugesen, J. Buchholz, and T. Dau (2010). “Externalization versus Internalization of Sound in Normal-hearing and Hearing-impaired Listeners”. In: *Fortschritte der Akustik, DAGA 2010, 36. Jahrestagung für Akustik*.

Ohl, B. (2009). “Externalization versus Internalization of Sound in Normal-hearing and Hearing-impaired Listeners”. Msc Thesis. Technical University of Denmark.

Oreinos, C. and J. M. Buchholz (2015). "Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones". In: *The Journal of the Acoustical Society of America* 137.6, pp. 3447–3465.

Parseihian, G., C. Jouffrais, and B. F. Katz (2014). "Reaching nearby sources: comparison between real and virtual sound and visual targets". In: *Frontiers in neuroscience* 8.269, pp. 1–13.

Patterson, R., I. Nimmo-Smith, J. Holdsworth, and P. Rice (1987). "An efficient auditory filterbank based on the gammatone function". In: *A Meeting of the IOC Speech Group on Auditory Modelling at RSRE*. Vol. 2. 7.

Pedersen, E. (2007). "Bestemmelse af taleforståelighed i støj (Determination of speech intelligibility in noise)". MA thesis. Odense: Syddansk Universitet.

Plack, C. J. (2005). *The sense of hearing*. Lawrence Erlbaum Associates Publishers.

Plack, C. J. (2013). *The sense of hearing*. Psychology Press.

Plenge, G. (1972). "Über das Problem der Im-Kopf-Lokalisation [On the problem of In Head Localization]". In: *Acustica* 26.5, pp. 241–252.

Plenge, G (1974). "On the differences between localization and lateralization". In: *The Journal of the Acoustical Society of America* 56.3, pp. 944–951.

Plomp, R (1976). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)". In: *ACUSTICA* 34.4, pp. 200–211.

Pulkki, V. (2007). "Spatial sound reproduction with directional audio coding". In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516.

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

Rhebergen, K. S. and N. J. Versfeld (2005). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners". In: *The Journal of the Acoustical Society of America* 117.4, pp. 2181–2192.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise". In: *The Journal of the Acoustical Society of America* 120.6, pp. 3988–3997.

Rindel, J. (2000). "The use of computer modeling in room acoustics". In: *Journal of Vibroengineering* 3.4, pp. 41–72.

Rindel, J. and C. Christensen (2003). "Room Acoustic Simulation and Auralization–How Close can we get to the Real Room?" In: *WESPAC 8, The Eighth Western Pacific Acoustics Conference, Melbourne*. Citeseer.

Roffler, S. K. and R. A. Butler (1968). "Factors that influence the localization of sound in the vertical plane". In: *The Journal of the Acoustical Society of America* 43.6, pp. 1255–1259.

Rychtáriková, M., T. van den Bogaert, G. Vermeir, and J. Wouters (2011). "Perceptual validation of virtual room acoustics: Sound localisation and speech understanding". In: *Applied Acoustics* 72, pp. 196–204.

Sakamoto, N., T. Gotoh, and Y. Kimura (1976). "On -Out-of-Head Localization in Headphone Listening". In: *J. Audio Eng. Soc* 24.9, pp. 710–716.

Schoeffler, M., J. Gernert, M. Neumayer, S. Westphal, and J. Herre (2015). "On the validity of virtual reality-based auditory experiments: a case study about ratings of the overall listening experience". In: *Virtual Reality*, pp. 1–20.

Schroeder, M. R. and K. H. Kuttruff (1962). "On Frequency Response Curves in Rooms. Comparison of Experimental, Theoretical, and Monte Carlo Results for the Average Frequency Spacing between Maxima". In: *The Journal of the Acoustical Society of America* 34.1, pp. 76–80.

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

Shinn-Cunningham, B. G. (2008). "Object-based auditory and visual attention". In: *Trends in cognitive sciences* 12.5, pp. 182–186.

Spencer, N. J., M. L. Hawley, and H. S. Colburn (2016). "Relating interaural difference sensitivities for several parameters measured in normal-hearing and hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 140.3, pp. 1783–1799.

Stein, B. E. and T. R. Stanford (2008). "Multisensory integration: current issues from the perspective of the single neuron". In: *Nature Reviews Neuroscience* 9.4, pp. 255–266.

Strelcyk, O. and T. Dau (2009). "Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing". In: *The Journal of the Acoustical Society of America* 125.5, pp. 3328–3345.

Strutt, J. W. (1907). "On our perception of sound direction". In: *Philosophical Magazine and Journal of Science* 13.6, pp. 214–232.

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2011). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 2125–2136.

Thompson, S. P. (1882). "On the function of the two ears in the perception of space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.83, pp. 406–416.

Toole, F. (1970). "In-Head Localization of Acoustic Images". In: *The Journal of the Acoustical Society of America* 48, p. 943.

Udesen, J., T. Piechowiak, and F. Gran (2014). "Vision Affects Sound Externalization". In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society.

Udesen, J., T. Piechowiak, and F. Gran (2015). "The Effect of Vision on Psychoacoustic Testing with Headphone-Based Virtual Sound". In: *Journal of the Audio Engineering Society* 63.7/8, pp. 552–561.

Valente, M., D. Fabry, and L. G. Potts (1995). "Recognition of speech in noise with hearing aids using dual microphones". In: *Journal of the American Academy of Audiology* 6.6.

Van den Bogaert, T., T. J. Klasen, M. Moonen, L. van Deun, and J. Wouters (2006). "Horizontal localization with bilateral hearing aids: Without is better than with". In: *The Journal of the Acoustical Society of America* 119.1, pp. 515–526.

Van den Bogaert, T., S. Doclo, J. Wouters, and M. Moonen (2008). "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids". In: *The Journal of the Acoustical Society of America* 124.1, pp. 484–497.

Vorländer, M. and J. E. Summers (2008). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality*. 1st ed. Springer-Verlag, Berlin.

Wagener, K. (2003). "Factors Influencing Sentence Intelligibility in Noise". PhD thesis. Universität Oldenburg.

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

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

procedure and masking parameters". In: *International Journal of Audiology* 44.3, pp. 144–156.

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

Wan, R., N. I. Durlach, and H. S. Colburn (2014). "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers". In: *The Journal of the Acoustical Society of America* 136.2, pp. 768–776.

Wang, D. (2005). "On ideal binary mask as the computational goal of auditory scene analysis". In: *Speech separation by humans and machines*. Ed. by P. Divenyi. USA: Springer. Chap. 12, pp. 181–197.

Wenzel, E. M., M. Arruda, D. J. Kistler, and F. L. Wightman (1993). "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123.

Werner, S. and A. Siegel (2012). "Effects of binaural auralization via headphones on the perception of acoustic scenes". In: *Proc. of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research*, pp. 215–222.

Westerlund, N., M. Dahl, and I. Claesson (2005). "Speech enhancement for personal communication using an adaptive gain equalizer". In: *Signal processing* 85.6, pp. 1089–1101.

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

Westermann, A. and J. M. Buchholz (2017a). "The effect of hearing loss on source-distance dependent speech intelligibility in rooms". In: *The Journal of the Acoustical Society of America* 141.2, EL140–EL145.

Westermann, A. and J. M. Buchholz (2017b). "The effect of nearby maskers on speech intelligibility in reverberant, multi-talker environments". In: *The Journal of the Acoustical Society of America* 141.3, pp. 2214–2223.

Westermann, A. and J. M. Buchholz (2015b). "The effect of spatial separation in distance on the intelligibility of speech in rooms". In: *The Journal of the Acoustical Society of America* 137.2, pp. 757–767.

Wettschureck, R, G Plenge, and F Lehringer (1973). “Entfernungswahrnehmung beim natürlichen Hören sowie bei kopfbezogener Stereofonie”. In: *Acustica* 29, pp. 260–272.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.

Wiggins, I. and B. Seeber (2012). “Effects of Dynamic-Range Compression on the Spatial Attributes of Sounds in Normal-Hearing Listeners”. In: *Ear and hearing* 33.3, pp. 399–410.

Wightman, F. L. and D. J. Kistler (1989a). “Headphone simulation of free-field listening. I: Stimulus synthesis”. In: *The Journal of the Acoustical Society of America* 85.2, pp. 858–867.

Wightman, F. L. and D. J. Kistler (1989b). “Headphone simulation of free-field listening. II: Psychophysical validation”. In: *The Journal of the Acoustical Society of America* 85.2, pp. 868–878.

Wightman, F. L. and D. J. Kistler (1992). “The dominant role of low-frequency interaural time differences in sound localization”. In: *The Journal of the Acoustical Society of America* 91.3, pp. 1648–1661.

Wong, L. L., E. H. Ng, and S. D. Soli (2012). “Characterization of speech understanding in various types of noise”. In: *The Journal of the Acoustical Society of America* 132.4, pp. 2642–2651.

Wouters, J., L. Litière, and A. Van Wieringen (1999). “Speech intelligibility in noisy environments with one-and two-microphone hearing aids”. In: *International Journal of Audiology* 38.2, pp. 91–98.

Yang, W. and M. Hodgson (2007). “Validation of the Auralization Technique: Comparative Speech-Intelligibility Tests in Real and Virtual Classrooms”. In: *Acta Acustica united with Acustica* 93.6, pp. 991–999.

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

Zahorik, P. (2002a). “Assessing auditory distance perception using virtual acoustics”. In: *The Journal of the Acoustical Society of America* 111, pp. 1832–1846.

Zahorik, P. (2002b). “Auditory display of sound source distance”. In: *Proceedings of the International Conference on Auditory Display*, pp. 326–332.

von Békésy, G. (1938). “Über die Entstehung der Entfernungsempfindung beim Hören.” In: *Akustische Zeitschrift*.

A

Speech intelligibility evaluation for mobile phones[a]

Abstract

In the development process of modern telecommunication systems, such as mobile phones, it is common practice to use computer models to objectively evaluate the transmission quality of the system, instead of time-consuming perceptual listening tests. Such models have typically focused on the quality of the transmitted speech, while little or no attention has been provided to speech intelligibility. The present study investigated to what extent three state-of-the art speech intelligibility models could predict the intelligibility of noisy speech transmitted through mobile phones. Sentences from the Danish Dantale II speech material were mixed with three different kinds of background noise, transmitted through three different mobile phones, and recorded at the receiver via a local network simulator. The speech intelligibility of the transmitted sentences was assessed by six normal-hearing listeners and model predictions were compared to the perceptual data. Statistically significant differences between the intelligibility of the three phones were found in stationary speech-shaped noise. A good correspondence between the measured data and the predictions from one of the three models was found in all the considered conditions. Overall, the results suggest that speech intelligibility models inspired by auditory signal processing can be useful for the objective evaluation of speech transmission through mobile phones.

[a] This chapter is based on Jørgensen, Cubick, and Dau (2015).

A.1 Introduction

Speech transmission through modern telecommunication devices, such as mobile phones, has traditionally been evaluated mainly in terms of speech quality. One reason for the focus on speech quality, rather than speech intelligibility, might be that mobile phone communication typically occurs at high signal-to-noise ratios (SNR) where speech intelligibility is not compromised. In situations with very poor intelligibility, people generally either terminate the conversation or put themselves in a position where the SNR is increased. However, in other mobile telecommunication situations, e.g., when the talker is situated in a car or a train, the transmitted speech signal can, in fact, be largely affected by the presence of background noise surrounding the talker, which is difficult to move away from. The listener at the receiving end might have difficulty understanding the transmitted speech, even if he or she was situated in a quiet room. In such conditions, estimates of speech intelligibility could provide an additional important performance parameter of the mobile telecommunication system, in addition to a speech quality evaluation.

The perceptual evaluation of speech quality and intelligibility, based on listening tests, has generally been considered to be more reliable than objective prediction models for complex environmental or conversational conditions (Gierlich and Kettler, 2006). However, perceptual evaluations are very time consuming, often requiring many listeners and hours of testing. Thus, a reliable objective tool for predicting speech intelligibility performance in a given acoustic condition would be very valuable in the development process of new telecommunication systems. The objective evaluation method recommended by the International Telecommunication Union is the "perceptual evaluation of speech quality" (PESQ; ITU-T P.862, 2001) model. This model was developed specifically for evaluation of speech quality and has been shown to correlate well with the perceptual quality ratings quantified by the mean opinion score (MOS; ITU-T P.800, 1996). Several studies have attempted to use PESQ for predicting speech intelligibility in addition to speech quality (e.g., Liu et al., 2006; Beerends et al., 2009; Kondo, 2011). For example, Liu et al. (2006) reported a good correlation of the PESQ metric to perceptual intelligibility scores in conditions with Gaussian noise and different kinds of noise suppression and speech coding strategies. However, the listening test used by Liu et al. (2006)

was based on a rating of listening effort, rather than a quantitative measure of the number of speech items that had been understood. It is therefore unclear to what extent the perceptual data reported by Liu et al. (2006) reflect aspects of quality rather than intelligibility. In contrast, Beerends et al. (2009) found a poor correlation of PESQ predictions to perceptual intelligibility scores of consonant-vowel-consonant stimuli for a large variety of conditions, including telecom distortions, such as low-bit-rate speech coding, bandwidth limitation, different types of background noise (white, babble, car), multiplicative noise, and room response distortions. The PESQ score was generally found to be at floor level for all conditions, while the perceptual score ranged from 15 to 90% correct. Similarly, Kondo (2011) observed a floor effect when comparing PESQ predictions to perceptual speech intelligibility obtained using a dynamic rhyme test in conditions with white noise, pseudo-speech noise, and babble noise. The perceptual intelligibility scores were obtained in the range between 20 to 95% correct. Thus, the PESQ metric as provided in ITU-T P.862 (2001) appears to be inappropriate for general speech intelligibility prediction and it might, therefore, be advantageous to consider prediction models designed for speech intelligibility.

Several objective speech intelligibility metrics have been proposed, the first of which and most widely used one is the articulation index (AI; ANSI S3.5-1969, 1969). The AI essentially measures the amount of audible speech, based on information about the SNR and the human hearing threshold in several frequency bands that cover the overall frequency spectrum of speech. The AI model was later modified to include aspects of hearing loss, and was published as a new standard under the name "speech intelligibility index" (SII; ANSI S3.5-1997, 1997). However, the SII is inherently limited to conditions with a stationary background noise, and fails in more realistic conditions with fluctuating noise backgrounds, such as speech babble in a cafeteria. To overcome this limitation, Rhebergen and Versfeld (2005) presented the extended speech intelligibility index (ESII), which could account for the speech intelligibility in conditions with various types of fluctuating noise interferers. However, the AI, SII, and ESII models were designed to account for the effects of reduced bandwidth of early telecommunication systems, and might not be directly applicable to conditions where the noisy speech mixture has been processed by nonlinear noise reduction and low bit-rate coding used in modern telecommunication

systems.

Recently, an alternative metric for predicting speech intelligibility was proposed (Jørgensen and Dau, 2011), which is based on a measure of the SNR in the envelope domain (SNR_{env}). The SNR_{env} measures the relative strength of the speech and the noise envelope fluctuations, inspired by earlier work indicating that such a metric might be related to speech intelligibility (Dubbelboer and Houtgast, 2008). The metric is computed using the speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011; Jørgensen et al., 2013), which effectively mimics key aspects of human auditory signal processing. The key difference between the ESII and the sEPSM is that the sEPSM analyses the temporal modulation characteristics of the noisy speech and those of the background noise, using the concept of a modulation filterbank. The model was demonstrated to account for conditions with various stationary and fluctuating unintelligible interferers as well as the effects of reverberation and nonlinear noise reduction where the ESII failed (Jørgensen et al., 2013). The sEPSM framework might therefore be applicable for predicting the speech intelligibility performance of mobile phones.

Another promising model for predicting the effect of nonlinear processing on speech intelligibility is the short-term objective intelligibility model (STOI; Taal et al., 2011). This model does not assume modulation-frequency selectivity in the preprocessing, and instead of using an SNR-based decision metric, such as in the ESII and the sEPSM approaches, the STOI model assumes a decision metric based on the correlation coefficient between the temporal envelope of the clean speech and that of the processed/noisy speech. Taal et al. (2011) demonstrated that the STOI model could account for intelligibility of noisy speech processed by an ideal binary mask (Wang, 2005), and might be another candidate for predicting the effect of mobile phone processing.

Several aspects of the mobile phone signal transmission chain may influence the quality and intelligibility of transmitted speech. These include the surrounding acoustic environment, the microphone characteristics, the digital signal processing in the phone (typically including noise reduction and echo-cancelling) as well as the digital transmission network (Gierlich and Kettler, 2006). Figure A.1 illustrates the very basic elements of such a transmission

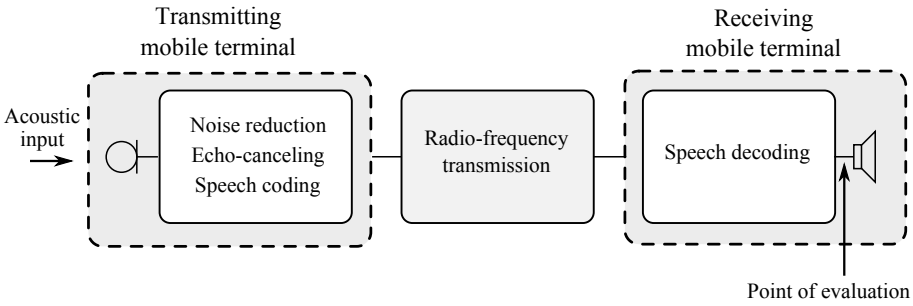


Figure A.1: Illustration of the basic elements of a transmission chain in a modern radio-frequency based telecommunication system, for example using mobile phones. The acoustic input is picked up by the microphone of the transmitting device, digitized, and typically processed by noise reduction and echo-cancelling algorithms. The signal is transformed via a speech coder and the resulting coefficients are transmitted using a radio-frequency based network. Finally, the signal is picked up by the receiving device, decoded, and played back via the devices' loudspeaker. The vertical arrow indicates the point of evaluation considered in the present study.

chain. Several studies have focused on the evaluation of different aspects of the transmission chain, such as speech coding/decoding algorithms (McLoughlin et al., 2002), echo-cancelling (Hänsler, 1994), noise reduction algorithms (Mattila, 2003; Westerlund et al., 2005), and effects of network transmission (ETSI TR 102 251, 2003). However, in order to include the combined effect of the various nonlinear steps of the transmission chain, an evaluation should consider the transmission chain as a whole, from the acoustic input through the transmitting phone to the signal picked up by the receiver (Gierlich and Kettler, 2006). In the present study, the transmitted signal was evaluated at the point just after it had been decoded by the receiving phone, as indicated by the vertical arrow in Fig. A.1. Thus, the loudspeaker characteristics or other components of the signal processing that would be specific to the receiving phone were not included. The aim was to obtain a realistic simulation of a telecommunication situation, by including the combined effects of the acoustic environment surrounding the transmitting mobile phone, the signal processing specific to the phone, and the digital transmission network.

The present investigation consisted of two parts. First, it was assessed whether the transmission of noisy speech through three different mobile phones

would lead to differences in intelligibility. Secondly, the main aim was to compare predictions using the sEPSM from Jørgensen et al. (2013), the STOI model from Taal et al. (2011), and the ESII from Rhebergen et al. (2006) to the measured data, in order to investigate to what extent these speech intelligibility models could predict the measured data and serve as objective evaluation methods of speech transmission through mobile phones.

A.2 Method

A.2.1 Stimuli

Sentences from the Danish Dantale II speech corpus (Wagener et al., 2003), spoken by a female talker, were used as the target speech stimuli. The speech corpus is a matrix test that consists of 160 five-word sentences with a grammatically correct structure (name + verb + numeral + adjective + object). All sentences are permutations of the 50 words of a base list with 10 sentences, thus making it very hard to memorize specific sentences, which allows reusing the sentences within a test session for the same listener (Wagener et al., 2003). The sentences were acoustically mixed with noise and recorded digitally using the setup shown in Fig. A.1. The setup consisted of a Brüel & Kjær 4128-D Head and Torso Simulator (HATS) geometrically centered between four loudspeakers in a standardized listening room (IEC 60268-13, 1985). The mobile phone under test was attached to the HATS using the Brüel & Kjær Handset Positioner Type 4606. Two-channel noise signals were supplied to the loudspeakers such that the two left loudspeakers played back the left channel and the right speakers played back the right channel of the signals. The four loudspeaker signals were de-correlated by introducing delays of 0 ms (front left), 11 ms (rear left), 17 ms (rear right), and 29 ms (front right) to the respective signals to make the resulting sound field more diffuse as recommended in ETSI EG 202 396-1 (2008). The mouth speaker of the HATS played back the target speech. The frequency response of the loudspeakers was equalized in a procedure similar to that described in ETSI EG 202 396-1 (2008). First, the transfer function from each loudspeaker to an omnidirectional 1/4" microphone (B&K 4938) at the position of the HATS was measured with pink noise. Then, each loudspeaker was equalized individually to have a flat frequency response between 120 Hz and 10 kHz (within ±3 dB) by filtering with the inverse of the measured frequency response. Once equalized,

all loudspeaker signals were given the same level, 6 dB below the desired playback level to yield the correct overall level when all loudspeakers were playing. Finally, the level of all loudspeakers playing simultaneously was adjusted to the desired overall playback level. The mouth speaker of the HATS was equalized to have a flat frequency response with respect to a 1/4" microphone at the mouth reference point of the HATS.

The mobile phone under test was connected to a Rohde and Schwarz CMD 55 Digital Radiocommunication Tester (DRT) via a locally established cellular network and the electrical output signal from the DRT was recorded using the Matlab software package via an RME Fireface UCX soundcard. Thus, the recorded stimuli reflected a one-way cellular telecommunication situation where the talking person was located in a noisy environment, such that both noise and target speech was transmitted via the mobile phone.

Individual recordings of all 160 sentences mixed with each noise type at a number of SNRs for each phone were stored digitally with a sampling rate of 44.1 kHz. In addition to the signal from the DRT, a reference signal from a Brüel & Kjær 4938 1/4-inch microphone positioned close to the input microphone of the mobile phone was recorded for all conditions. In this setup, and throughout this paper, the SNR was defined as the relative speech and noise levels measured with the reference microphone. The noise level was kept constant at 70 dB SPL throughout the recordings, the SNR was adjusted by changing the level of the speech.

A.2.2 Perceptual evaluation

Conditions

Three commercially available mobile phones from three different manufacturers were considered, denoted here as A, B, and C. The phones were released on the market in the years 2002, 2008, and 2010, respectively. All phones were state-of-the art models when released and phone B and C were so-called smart-phones. These phones were chosen because it was expected that the diverse release date could lead to different performance of speech intelligibility. There was no control over the signal processing in either of the phones.Three types of background noise were considered: Dantale II Speech shaped noise (SSN; Wagener et al.,

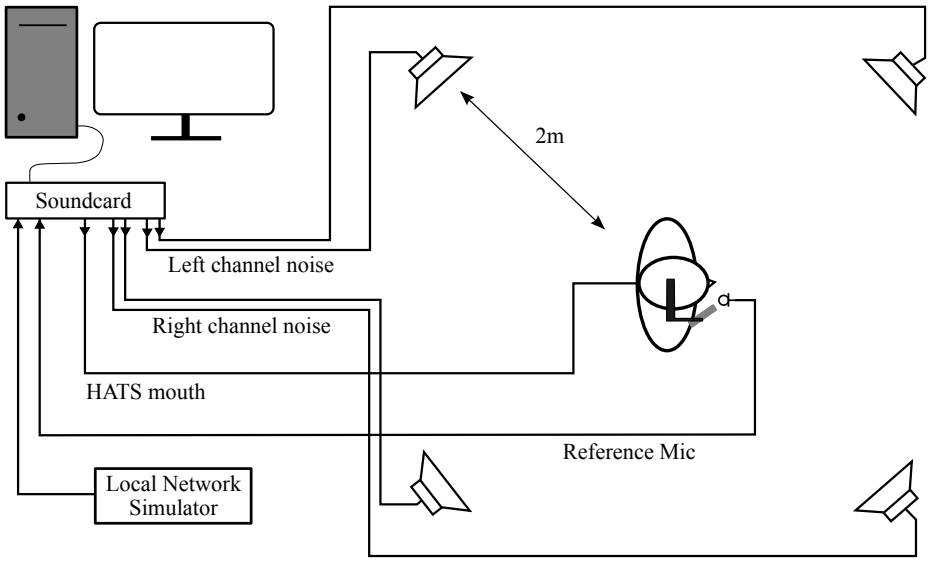


Figure A.1: Sketch of the setup used to simulate a one-way telecommunication situation in the present study. See main text for further details.

2003), "Traffic" (Outside-Traffic-Crossroads-binaural, street noise with cars passing by recorded from the pavement), and "Pub" (Pub-noise-binaural-V2k, babble noise with impulsive elements like clink of glasses) noise from the noise database provided in ETSI EG 202 396-1 (2008). The SSN and the Pub noise had long-term spectra similar to speech with most of the energy in the frequency range between about 120 Hz and 1 kHz and a roll-off of roughly 6 dB/octave for SSN, and 8 dB/octave for Pub above 1 kHz. The SSN was considered stationary, whereas the Pub noise fluctuated with the natural rhythm of the multiple talkers. The Traffic noise had pronounced low-frequency content with maximal energy around 50 Hz and a roughly constant decay of 12 dB/oct towards high frequencies. The Traffic noise was not completely stationary as it consisted of passing cars. In addition to the phone conditions, a reference condition with SSN background was included consisting of the broadband signal from the reference microphone, denoted as "Ref". Moreover, a condition with the broadband signal from the reference microphone filtered with the modified intermediate reference system (IRS; ITU-T P.830, 1996) transfer function (bandpass filter with −10 dB cut-off frequencies of 260 Hz and 3750 Hz), denoted as "BP" was included to evaluate the effect of the bandwidth of the transmission channel only. All conditions were evaluated at four SNRs, which were chosen to cover

the range from about 40% to 90% intelligibility in a given condition, based on pilot measurements. All conditions and SNRs were tested twice for each listener. In total, 2 x 44 (3 phones x 3 noises x 4 SNRs + (1 Ref + 1 BP) x 4 SNRs) conditions were evaluated per listener.

Apparatus and procedure

The perceptual evaluation was conducted in a double-walled sound insulated booth where the listener and the experimenter were seated. The sentences were presented to the listener's right ear via Sennheiser HD 650 headphones, which were equalized to have a flat frequency response at the ear reference point. The stimuli were filtered with the modified IRS receive transfer function to simulate a standard acoustic output of a receiving mobile phone. In any given trial, a noisy speech stimulus was presented to the listener, with the noise starting 1.6 seconds before the sentence and ending 1.2 seconds after it. The stimulus was faded in and out using 100 ms Hanning ramps. The gain of the playback system was adjusted independently for the Ref and BP conditions, such that sentences mixed with SSN at 0 dB SNR were presented at 70 dB SPL. Similarly, the gain was adjusted such that the sentences mixed with SSN at 0 dB SNR transmitted through phone A were presented at 70 dB SPL. This level was then kept fixed and used for all other mobile phone conditions. Therefore, the presentation level depended on the noise type, the phone type and the SNR condition. This led to level differences between the phones of up to 13 dB in some conditions. This difference was not expected to have an influence on speech intelligibility, because Wagener (2003) found no significant dependence on the presentation level for the intelligibility threshold when the Dantale II test was performed at presentation levels ranging from 45 to 80 dB SPL. The task of the listeners was to repeat as many words of the presented sentences as possible. The listeners were allowed to guess. The experimenter noted the number of correct words on a computer screen hidden from the listener.

The test was divided into three sessions of approximately two hours per listener. Two lists of 10 sentences were used for each phone, noise, and SNR configuration. Additional 80 sentences were used for training before the first and the second session, and 40 sentences before the third session, in order to familiarize the listeners with the test conditions. The noise and phone type conditions were balanced over the test subjects, the SNRs and the sentence lists

were randomly permuted with the restriction that no list could be reused within eight runs to avoid learning effects. Due to the large number of conditions, test lists were used repeatedly, though maximally 4 times per 2 hour test session.

Listeners

Six native Danish listeners with audiometric thresholds below 20 dB HL from 250 Hz to 8 kHz participated in the evaluation and were paid for their participation. None of the listeners had previous experience with psychoacoustic testing.

A.2.3 Modelling

The model predictions were based on a subset of 30 of the 160 sentences for each condition. The predictions for each sentence were performed separately in all conditions, and the results were averaged across the 30 sentences. The stimuli were truncated in time such that the onsets and endings for the speech and the noise were the same.

Separation of speech and noise components after transmission

The sEPSM and the ESII models considered here require separate access to the speech and the noise at the output of the transmission system. However, the separate transmission of the speech and the noise through the mobile phones would not reflect the situation considered in the perceptual evaluation, since the nonlinear behaviour of the transmission system could have affected the noisy speech mixture (used in the perceptual evaluation) in a different way than the speech and the noise alone. Therefore, estimates of the noise and the speech components were obtained from the noisy speech mixture after mobile phone transmission. This was achieved using a method developed by Hagerman and Olofsson (2004) for separating speech and noise from a mixture transmitted through a hearing aid with noise reduction processing. The method is briefly described in the following. Two noisy speech signals a_{in} and b_{in} were obtained from the speech and noise signals as:

$$a_{\text{in}}(t) = s(t)+n(t) \text{ and} \tag{A.1}$$

$$b_{\text{in}}(t) = s(t)-n(t), \tag{A.2}$$

where s and n denote the speech and the noise components at the input of the mobile phone, respectively and t represents time. Specifically, the phase of the digital noise signal was shifted by 180 degrees for $b_{\text{in}}(t)$ compared to $a_{\text{in}}(t)$. The separated speech and noise components, $s'_{\text{out}}(t)$ and $n'_{\text{out}}(t)$, were then obtained as:

$$s'_{\text{out}}(t)+\frac{1}{2}E_1(t) = \frac{1}{2}(a_{\text{out}}(t)+b_{\text{out}}(t)) \text{ and} \tag{A.3}$$

$$n'_{\text{out}}(t)+\frac{1}{2}E_2(t) = \frac{1}{2}(a_{\text{out}}(t)-b_{\text{out}}(t)), \tag{A.4}$$

where $a_{\text{out}}(t)$ and $b_{\text{out}}(t)$ denote the recorded signal mixtures at the end of the transmission line (point of evaluation, see Fig. A.1). The error terms, $E_1(t)$ and $E_2(t)$ can be estimated using the methods described by Hagerman and Olofsson (2004), but were neglected here for simplicity.

sEPSM-based simulations

The sEPSM framework assumes *a priori* information about the noise component of a noisy speech mixture. Thus, for the sEPSM-predictions, the inputs to the model for a given condition were the transmitted mixture, a_{out}, and the estimated noise component n'_{out}. From these inputs, the overall SNR_{env} was computed and converted to a sensitivity index, d', of a statistically "ideal observer" using the relation

$$d' = k(\text{SNR}_{\text{env}})^q \tag{A.5}$$

where k and q empirically determined constants. The value of q is set to 0.5 as suggested in Jørgensen et al. (2013) and k is assumed to depend on the speech material used. Finally, d' is converted to the probability of recognizing the speech for the ideal observer using an m-alternative forced choice model (Green and Birdsall, 1964) in combination with an unequal-variance Gaussian model. Conceptually, the ideal observer is assumed to compare the input speech item with m stored alternatives and to select the item (x_S) that yields the largest similarity. The $m-1$ remaining items are assumed to be noise, one of which, $x_{N,\text{max}}$, has the largest similarity with the input speech item. In this model, the value of x_S is a random variable with mean d' and variance σ_S^2. Similarly, the value of $x_{N,\text{max}}$ is a random variable with mean μ_N and variance σ_N^2. The

selected item is considered to be correct if the value of x_S is larger than $x_{N,max}$. The corresponding probability of being correct is estimated from the difference distribution of x_S and $x_{N,max}$

$$P(c) = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right) \tag{A.6}$$

where Φ denotes the cumulative normal distribution. The values of σ_N^2 and μ_N are determined by the response-set size, m, of the speech material (detailed expressions are given in Jørgensen and Dau (2011)). The value of m can be determined exactly if the material is based on a closed-set paradigm with a fixed number of response alternatives, such as the matrix-test paradigm used here, where m equals 50. The value of σ_S^2 was obtained empirically for Dantale II in Jørgensen et al. (2013). The value of k was determined by optimizing the fit of the sEPSM-predictions to the perceptual data in the Ref condition. The complete set of parameters were $[k = 0.66, q = 0.5, m = 50, \sigma_S^2 = 0.9]$. All model parameters were then kept fixed in all other conditions.

STOI-based simulations

The STOI-model assumes *a priori* information about the clean speech signal, so the inputs to the model were the transmitted mixture, a_{out}, and the clean speech $s(t)$. The model output was a scalar value, d, between 0 and 1, which is expected to have a monotonic relation with speech intelligibility. d was mapped to intelligibility scores using a logistic function suggested in Taal et al. (2011):

$$P_{correct} = \frac{100}{1 + \exp(a \cdot d + b)} \tag{A.7}$$

where a and b are free parameters. In this study, the values of a and b were determined as the minimum-mean-square error fit of the logistic function to the STOI-predictions vs. perceptual data in the Ref condition. The best fitting values were $a = -21.4785$ and $b = 12.8334$. These values were fixed for all other experimental conditions.

ESII-based simulations

The ESII requires information about the speech component and the noise component at the output of the transmission system under test. However, unlike the

STOI and the sEPSM, the ESII does not consider the speech signal itself, but uses a stationary speech-shaped noise signal as a probe for the speech component (Rhebergen and Versfeld, 2005; Rhebergen et al., 2006). Specifically, the inputs to the ESII in a given condition were the transmitted SSN-speech probe with a root-mean-squared (rms) level equal to the rms level of $s'_{\text{out}}(t)$ and the separated noise component $n'_{\text{out}}(t)$. The output of the ESII was an SII-value between 0 and 1, which was transformed to a percentage of correct responses using the transfer function suggested by Amlani et al. (2002):

$$P_{\text{correct}} = 1 - 10^{(-SII+K)/Q}. \tag{A.8}$$

K and Q were free parameters with values that were obtained using a minimum-mean-square fit of the transfer function to the ESII-predictions vs. perceptual data in the Ref condition. The optimal values for the parameters were found to be $K = 0.1851$ and $Q = 0.2234$, which were then used for all other conditions.

A.3 Results

A.3.1 Perceptual data

The percentage of correct responses for the individual listeners (averaged across the two repeated presentations) are shown in Fig. A.1, for the Ref condition (left panel) and the BP condition (right panel). A psychometric function with two parameters (Wagener et al., 2003), the slope (S_{50}) and the 50-% point (SRT_{50}), was fitted to the mean results of each individual listener in a given condition:

$$P(SNR) = \left[1 + e^{-4S_{50}\cdot(SNR-SRT_{50})}\right]^{-1}. \tag{A.9}$$

The parameters were averaged across the listeners to obtain a group-mean psychometric function. The obtained functions for the Ref and the BP conditions are shown as solid black lines in Fig. A.1. Similarly, Fig. A.2 shows the percentage of correct responses for the individual listeners and average psychometric functions for the conditions with SSN (left column), Pub noise (middle column), and Traffic noise (right column), and for the conditions with Phone A (top row), Phone B (middle row), and Phone C (bottom row).

It was not possible to obtain a psychometric function with a range from 0 to 100% for listener UCR in the conditions with Phone C in Pub-noise. In order to

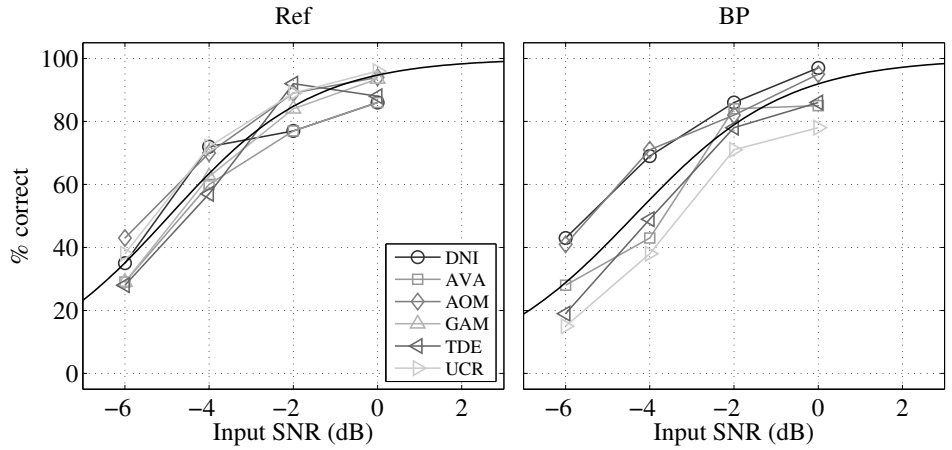


Figure A.1: Percentage of correct responses for the individual listeners for the Ref (left panel) and the BP conditions with SSN (right panel), and the corresponding group-mean psychometric functions (solid black line). No data was collected for subject GAM in the BP conditions.

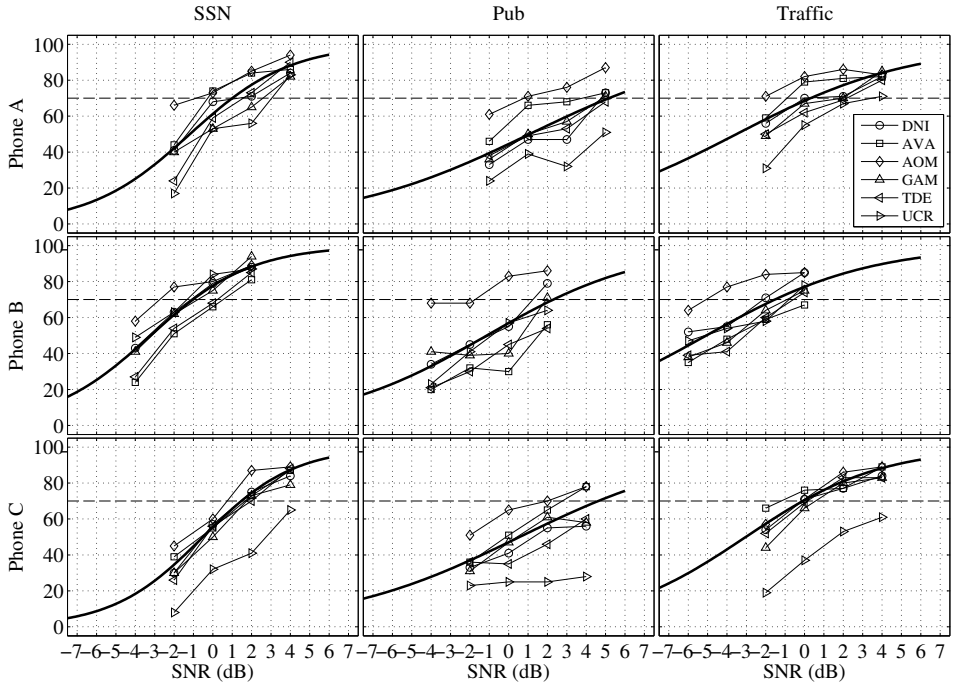


Figure A.2: Percentage of correct responses for the six individual listeners for the conditions with SSN (left column), Pub noise (middle column), and Traffic noise (right column), and for the conditions with Phone A (top row), Phone B (middle row), and Phone C (bottom row). The corresponding group-mean psychometric functions are indicated by solid black lines.

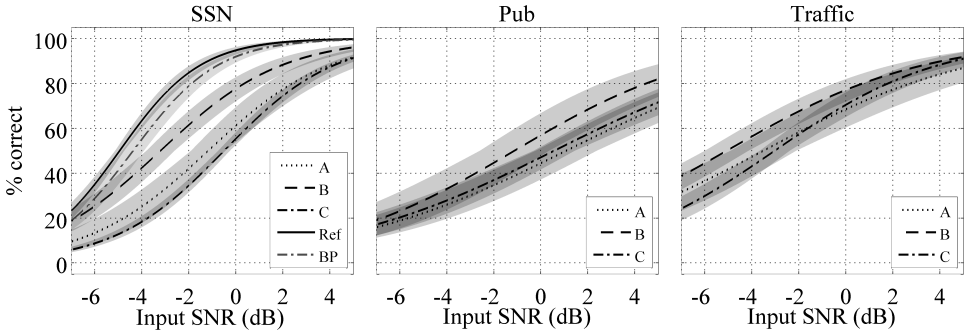


Figure A.3: Psychometric functions for all the considered conditions with SSN (left panel), with Pub noise (middle panel), and with Traffic noise (right panel). The shaded area surrounding each function represents one standard error of the parameters of the psychometric function.

have the same number of listeners for all conditions, this listener was excluded from the following statistical tests. A statistical analysis based on the remaining five subjects may seem weak, but it was sufficient to test the main hypothesis of the paper.

To ease comparison of the data across the phones, Fig. A.3 shows the psychometric functions for all the considered conditions with SSN (left panel), with Pub noise (middle panel), and with Traffic noise (right panel). The shaded area surrounding each function represents one standard error of the function parameters. The psychometric functions for the Ref and BP conditions were clearly above those for the mobile phones for input SNRs at and above −4 dB. This demonstrates worse intelligibility for the phone conditions compared to the Ref condition. Moreover, the slopes of the functions in the Ref and BP conditions were steeper than those of the phone conditions. For the SSN conditions (left panel), the psychometric functions for the three phones differed in their horizontal position. In contrast, the functions were much closer to each other in the conditions with Pub (middle panel) and Traffic (right panel) noise. This suggests that it is easier to discriminate between the performance of the three phones in the SSN conditions compared to the Pub and Traffic-noise conditions.

In order to quantify the differences between the perceptual results, the following analysis was based on the speech reception thresholds determined from the psychometric functions as the SNR corresponding to 70% correct (SRT_{70}). The SRT_{70} was chosen because this point was roughly in the middle of

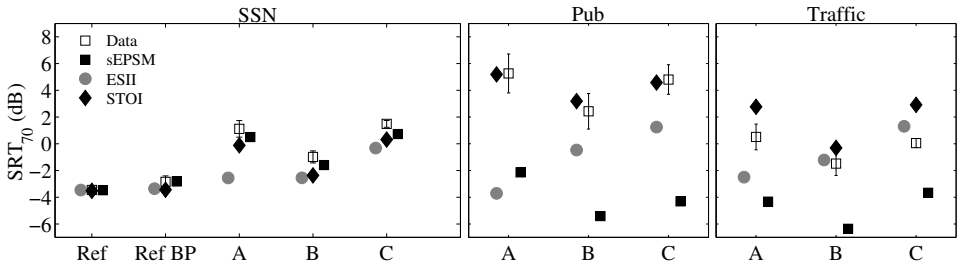


Figure A.4: SRT_{70} obtained from the perceptual psychometric functions (open squares), for all the conditions with SSN (left panel), with Pub noise (middle panel), and with Traffic noise (right panel). The vertical bars denote one standard error. Predictions from the sEPSM are indicated by the filled black squares, predictions from STOI are shown as filled black diamonds, and predictions from the ESII are indicated by the filled gray circles.

the measured range of percent correct. Figure A.4 shows SRT_{70} (open squares) obtained from the perceptual data for the conditions with SSN (left panel), Pub noise (middle panel), and Traffic noise (right panel). The vertical bars denote one standard error.

For the SSN conditions, the SRT_{70} for the Ref condition was obtained at an SNR of -3.5 dB, followed by a slightly higher SRT_{70} for the BP condition at -3.0 dB. The SRT_{70} for the three phones were obtained at higher SNRs than the reference conditions, with the lowest SRT_{70} (best intelligibility) obtained for phone B at an SNR of -0.9 dB, followed by phones A and C, which were both obtained at an SNR of 1.2 dB. For the conditions with Pub noise, the SRT_{70}s were obtained at SNRs of 4.2 dB, 2.4 dB, and 4.8 dB, for phones A, B, and C respectively. For the conditions with Traffic noise, the SRT_{70}s were obtained at SNRs of 0.5 dB, -1.6, and 0.1 dB for phones A, B, and C respectively. Thus, similar patterns of results were found for the Pub noise with generally higher SRT_{70}, and for the Traffic noise with generally lower SRT_{70}. For any given phone, the lowest SRT_{70} was obtained for the Traffic noise followed by the SSN and the Pub noise.

A one-way analysis of variance performed on the data for the three phones revealed that the SRT_{70} was significantly different across the phones for the SSN conditions ($p < 0.05$), but not for the Pub and the Traffic noises. A multiple comparison analysis with Bonferroni correction performed on all SSN conditions showed that the SRT_{70} for the Ref condition was significantly lower than the SRT_{70}s for all three phones ($p < 0.05$). Thus, the transmission chain led to

significant decreases of intelligibility for all phones, with respect to the reference condition. Moreover, the SRT_{70} for phone B (-0.9 dB) was significantly lower than for phone C (1.2 dB). Finally, the SRT_{70}s for the Ref and the BP conditions were not significantly different[a], demonstrating that band limiting imposed by the IRS-filter had very little influence on the obtained intelligibility.

A two-way analysis of variance with phone-type and noise-type as factors and individual listener SRT_{70}s as observations revealed a significant effect of phone type [$F_{2,36} = 5.1$, $p = 0.0112$] and noise type [$F_{2,36} = 20.38$, $p < 0.0001$], which means that the listener group average cannot be considered equal across phones, nor across noises for a given phone. There was no significant interaction, i.e., the relative difference between the SRTs for the phones was similar for the three noise types.

A.3.2 Model predictions

Psychometric functions (Eqn. A.9) were fitted to the predicted results and the corresponding predicted SRT_{70} were determined. Figure A.4 shows predicted SRT_{70} obtained with the sEPSM (black filled squares), the STOI (black filled diamonds), and the ESII (gray filled circles) for all considered conditions. The sEPSM accounted well for the data in the SSN conditions, with a root-mean-squared error (RMSE) of 0.58 dB. Moreover, the sEPSM predicted the same pattern of results across phone type as seen in the data for the Pub and Traffic noises, but with a large vertical offset of about -8 dB for the Pub noise and -5 dB for the Traffic noise, so that it does not predict the trend across noise type. The ESII predicted the right trend for phones B and C, however, it failed to predict the correct pattern of results across the three mobile phones seen in the data. The STOI-model accounted well for the relative performance across the phones for all noise types. However, the predictions showed a vertical offset of 2 dB in the conditions with Traffic noise. Table A.1 shows the across phone RMSE between the data and predictions for all models and noise conditions. Overall, the STOI-model provided the best prediction accuracy in the conditions considered here, predicting the trends across both noise and phone types.

[a] Subject GAM was excluded from this test because no data was collected for this subject in the BP conditions.

Table A.1: Root mean squared error (RMSE) in dB between measured and predicted SRT_{70} from the three mobile phones in the three noises considered in the present study.

	SSN	**Pub**	**Traffic**
sEPSM	0.58	8.2	4.5
ESII	2.2	5.8	1.9
STOI	1.0	0.46	2.2

A.4 Discussion

A.4.1 Simulation of a realistic one-way communication situation

One aim of the present study was to investigate to what extent speech intelligibility performance varied across three commercially available mobile phones in different noisy scenarios. The scenarios included different (noisy) acoustic environments where the mobile phones were transmitting speech via a locally established cellular network. A key element of the communication system simulation was that the gain of the playback system was fixed across the phones for the perceptual and objective evaluation procedures. This implied that the overall level of the presented stimuli varied across the phones, reflecting the effect of the noise reduction and the signal processing specific to a given phone. This level variation was considered an inherent part of the system, contributing to the realism of the simulation, and was therefore not compensated for in the playback system. Wagener (2003) measured the SRT using the Dantale II material and SSN at different overall presentation levels. They found no dependence of the measured SRTs on the overall presentation level, supporting that the differences in presentation level had negligible effect on the intelligibility scores measured in the present study.

The present study attempted to evaluate the mobile phone transmission system as a whole, from the acoustic input through the transmitting phone to the signal picked up by the receiver. One drawback of this setup was that it was difficult to disentangle which aspects of the transmission led to the differences in the performance across the phones. However, such a level of analytical detail was sacrificed in order to provide an overall impression of the performance of a given phone in a realistic communication situation.

A.4.2 Perceptual evaluation of speech intelligibility in modern telecommunication

The data from the perceptual evaluation showed a very similar pattern of results across phones for the three different noise types, i.e. the SRT_{70} was about 2 dB lower for Phone B compared to Phone A and C. Moreover, the conditions with SSN provided the lowest variability across listeners. The observation that the pattern of results for the different phones were similar for the different noise types suggests that it may be sufficient to use SSN for assessing the relative performance across the devices. This is in agreement with a recent study (Wong et al., 2012), which concluded that the SSN could be used to evaluate differences in intelligibility relevant to more realistic background noises.

The maximal difference in SRT_{70} obtained between the three phones with the SSN amounted to about 2 dB, which was reflected in a difference of 25% at an SNR of 0 dB in the psychometric functions for phone B and C. Such a difference might be crucial in an everyday communication situation, and motivates the use of intelligibility as a critical performance parameter of mobile phones.

A.4.3 Performance of the prediction models

Another aim of the present study was to investigate to what extent different objective speech intelligibility models could replace a panel of human listeners for evaluating the performance of a modern mobile phone.

The sEPSM framework accounted qualitatively for the differences in SRT_{70} across the three different mobile phones considered in this study, in all three considered background noise conditions, i.e. the predicted rank order across the phones was in agreement with the data. Moreover, the model accounted quantitatively for the data obtained with SSN showing an RMS error of only 0.51 dB. However, the predictions for the Pub and Traffic noises were offset vertically by -8 and -5 dB, respectively. For all mobile phones, the sEPSM predicted a better intelligibility for the Pub noise compared to SSN, which is in contrast to the data. Thus, the predicted rank order across the three noises was *not* in agreement with the data, i.e., the sEPSM was successful only in the conditions with SSN.

The predictions from the ESII generally showed the correct tendency for

Phones B and C in all three noise types. However it did not predict the correct trend for Phone A, which meant that the model failed to accurately account for the different SRT_{70} across the mobile phones in general. It was not possible to determine what characteristic of the stimuli from Phone A led to the poor performance of the ESII with this phone. Since the sEPSM performed reasonably well for Phone A, at least in the SSN, it could not be attributed to a general problem with the stimuli for Phone A.

Predictions from the STOI model were in good agreement with the perceptual data for all the considered conditions. In contrast to the sEPSM and ESII, this model did not require the use of the method from Hagerman and Olofsson (2004) to separate the speech and noise signals after the transmission through the phones. The success of the STOI indicates that the use of the separation method might have had a negative effect on the predictive performance of the sEPSM and ESII models. On the other hand, the performance of the sEPSM (using the Hagerman and Olofsson approach) was superior to the STOI in the SSN conditions. This suggests that the implementation of the separation method was accurate, but that the separation method might have been less appropriate for non-stationary noises such as the Pub and Traffic noises compared to the SSN.

The STOI and the sEPSM provided very similar predictions in the conditions with SSN, and the two models do have common aspects. For example, STOI effectively measures the temporal correlation of the modulation content from the envelope waveforms of the clean and transmitted speech signals, whereby any reduction of the correlation may be assumed to result from noise modulations or other non-speech modulations. This is conceptually similar to the SNR_{env} metric, which effectively measures the ratio of speech and noise modulation power. A main difference is that the sEPSM framework makes specific assumptions about the source of the speech degradation, i.e., modulation masking by noise, while this is not clear in the correlation coefficient. Moreover, the sEPSM includes additional aspects of human auditory processing, in the form of the perceptually and physiologically motivated modulation filterbank, which might be crucial in other conditions, such as reverberation, where the STOI metric has limitations (Taal et al., 2011). On the other hand, the STOI did not have to use any separation method, which greatly simplified the practical aspect of

predicting the intelligibility of the phones.

Another model that might be considered for the purpose of evaluating intelligibility of mobile phones is the telecommunication-version of the speech transmission index (STITEL; IEC 60268-16:2003, 2003). However, this model uses a probe signal consisting of modulated noise bands, and not the actual speech stimuli as used for the perceptual experiments. The nonlinear operation of the mobile phones and the transmission chain as a whole does not necessarily affect the probe and the actual speech in the same way and the STITEL might not be ideal in this respect. Alternatively, a speech-based version of the STI might be used, such as that suggested by Houtgast and Steeneken (1985). However, simulations using the speech-based STI failed to account for the perceptual data (not shown here explicitly), because the noisy speech stimuli became more modulated after transmission through the phones, compared to before the transmission (the Ref-condition), thus predicting improved intelligibility with the STI, so that this model would not be appropriate for predicting intelligibility of the mobile phones considered in this study.

A.5 Summary and conclusions

Speech intelligibility performance was assessed for three commercially available mobile phones in three different noise scenarios. The results showed statistically significant differences in speech intelligibility across the mobile phones of up to 2 dB of SRT_{70}. Transmission of the speech via the considered mobile phones led to a decreased intelligibility, relative to the reference condition without transmission through the mobile phones. A generally good correspondence between the predictions from one of three objective speech intelligibility models, the STOI, and the measured perceptual intelligibility was obtained in the considered conditions. The sEPSM provided accurate predictions only in the conditions with SSN. This suggests that the these models might be useful for objective speech intelligibility prediction in the development and evaluation of mobile telecommunication systems.

A.6 Acknowledgements

The authors thank Wookeun Song, Lars Birger Nielsen, and Claus Blaabjerg from Brüel & Kjær for providing measurement equipment and support for recording of the stimuli. This work was supported financially by the Danish Sound Technology Network, and the Danish hearing aid companies, Widex, Oticon and GN Resound.

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: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.

Vol. 22: *Federica Bianchi*, Complex-tone pitch representations in the human auditory system, 2016.

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.

Vol. 25: *Gusztáv Lőcsei*, Lateralized speech perception with normal and impaired hearing, 2016.

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation with cochlear implants, 2016.

Vol. 27: *Henrik Gert Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017

Vol. 28: *Richard McWalther*, Perceptual and Neural Response to Sound Textures, 2017

The end.

To be continued...

Spatial hearing, i.e. the ability to localize sound sources in space, is one of the most remarkable capabilities of the human auditory system. It allows us to be aware of events happening in our environment and to react to them even though they might be out of sight. Spatial hearing not only helps us to navigate in complex environments (such as in busy traffic situations), but also facilitates speech communication in situations with more than one talker. This thesis particularly focuses on auditory distance perception and externalization (the perception of sounds outside the head) as well as on speech intelligibility in complex environments with multiple sound sources. It was found that both distance perception and externalization were influenced by the room in which the experiment was conducted. It was also found that hearing aids change the spatial perception of the experimental setup and that speech intelligibility in normal-hearing listeners is decreased with hearing aids compared to the conditions without hearing aids. These findings might have implications for sound reproduction techniques and novel spatial sound stimulation paradigms. Furthermore, this work might be valuable for the design of future hearing-aid signal processing strategies.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk