Downloaded from orbit.dtu.dk on: May 02, 2024

DTU Library

Assessing hearing device benefit using virtual sound environments

Mansour, Naim

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Mansour, N. (2021). *Assessing hearing device benefit using virtual sound environments*. DTU Health Technology. Contributions to Hearing Research Vol. 46

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

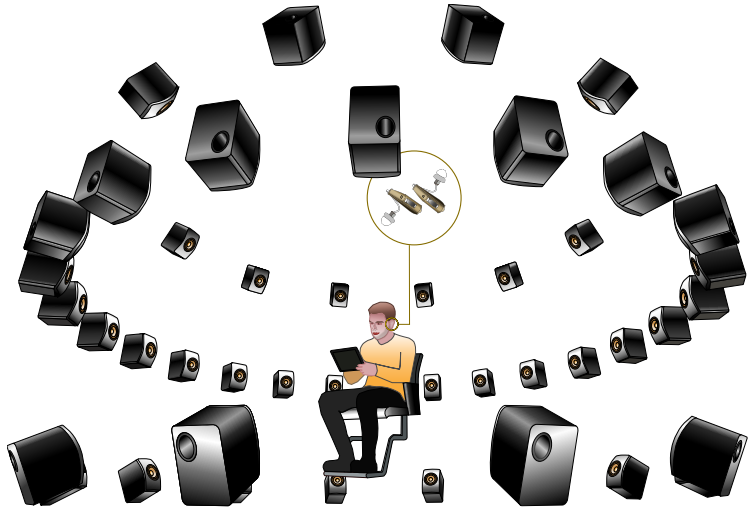
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 46

Naim Mansour

Assessing hearing device benefit using virtual sound environments



DTU Health Tech
Department of Health Technology

Assessing hearing device benefit using virtual sound environments

PhD thesis by
Naim Mansour

Technical University of Denmark

2021

© Naim Mansour, 2021

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

The project was partly financed by the Technical University of Denmark (2/3) and by Widex A/S (1/3).

Supervisors

Professor Torsten Dau
Associate Professor Tobias May
Doctor Marton Marschall
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kongens Lyngby, Denmark

Doctor Adam Westermann
Widex A/S
Lynge, Denmark

For MMM

Abstract

The investigation into people's ability to understand speech in noisy everyday situations, particularly those affected by hearing loss, constitutes an important area of hearing research. Hearing devices, such as hearing aids, attempt to restore a hearing-impaired person's real-world hearing ability. However, many psychoacoustic tests currently in use to evaluate speech intelligibility and improve hearing aid performance do not take the acoustic properties of complex real-world sound scenes into account, instead relying on artificial target speech and background noise signals presented over headphones or small sets of loudspeakers. While such laboratory settings provide highly controlled and reliable results, they typically do not capture acoustic characteristics of real-world environments such as reverberation and moving sound sources, and they do not fully reflect how people experience their real-world auditory reality.

This thesis aimed to bridge the gap between laboratory-based hearing tests and real-world listening by evaluating hearing using loudspeaker-based virtual sound environments (VSEs). Such a VSE is reproduced inside a spherical array of loudspeakers, capable of presenting spatialized sound fields to a listener positioned in the center with a high level of physical accuracy. By employing VSEs in combination with spatially recorded real-world noise signals and spatialized target speech, acoustically realistic speech intelligibility tasks were designed and implemented. This included the development of a method for in-situ, realistic conversational signal-to-noise ratio estimation, intended to characterize a talker's real-world speech levels. Measured speech reception thresholds (SRTs) were shown to be elevated for realistic VSE conditions compared to more artificial headphone and spatialized artificial noise conditions for both normal-hearing and hearing-impaired listeners. However, the hearing-impaired listeners' SRTs increased more between the artificial conditions and the realistic VSE condition than those of the normal-hearing listeners. Speech recognition scores obtained at the normal-hearing conversational signal-to-noise ratio provided percentage-correct scores relating speech intelligibility performance to communication ability in the real world.

Furthermore, it was shown that hearing aid dynamic range compression benefited speech intelligibility more in the realistic VSE condition compared to more artificial conditions, likely as a consequence of the acoustic properties of the speech and noise signals and their effect on the hearing aid signal processing. Finally, a method for guided ecological momentary assessment (EMA) was conceived to evaluate subjective, listener-reported hearing ability in a way that would reduce the data variability found in conventional EMA. The proposed method was shown to result in consistent ratings of hearing ability across a group of normal-hearing participants. The ratings were found to be reproducible inside acoustically matched, realistic VSEs.

Overall, this thesis showed the ability of VSE-based laboratory environments to provide increased acoustic realism in psychoacoustic listening tasks, rendering more ecologically valid results, for both normal-hearing and hearing-impaired individuals. The development of increasingly realistic VSE-based hearing and hearing aid evaluation tests has the potential to increase the benefit hearing devices provide to users in their everyday life.

Resumé

Udforskningen af menneskers evne til at forstå tale i støjende hverdagssituationer, især for dem, der er ramt af høretab, udgør et vigtigt område indenfor høreforskningen. Høreapparater forsøger at genoprette en hørehæmmet persons evne til at høre. Men mange psykoakustiske tests, der bruges til at evaluere taleforståelighed og forbedre høreapparatets ydeevne, tager ikke højde for de akustiske egenskaber i støjende hverdagssituationer, og bruger i stedet kunstige mål- og baggrundsstøjsignaler, der præsenteres over hovedtelefoner eller fra få højttalere. Selvom sådanne laboratorieopsætninger giver kontrollerede og pålidelige resultater, fanger de typisk ikke de akustiske egenskaber ved rigtige miljøer såsom efterklang og lydkilder i bevægelse, og afspejler ikke, hvordan folk oplever deres virkelige auditive virkelighed.

Denne afhandling har til formål at bygge bro mellem laboratoriebaserede høretests og hørelsen i den virkelige verden ved hjælp af højttalerbaserede virtuelle lydmiljøer (VSE'er). Et VSE gengives i en sfærisk højttaler opsætning, der er i stand til at gengive rumlige lydfelter for en lytter placeret i midten med et højt niveau af fysisk nøjagtighed. Ved at anvende VSE'er i kombination med optagede signaler fra den virkelige verden og en spatialiseret taletest blev akustisk realistiske taleopfattelsesopgaver designet og implementeret. Dette omfattede udvikling af en metode for in-situ, realistisk signal-støj-forholdsestimering. Målte tærskler for talemodtagelse (SRT'er) viste sig at være forhøjede for realistiske VSE-forhold sammenlignettil mere kunstige hovedtelefoner og spatialiserede kunstige støjforhold for både hørehæmmede og normalt hørende lyttere. De hørehæmmede lytteres SRT'er steg imidlertid mere mellem de kunstige forhold og den realistiske VSE-tilstand end de, der hører normalt. Tale forståelighedsscore målt ved realistiske signal-støj-forhold viste forholdet mellem taleforståelighed og kommunikationsevne i den virkelige verden.

Det blev vist, at kompression i høreapparatets dynamik gavner taleforståelighed mere i den realistiske VSE-miljøer sammenlignet med kunstige lytteforhold, sandsynligvis som en konsekvens af tale- og støjsignalernes akustiske egenskaber og deres virkning på høreapparatets signalbehandling. Endelig blev en metode til guidet momentan vurdering (EMA) brugt til at evaluere subjektiv høreevne på en måde, der ville reducere variabiliteten, der findes i konventionel EMA. Den foreslåede metode resulterede i ensartede vurderinger af høreevnen på tværs af en gruppe deltagere med normal hørelse. EMA vurderingerne var reproducerbare i akustisk matchede, realistiske VSE'er.

Samlet set viste denne afhandling VSE-baserede laboratoriemiljøers evne til øge akustisk realisme i psykoakustiske lytteopgaver. Udviklingen af mere og mere realistiske VSE-baserede evalueringer kan potentialet øge effektiviteten af høreapparaters signalbehandling.

Acknowledgments

I want to thank my supervisors, Adam, Marton, Tobias and Torsten, for supporting me throughout this project and for the many fruitful discussions. Thanks also to Jörg, my supervisor during the external research stay, for his creativity and guidance in challenging times. A special thanks goes to all my colleagues and friends at Hearing Systems, which have meant a lot to me and made my working environment truly enjoyable.

Thanks to Widex for co-funding the project and providing me with valuable resources and expertise, in a very welcoming way. At DTU, much of my research took place inside the AudioVisual Immersion Lab, and I want to thank Axel for teaching me how to operate it.

Last but not least, thanks to my family for always backing me, and to Maria for taking this journey with me, through all of its ups and downs.

Related publications

Journal papers

- Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (**2021**). "A method for realistic, conversational signal-to-noise ratio estimation", J. Acoust. Soc. Am. **149(3)**, 1559-1566 10.1121/10.0003626

- Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (**submitted**). "Speech intelligibility in a realistic virtual sound environment".

- Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**submitted**). "The effect of hearing aid dynamic range compression on speech intelligibility in a realistic virtual sound environment".

- Mansour, N., Westermann, A., Marschall, M., May, T., Dau, T., and Buchholz, J. (**submitted**). "Guided ecological momentary assessment in real and virtual sound environments".

Published abstracts

- Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**2019**). "Speech and background levels in a realistic sound environment", 11th Speech in Noise Workshop, Ghent, Belgium, January 2019.

- Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**2019**). "A method for conversational signal-to-noise ratio estimation in real-world sound scenarios", J. Acoust. Soc. Am. 145, 1873, Louisville, KY, May 2019. 10.1121/1.5101769

- Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**2019**). "The effect of hearing aid signal processing on speech intelligibility in a realistic virtual sound environment", J. Acoust. Soc. Am. 148, 2721, Acoustics Virtually Everywhere, December 2020. 10.1121/1.5147549

Conference papers

- Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**2019**). "Speech Intelligibility in a Realistic Virtual Sound Environment", Proceedings of the 23rd International Congress on Acoustics, Aachen, Germany, September 2019. 10.18154/RWTH-CONV-239341
  EAA Best Paper and Presentation Award for young researcher

Contents

List of figures

List of acronyms

4FAHL:	Four-Frequency Average Hearing Loss
ANOVA:	ANalysis-Of-VAriance
AVIL:	AudioVisual Immersion Lab
CM:	Cheek Microphone
DRC:	Dynamic Range Compression
DRR:	Direct-to-Reverberant Ratio
EMA:	Ecological Momentary Assessment
FFC:	Free-Field Correction
HA:	Hearing Aid
HATS:	Head-And-Torso Simulator
HI:	Hearing Impaired
HINT:	Hearing In Noise Test
HOA:	Higher-Order Ambisonics
LTAS:	Long-Term Average Spectrum
NH:	Normal Hearing
RIR:	Room Impulse Response
RMS:	Root-Mean-Square
SI:	Speech Intelligibility
SNR:	Signal-to-Noise Ratio
SPL:	Sound Pressure Level
SRS:	Speech Recognition Score
SRT:	Speech Reception Threshold
VAD:	Voice Activity Detector
VSE:	Virtual Sound Environment

1

General introduction

A healthy auditory system enables people to reliably and almost effortlessly navigate the world around them, to communicate with others and appreciate music. When hearing functions normally it is taken for granted, yet impairments to this ability can have severe consequences on the personal and professional lives of those affected (National Research Council, 2004). As reported by the World Health Organization, nearly 500 million people (over 6% of the world's population) are currently suffering from disabling hearing loss, an estimate which is expected to rise to over 900 million by the middle of the 21st century (Davis and Hoffman, 2019). Finding ways to help alleviate or overcome hearing impairments has therefore never been more important.

1.1 Hearing in the real world

The world is a complex place, also from an auditory point of view. In hearing research, the psychoacoustic properties of a real-world, multi-talker, reverberant sound scene are often studied and typically qualified as elements in a "cocktail party" environment (Cherry, 1953). By listening binaurally and using higher-level, cognitive segregation mechanisms, a well-functioning human auditory system has the ability to navigate such multi-talker scenes and selectively focus on the speech sounds of interest (Arons, 1992). But even normal-hearing people can struggle to understand speech in noisy environments. While increasing background noise levels generally cause an increase in conversational speech levels (known as the Lombard effect, Lombard, 1911), the signal-to-noise ratio (SNR) between speech and noise levels tends to decrease at a fixed talker distance (Weisser and Buchholz, 2019). When increasing conversational speech levels is no longer comfortable or socially acceptable, people will decrease their talking distance to continue being able to understand their conversational partner(s). The values and dynamics of real-world conversational SNRs between normal-hearing (NH) people in real-world scenes are indicative of the challenge

that a certain real-world sound scene poses to successful speech communication. Several studies have attemped to capture and characterize conversational SNRs (Pearsons et al., 1977; Smeds et al., 2015; Wu et al., 2018), yet it has remained unclear whether the used methodologies accurately represented the range of commonly experienced SNRs.

For people affected by a hearing loss, understanding speech in a cocktail party-like scene is substantially more challenging, as their loss of audibility combined with supra-threshold distortions reduces the overall sensitivity to sound and typically diminishes spatial, temporal and spectral resolution. Research on speech intelligibility (SI) has shown that because of these deficits, hearing-impaired (HI) individuals require higher conversational SNRs in order to properly understand speech in noisy backgrounds compared to NH listeners (Bradley et al., 1999). As such, any treatment or device that attempts to restore a HI individual's hearing should aim to re-establish the individual's SI at NH conversational SNRs.

1.2 The role of hearing devices

Over the course of the last 120 years, advances in science, engineering and technology have seen the birth and subsequent development of electro-acoustic devices that aim to restore, either partially or fully, normal hearing ability in people with a hearing loss (Alexander, 1998). Particularly during the last five decades, these hearing devices, or hearing aids (HAs), have seen a dramatic increase in sophistication, miniaturization and versatility (Mills, 2011), resulting in the tiny, digital machines manufactured today.

A typical modern HA works by recording sounds in the user's environment using microphones mounted on the HA shell, amplifying them and playing them back into the ear using a small loudspeaker, or HA receiver. The core sound amplification algorithm inside a HA operates in a non-linear, frequency- and level-dependent way, configured via prescription rules which aim to compensate for the HI person's hearing deficits as indicated by their pure-tone audiogram. The primary purposes of a prescription rule, such as the Desired Sensation Level 5 rule (DSL 5, Scollie et al., 2005) and the National Acoustics Laboratory-Non Linear 2 rule (NAL-NL2, Keidser et al., 2011), are to restore the HI user's ability to hear sounds, i.e. audibility, to that of a NH listener, while simultaneously improving their ability to understand speech, i.e. speech intelli-

gibility (SI). These prescription rules are commonly derived using models of SI, such as the Speech Intelligibility Index (SII , ANSI, 1997), and loudness (Moore and Glasberg, 1997; Moore and Glasberg, 2004), which are largely based on psychoacoustic listener data. Similarly, the potential benefit of a "fitted" HA is often assessed using psychoacoustic listening paradigms like SI tasks or speech-in-noise tasks in general. It is therefore important that psychoacoustic listening tasks, particularly those tasks that focus on aspects of speech-in-noise performance, produce controlled estimates of hearing ability that reflect real-world experience.

1.3 Experimental control versus realism

The empirical scientific investigation of any complex, real-world system is subject to a fundamental trade-off between experimental control and realism. A high degree of control over the experimental environment and stimuli permits the researcher to precisely relate the relative contributions of underlying phenomena to the value of an observed outcome measure. However, this type of investigative approach is sensitive to confounding effects because of its over-simplification of reality, and it is not always clear how (if at all) its conclusions can generalize to more realistic settings. Conversely, conducting experiments in realistic, yet uncontrolled, ways may limit the degree to which the contribution of underlying phenomena can be quantified reliably, because of the large variability in experimental conditions which decrease their reproducibility.

In the context of psychacoustic research, and particularly for speech-in-noise paradigms, this trade-off typically occurs along dimensions of the speech and noise stimuli as well as the presentation method. The most controlled way of presenting the stimuli in a speech-in-noise task is by using headphones. Using stationary noise as the masker, headphones can present speech stimuli ranging from individual words (Fogerty and Humes, 2010; Studebaker et al., 1999), to matrix-based sentences (Elberling et al., 1989), brief natural sentences (Nielsen and Dau, 2009) to potentially conversational speech. The same speech types can be combined with increasingly realistic noise stimuli, by combining modulated noise with matrix-based sentences (Hopkins and Moore, 2009) or brief sentences (Festen and Plomp, 1990), by combining noise obtained through simulations with words (Yang and Bradley, 2009) or brief sentences (Fogerty et al., 2020), by combining recorded noise with brief sentences

(Compton-Conley et al., 2004) or conversational speech (Sørensen et al., 2019) and possibly even by using environmental noise if open headphones were used. While headphone-based setups are highly controlled, they are generally limited in the accuracy with which they can spatialize stimuli and simulate the effects of head movement. In addition, wearing HAs inside headphones is practically challenging.

To address some of the limitations of headphones as a reproduction tool, loudspeakers have been commonly used to present the speech and noise stimuli instead. Studies have employed a similar range of speech of noise types, for example by combining stationary noise with word stimuli (Dirks and Wilson, 1969), stationary noise with brief sentences (Grange and Culling, 2016), recorded noise with word stimuli (Litovsky, 2005) or environmental noise with brief sentences (Hawley et al., 1999). Loudspeaker setups can reveal effects of spatial release from masking and binaural interaural time and level differences, yet they are generally not capable of accurately reproducing the physical sound field experienced by listeners and HAs in the real world. Furthermore, simulating realistic, moving interfering sound sources is difficult which means that reproduced scenes may lack real-world complexity and dynamics.

The most realistic way of reproducing speech-in-noise task stimuli is by conducting the task in the real world. Real-world speech-in-noise paradigms, although much less prevalent, have been developed, for instance by using environmental noise combined with word stimuli (Brungart et al., 2020) or conversational speech (Astolfi and Filippi, 2004). Even though environmental noise is always present in the real world, combinations can still be made by using loudspeaker setups placed in the real-world scene to simulate, for example, prerecorded interferers in the scene (Oreinos and Buchholz, 2016).

Figure 1.1 shows a three-dimensional depiction of the control-realism trade-off within speech-in-noise tasks. The two horizontal axes represent choices for the speech and noise stimuli while the vertical axis represents the ways of reproducing them. The elements on each axis are arranged ranging from highly controlled (red, at the origin of the coordinate system) to highly realistic (green, at the edges). The grid points indicate specific speech-noise-reproduction combinations. The brighter points represent specific combinations that are associated with a study mentioned above. The colored planes group a method of reproduction by a corresponding color. Even though the speech, noise and reproduction methods indicated represent just the most prevalent choices

for the three parameters, they render many potential ways for evaluating a person's speech-in-noise performance, each trading off control with realism in a slightly different manner. All of the studies cited above are positioned in Fig. 1.1 according to their control-realism trade-off. They represent only examples, and many more of the illustrated combinations can be found in literature.

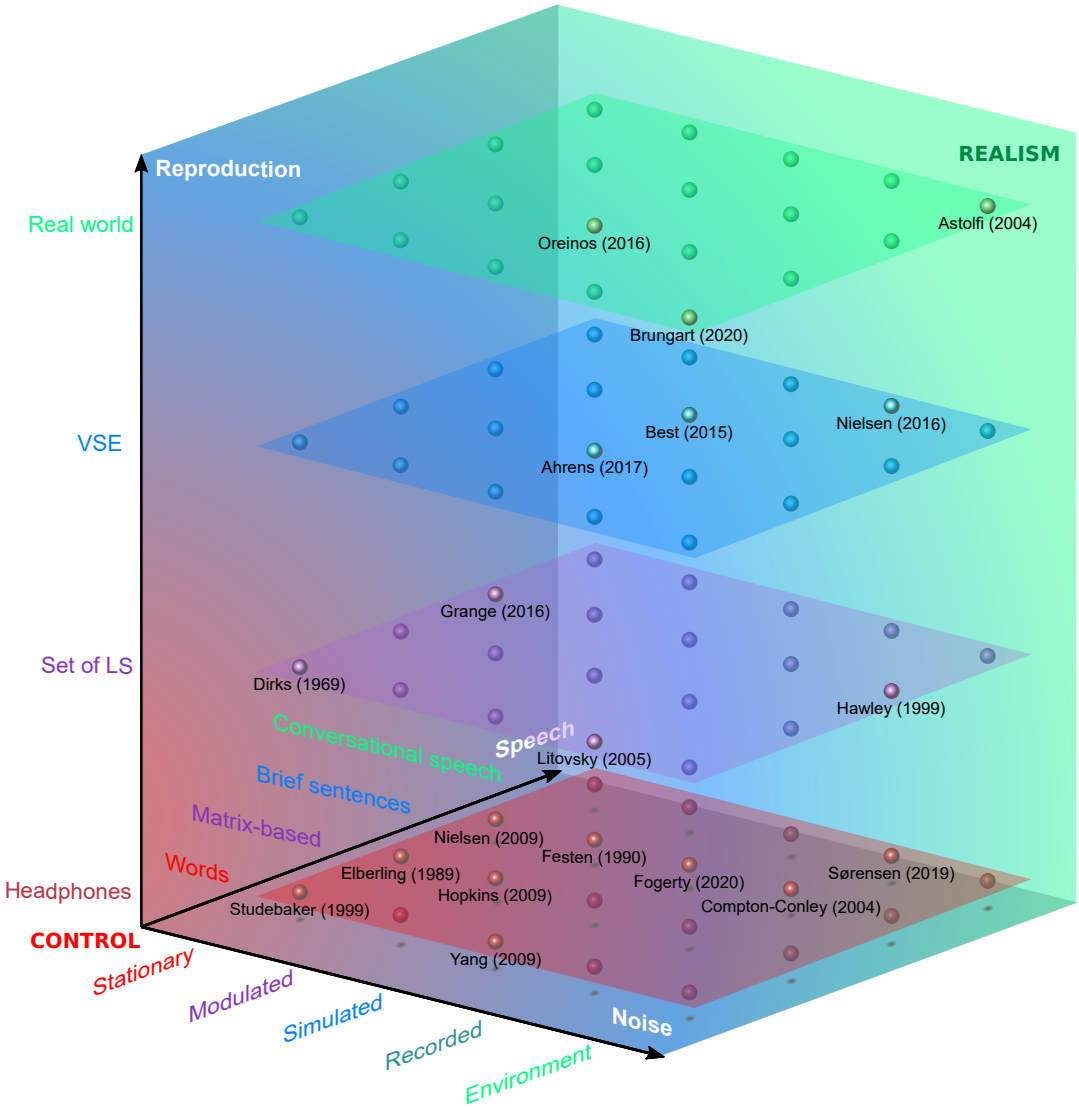


Figure 1.1: Graphical depiction of the trade-off between control and realism in speech-and-noise paradigms, with respect to the speech and noise stimuli types (horizontal axes) as well as their method of reproduction (vertical axis). The speech and noise types and the reproduction method range from highly controlled (red, origin) to highly realistic (green, edges). The grid points indicate specific speech-noise-reproduction combinations. The brighter points, associated with a reference, represent specific combinations that are referred to in the text. The colored planes group a method of reproduction of the corresponding color.

There are many possible ways of balancing experimental control with realism in speech-in-noise research, and any individual combination may depend on the specific hypotheses of the investigators. However, in recent years, a new way of reproducing stimuli in speech-in-noise tasks has gained traction in an attempt to achieve an optimal trade-off with respect to control and realism: the virtual sound environment (VSE).

1.4 The Virtual Sound Environment

A VSE can be defined as an array of spatially arranged loudspeakers, most commonly configured as a horizontal ring or a three-dimensional sphere, that work together to provide a desired spatial sound field to a listener positioned at the center of the array. These sound fields can be reproduced to be perceptually accurate to a human listener, using parametric techniques like directional audio coding (Pulkki, 2007), or to be physically (and by extension perceptually) accurate, using analytic techniques like wave field synthesis (WFS, Berkhout et al., 1993) and higher-order Ambisonics (HOA, Gerzon, 1973).

Of the techniques targeting physical accuracy, necessary for evaluating the non-human auditory processing of HAs, HOA-based methods are generally preferred over WFS due to their efficiency and robustness at higher frequencies (Daniel et al., 2003). The Ambisonic reproduction of a given sound field relies on its decomposition into "spherical harmonics" based on the expression of an acoustic pressure field as a Bessel-Fourier series. Similar to the Fourier transform for one-dimensional sound waves, the spherical harmonic functions form an orthonormal base which is modified by weighted spherical Bessel functions that define the sound field pressure in a spherical coordinate system. Depending on the order of the Ambisonic reproduction, the reproduced sound field is physically accurate, up to the associated spatial and temporal aliasing frequencies, within an area at the center of the array. This area is commonly referred to as the "sweet spot" and it increases with increasing Ambisonic order.

First developed several decades ago, Ambisonic reproduction techniques have recently gained popularity due to the successful practical expansion into higher orders, beyond the first-order systems originally developed, made possible by increasing computational power and the decreasing cost of microphone and loudspeaker array systems. The reproduction accuracy of current spherical loudspeaker arrays is mainly limited by the total number of loudspeakers M

required for a certain Ambisonic order N by the relation $N > (M+1)^2$ (Ward and Abhayapala, 2001). The weighted "B-format" functions necessary for Ambisonic reproduction can either be obtained using virtual source encoding, e.g. based on a room acoustic simulation, or by recording a desired sound field with a spherical microphone array. The latter approach is becoming increasingly widespread due to advances in microphone array technology, producing arrays that can record up to 4th, and recently even 7th order Ambisonics (Elko, 2018).

VSEs have been used in hearing research, together with a variety of reproduction techniques including HOA, and have been shown to be powerful tools for investigating spatial hearing and hearing aids in a controlled, yet more realistic way (Cubick and Dau, 2016; Minnaar et al., 2010). As illustrated in Fig. 1.1, studies have combined simulated HOA reproductions with matrix-based speech (Ahrens et al., 2017) and brief sentences (Best et al., 2015; Westermann and Buchholz, 2015), as well as spatially recorded HOA reproductions with conversational-style speech (Nielsen et al., 2016).

1.5 Designing VSE-based speech-in-noise tasks

To increase the ecologically validity of VSE-based speech-in-noise tasks, attention should be paid to stages in their design beyond solely the reproduction. This includes the way in which speech and noise stimuli are selected and acquired, as well as the specific task a listener is asked to perform.

1.5.1 Stimulus selection

Speech and noise stimuli are usually selected based on a restricted set of properties they possess, such as noise modulations or word meter, to investigate the impact of these properties on the considered outcome measures. When increased stimulus realism is targeted, by e.g. simulating or recording background noise, the ecological validity of the noise content is arguably important as well. Therefore, it may be a good idea to select background noise stimuli that represent situations that people actually experience in their life.

Recent studies have categorized real-world sound scenarios of HA users based on ecological momentary assessment (EMA) (Smeds et al., 2020; Wolters et al., 2016). The methodology of EMA consists of data collection through questionnaires that are presented to a participant (e.g. via a smartphone app) at

regular intervals in their everyday life. By asking participants to rate the relative importance, occurrence and difficulty of a scenario, the researchers were able to quantify which scenarios were of greatest interest to reproduce in the context of more ecologically valid speech-in-noise tasks.

1.5.2 Stimulus acquisition

A selected sound scenario can either be acquired through room acoustic simulations or recordings. Simulations, made with software packages like ODEON (ODEON A/S, 2020), have the advantage that they are fully controlled in terms of room acoustic properties and source/receiver positioning, and there are virtually no limits to the possible spatial geometry. However, the accuracy of the simulation is limited by the complexity of the room acoustic model and the approximations made by the software (e.g. ray-tracing order) and it is difficult to simulate complex, moving sources in a realistic way. On the other hand, spatial recordings made with spherical microphone arrays capture the scene "as-is", i.e. with much less flexibility with regard to room acoustic properties and positioning, but inherently tracking all present sources. Both approaches have preferred use cases, yet it has been shown that spatial recordings provide a better room acoustic approximation to a real-world reference scene than simulations (Ahrens et al., 2019).

1.5.3 Listening task

Designing a controlled yet realistic listening task that mimics real-world speech communication behavior is not straightforward, and most studies have instead focused on evaluating aspects of SI instead. Adaptive SI paradigms that produce speech reception thresholds (SRTs) as a measure of SI are widely used. Indeed, most of the references in Fig. 1.1 descibe SI experiments. The most realistic target speech stimuli used in these paradigms are brief natural sentences, even though efforts have been made to include aspects of conversational speech into the speech materials (Miles et al., 2020). However, it remains, as of yet, unclear if and how well the outcome measures of SI tasks can be related to people's subjective reports of real-world hearing ability. Newer methods like EMA may be able to bridge this gap between objective speech-in-noise assessments and subjective real-world reporting, if they can be modified for use in VSEs.

1.6 Overview of the thesis

This thesis investigates the impact of increased ecological validity on the results of VSE-based speech-in-noise paradigms as well as the potential impact and behavior of HA processing in such paradigms.

In *Chapter 2*, a method for in-situ, realistic SNR estimation is developed, attempting to characterize the speech levels of people in natural, real-world conversation in a more ecologically valid way. The method relies on a two-channel approach, using a lavalier microphone mounted to the cheek of a target talker and a second microphone placed next to the receiver. This approach is derived from a theoretical, room acoustic model and compared to an existing single-channel approach using simulations and real-world recordings.

Chapter 3 presents an approach for evaluating SI in a realistic VSE, focused on increasing the ecological validity of the stimuli selection, acquisition and reproduction. A real-world office meeting scenario is recorded with a spherical microphone array and reproduced inside a 64-channel loudspeaker array using HOA to provide the noise stimulus in a spatialized SI task. NH and HI listeners are evaluated and the results are compared with those obtained with tasks using more artificial stimuli and reproductions.

In *Chapter 4*, two real-world-recorded VSEs, as well as a reference condition employing artificial stimuli, are used to evaluate the effectiveness of a HA dynamic range compression processing strategy on the SI performance of HI listeners. The HA processing is simulated using a real-time "master" HA, and the results are compared to an unaided reference condition. An instrumental HA analysis is carried out to investigate the impact of the stimulus properties on the performance of the HA.

Chapter 5 introduces a guided approach to EMA that attempts to overcome some limitations to the applications of traditional EMA in hearing research and allow it to be applied inside realistic VSEs. In the method, a guide accompanies the participant to a known real-world location and assists the participant in carrying out various listening tasks, combined with EMAs. The procedure is then repeated inside two VSE-based laboratory environments to assess the consistency of the subjective assessments between the real world and the lab.

Finally, the concluding *Chapter 6* provides a general summary of the findings, including a discussion of the implications of the results. Several perspectives for future directions in hearing research employing realistic VSEs are given.

2

A method for realistic, conversational signal-to-noise ratio estimation[a]

Abstract

The analysis of real-world conversational signal-to-noise ratios (SNRs) can provide insight into people's communicative strategies and difficulties, and guide the development of hearing devices. However, measuring SNRs accurately is challenging in everyday recording conditions, where only a mixture of sound sources can be captured. This study introduces a method for accurate in-situ SNR estimation, where the speech signal of a target talker in natural conversation is captured by a cheek-mounted microphone, adjusted for free-field conditions, and convolved with a measured impulse response to estimate its power at the receiving talker. A microphone near the receiver provides the noise-only component through voice activity detection. The method is applied to in-situ recordings of conversations in two real-world sound scenarios. It is shown that broadband speech level and SNR distributions are estimated more accurately by the proposed method compared to a typical single-channel method, especially in challenging, low-SNR environments. The application of the proposed two-channel method may render more realistic estimates of conversational SNRs and provide valuable input to hearing instrument processing strategies whose operating points are determined by accurate SNR estimates.

[a] This chapter is based on Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (2021); A method for realistic, conversational signal-to-noise ratio estimation. The Journal of the Acoustical Society of America.

2.1 Introduction

Speech communication is a complex phenomenon that combines auditory, visual and cognitive processes to enable people to transmit and receive information. Such a conversation often occurs in noisy backgrounds, where a speech source of interest, i.e. the target talker signal, is accompanied by interfering sources (e.g. noise or competing talkers) and reverberation. Levels of conversational speech have been shown to strongly depend on the background noise level, as people raise their voice in increasingly loud surroundings to remain intelligible (Lombard, 1911). At the same time, the ratio of the average speech power arriving at the listener to the power of the background noise, i.e. the signal-to-noise ratio (SNR), is known to decrease at a fixed talker distance when the background noise level increases, i.e. people do not continue to increase their speech power indefinitely (Weisser and Buchholz, 2019).

Knowledge of the SNR distributions that occur in real-world conversations is important, since these SNRs affect a person's ability to understand speech in noisy environments. Developing more realistic listening tasks therefore demands accurate estimates of real-world speech levels and corresponding SNRs. Furthermore, the processing of hearing aids (HAs) strongly depends on the input signal levels. For example, the output SNR of a fast-acting dynamic range compression system depends on the input SNR, potentially impacting HA performance (Naylor and Johannesson, 2009). Accurate conversational SNR estimates would allow a HA to be tailored to the environment of its user (May et al., 2018).

Several studies have focused on the estimation of real-world SNRs. Specifically with regard to broadband, long-term estimates of conversational SNRs, two notable studies exist. In one of the studies, Pearsons et al. (1977) recorded conversations between two normal-hearing (NH) talkers at the ear of one of the participants in a diverse range of conditions, selected by the researchers. In the study by Smeds et al. (2015), HA recordings (Wagener et al., 2008) obtained by HA users in various situations of their daily lives were analyzed. Figure 2.1 shows the resulting broadband SNR distributions of the two studies (adapted from Wu et al., 2018). The blue and red bars represent the results from Pearsons et al. (1977) and Smeds et al. (2015), respectively. The purple shade indicates areas where the distributions overlap.

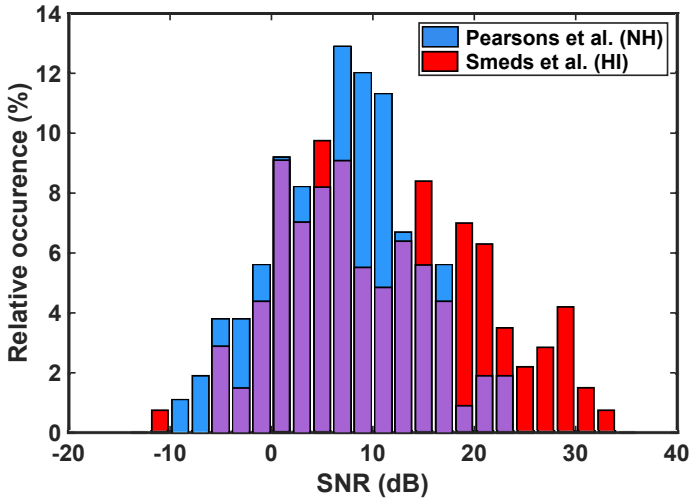


Figure 2.1: Distributions of speech-in-noise SNRs from Pearsons et al. (1977), indicated by the blue bars, and Smeds et al. (2015), indicated by the red bars. The purple shade indicates areas of overlap between the two distributions.

Both distributions reveal mostly positive SNRs across listening situations. The Pearsons et al. distribution is shifted slightly toward lower SNRs compared to the Smeds et al. distribution, most likely because Pearsons et al. collected data from NH participants who commonly communicate relatively easily at lower SNRs and may therefore not avoid such challenging acoustic conditions, unlike the HI participants (even if aided) in the Smeds et al. study.

While there were differences between the studies in terms of the methodology and hearing status of the participants, the SNRs were estimated in a similar way, using recordings made with a single microphone at the receiver position. Specificially, the root-mean-square (RMS) level of the clean speech was estimated by subtracting the average power of the noise-only segments from the average power of the noisy speech. These speech-in-noise and noise-only segments were hand-labeled by a human listener. The SNR was then obtained by dividing the estimated speech power by the noise-only power. This approach assumes that the speech and noise components in the recording are uncorrelated and that the estimated noise power in the noise-only segments reflects the noise power in the speech-and-noise segments. Both assumptions do not necessarily hold in real-world conditions with multiple interacting talkers in fluctuating background noise. Furthermore, it has been shown that at sufficiently negative SNRs, when the speech power becomes indistinguishable from the random fluc-

tuations in the noise power, this single-channel approach no longer provides accurate estimates since the SNR distribution essentially reflects the magnitude distribution of those fluctuations (Kim and Stern, 2008). In practice, the method relies on the accurate labeling of speech-in-noise and noise-only segments, which may become inaccurate at very low SNRs.

Here, a two-channel method is proposed to estimate real-world, in-situ conversational SNRs. The method extends the single-channel approach by introducing a cheek-mounted lavalier microphone to accurately capture the speech-only component of the target talker, in addition to the microphone at the receiver. A free-field correction and a room impulse response convolution were applied to this cheek microphone recording to obtain the target-speech-only signal at the receiving talker. From this signal, the SNR of the target talker at the receiver was derived by division with a noise-only signal, recorded at the ear of a mannequin standing next to the receiver. Accurate target speech labeling was employed based on the high-SNR cheek microphone signal, allowing for a reliable selection of segments where target speech was present, even in challenging situations containing speech-on-speech masking. The two-channel method was evaluated in room acoustic simulations of two real-world scenes, where theoretical, "true" SNR estimates could be calculated, and compared to the single-channel approach of Pearsons et al. (1977) and Smeds et al. (2015). In addition, both methods were evaluated for real-world recordings in the same two scenes.

2.2 Methods

2.2.1 SNR estimation principle

Figure 2.2 illustrates the conversational SNR estimation of a speech signal S produced by a target talker T at the location of a receiver R (yellow heads) in the presence of background noise N (blue rectangle). All signals are expressed in the frequency domain. S_R denotes the speech signal of the target talker at the position of the receiver.

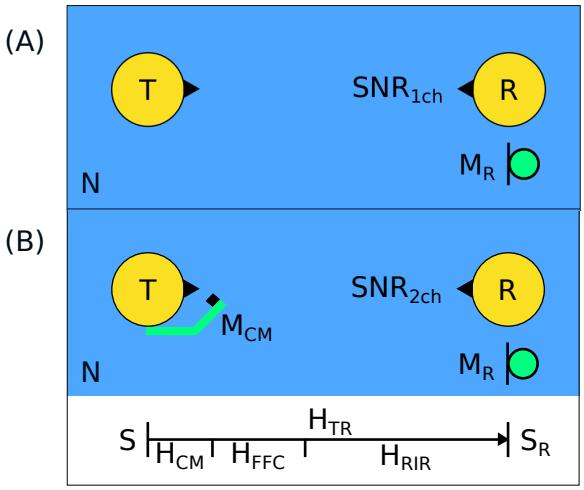


Figure 2.2: Conversational SNR estimation principle of a target talker T and their speech signal S at a receiver R (yellow heads) in a real-world containing background noise N (blue rectangle), for a single-channel method yielding SNR_{1ch} (panel A) and the proposed two-channel method yielding SNR_{2ch} (panel B). S_R denotes the speech signal of the target talker at the position of the receiver. M_{CM} and M_R represent a cheek microphone and receiver microphone (green stick and circle), respectively. H_{TR} denotes the transfer function between T and R, made up of the transfer function between T and M_{CM}, H_{CM}, a free-field correction transfer function H_{FFC} and a room impulse response transfer function, H_{RIR}.

The "true" SNR, SNR_{True}, is the ratio between the average power of S_R, $P(S_R)$, and the receiver noise-only power $P(N)$:

$$SNR_{True} = \frac{P(S_R)}{P(N)} \tag{2.1}$$

Neither $P(S_R)$ nor $P(N)$ can be measured in a real scene, since the target speech is mixed with the background noise by the time it arrives at the receiver. As illustrated in panel A of Fig. 2.2, a typical single-channel method uses a single receiver microphone M_R (green circle) to approximate $P(S_R)$ as $\tilde{P}(S_R)$, by capturing the noisy target speech power at the receiver $P([S+N]_R)$ and subtracting an estimate of the noise power $\tilde{P}(N)$ from it. $\tilde{P}(N)$ is obtained by estimating the noise power in speech gaps where the target talker and receiver are silent. Division of $\tilde{P}(S_R)$ by $\tilde{P}(N)$ then yields the single-channel SNR:

$$SNR_{1ch} = \frac{\tilde{P}(S_R)}{\tilde{P}(N)} = \frac{P([S+N]_R) - \tilde{P}(N)}{\tilde{P}(N)} \tag{2.2}$$

The proposed two-channel method, illustrated in the panel B of Fig. 2.2, estimates $P(S_R)$ directly by applying the room acoustic transfer function between T and R, H_{TR}, to S. To account for H_{TR}, a cheek(-mounted) microphone (green stick) worn by the target talker M_{CM} was used to capture the target speech (H_{CM}). Next, a fixed free-field correction (FFC) transfer function H_{FFC}, measured at a distance of 0.5 m, was applied to the recorded target speech to correct for near-field and head scattering effects due to the close distance of M_{CM} to the mouth of the target talker. Finally, convolution with an in-situ measured room impulse response (RIR), measured between T and R and calibrated to account for the attenuation caused by H_{FFC}, resulted in S_R (H_{RIR}). Division of the average power of S_R by $\tilde{P}(N)$, estimated in the same way as for the single-channel method, then yielded the 2-channel SNR:

$$\begin{aligned} SNR_{2ch} &= \frac{P(S_R)}{\tilde{P}(N)} = \frac{P(S \cdot H_{TR})}{\tilde{P}(N)} \\ &= \frac{P(S \cdot H_{CM} \cdot H_{FFC} \cdot H_{RIR})}{\tilde{P}(N)} \end{aligned} \tag{2.3}$$

Assuming that M_{CM} captures negligible background noise and that the speech power is the same at R and M_R, S_R can be obtained by the two-channel method. This is the main difference from the single-channel method and implies that the only deviations to SNR_{True} will be caused by the approximation $\tilde{P}(N) = P(N)$ if the assumptions for the speech signal, mentioned above, are fulfilled. This approximation for the noise power only holds if N is isotropic in space between R and M_R and stationary over time. In addition, the two-channel method allows for an accurate detection of the target talker speech segments even at low SNRs by using a voice activity detector (VAD) applied to the M_{CM} signal, which is not possible with the single-channel method.

In the following, each step in the proposed method is outlined in detail. All signals were sampled at a rate of 48 kHz and a resolution of 24 bit. Levels of speech and background noise as well as SNRs were derived from their broadband average power, in dB.

2.2.2 Microphone measurements and voice activity detection

The cheek microphone (DPA 4066, DPA Microphones, Lillerød, Denmark) used to capture the target speech signal S was mounted at a 5-cm distance next to the target talker's mouth, representing H_{CM}. It was assumed that, at this distance,

the power in the speech signal picked up by M_{CM} could be entirely attributed to S and that the dynamic range of the signal would be sufficient to accurately separate target speech segments. Energy-based VADs (Kinnunen and Li, 2010) were applied to both the M_{CM} and M_R signals. The obtained binary speech detection masks were used to exclude the speech of R and the noise N from the signal in M_{CM} and to exclude the speech of T and N from the signal in M_R. The VAD applied to M_{CM} estimated the short-term energy of S by segmenting the recording into frames of 20 ms duration and subsequently applying a threshold to this short-term energy, relative to its maximum value, to identify frames which contained relevant target activity. This threshold was set to the difference in dB between the 95th and 50th percentile of the short-term energy in order to adaptively separate the target speech energy distribution (peaking in the 95th percentile) from the background noise distribution (assumed to be distributed around the 50th percentile). Speech gaps longer than 200 ms (Demol et al., 2007) were not considered to be part of T, ensuring that the estimated speech power would not be affected by silence gaps.

The right-ear microphone of a Knowles Electronic Manikin for Acoustic Research (KEMAR, GRAS Sound & Vibration A/S, Holte, Denmark) mannequin with ear canals was used as M_R to estimate the noise-only signal N in a way that captures the effects of head and pinnae shape present in human listening. The receiver speech was subsequently removed using the same VAD applied directly to the M_R signal, but with a fixed threshold energy at 15 dB below the global maximum of the short-term energy, equal to the lower speech range boundary used in the computation of the speech transmission index (Houtgast et al., 1980). A fixed threshold was used in M_R, but not in M_{CM}. The target speech S contained in M_{CM} had a larger and more strongly varying dynamic range between frames than the receiver speech in M_R, due to the closer proximity of M_{CM} to T. This required an adaptive threshold to ensure the proper detection of the target speech. As was verified, applying a fixed threshold to the M_{CM} signal would have resulted in an underestimation of speech activity. The M_{CM} and M_R recordings were time-aligned to compensate for the acoustic delay through cross-correlation (Stoica, Moses, et al., 2005), allowing for the usage of both VAD masks in both microphone signals to remove R speech and T speech, respectively.

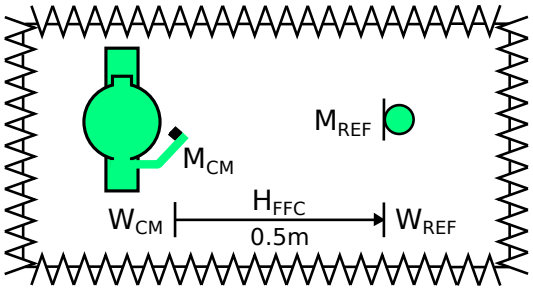


Figure 2.3: FFC measurement setup, including the KEMAR with mounted cheek microphone M_{CM} and reference microphone M_{REF} at a 0.5 m distance inside an anechoic enclosure. WN and WN_{REF} denote the white noise stimulus at the position of M_{CM} and M_{REF}, respectively. H_{FFC} denotes the transfer function between M_{CM} and M_{REF}.

2.2.3 Free-field correction

The near-field signal produced by the target talker's mouth was corrected for free-field conditions using the measurement setup illustrated in Fig. 2.3. The transfer function H_{FFC} between the position of M_{CM} mounted on the KEMAR (mannequin icon) and that of a reference pressure field microphone M_{REF} (GRAS AG40, GRAS Sound & Vibration A/S, Holte, Denmark), positioned upright at a distance of 0.5 m to the KEMAR, was measured inside an anechoic chamber. The KEMAR mouth simulator produced white noise, recorded by M_{CM} as W_{CM}, at a sound pressure level (SPL) of 90 dB at M_{REF}'s position, recorded as W_{REF}. A frequency-domain transfer function H_{FFC} was then derived from the ratio of the frequency-dependent cross-power spectral density of W_{CM} and W_{REF}, $P(W_{CM}, W_{REF})$ and the auto-power spectral density of W_{REF}, $P(W_{REF}, W_{REF})$:

$$H_{FFC} = \frac{P(W_{CM}, W_{REF})}{P(W_{REF}, W_{REF})} \tag{2.4}$$

H_{FFC} was smoothed in the frequency domain over critical bands using a 4th-order gammatone kernel G_s resembling the critical bands of the human auditory system, to avoid over-fitting H_{FFC} to the exact M_{CM} position and head shape used in the measurement. The original and smoothed magnitude responses of H_{FFC} are plotted between 100 Hz and 24 kHz, in Fig 2.4. Finally, a linear-phase finite-impulse response filter (FIR) was designed using the smoothed magnitude response, consisting of $n = 256$ taps and applying Hamming windowing to obtain $h_{FFC}[n]$ as time-domain representation of H_{FFC}:

$$h_{FFC}[n] = FIR\left[\sqrt{(G_S(|H_{FFC}|^2)}\right] \tag{2.5}$$

The target and realized filter magnitude responses were compared to evaluate that the chosen filter length was sufficient to correct for the main features of the transfer function. The M_{CM}-M_{REF} distance of 0.5 m was chosen to ensure a high dynamic range in WN_{REF} despite the power limitations of the mouth simulator. This resulted in highly coherent input signals to both microphones, as is necessary for reliably estimating H_{FFC}.

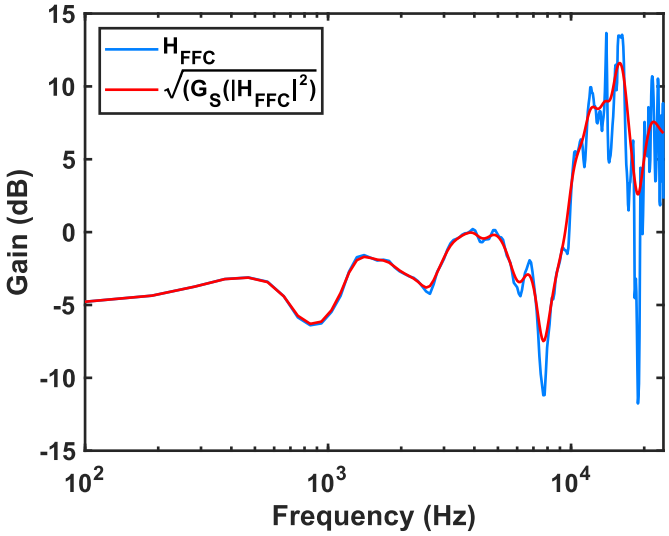


Figure 2.4: Magnitude response of the transfer function H_{FFC} and its smoothed version $\sqrt{G_S(|H_{FFC}|^2)}$, between 100 Hz and 24 kHz.

2.2.4 Real-world measurement setup and RIR measurement

The two-channel SNR measurement setup was realized in two real-world environments: an office meeting and a public lunch scenario. Panels A and B of Figure 2.5 show a top-down illustration of the measurement setup. In the office meeting, twelve normal-hearing participants were present in a typical office conference room of approximately 25 m^2, seated and standing around a large square table. The participants were coworkers who knew each other well. They were asked to converse naturally in pairs for a period of 5 minutes about everyday topics provided to them on a list, to generate the background noise (blue heads) while the male target T and receiving talker R (red heads) were having the conversation of interest at a distance of 2.4 m. Both the cheek microphone M_{CM} and the right ear of the KEMAR M_R were connected to a

sound card (Fireface 800, RME, Haimhausen, Germany) controlled by a laptop. The M_{CM} and M_R inputs were clock-synchronized to sample precision. The setup was similar in the lunch scenario, except that the twelve participants were now seated at narrower lunch tables in a large open-plan canteen of approximately 800 m^2, and the T-R distance was only 1 m. The single-channel SNR estimation method was applied in both scenes as well, using only the M_R recording. However, it used the VAD masks derived by the two-channel method to classify S_R and N segments in the M_{CM} and M_R signals, ensuring manual labeling errors would not affect classification performance.

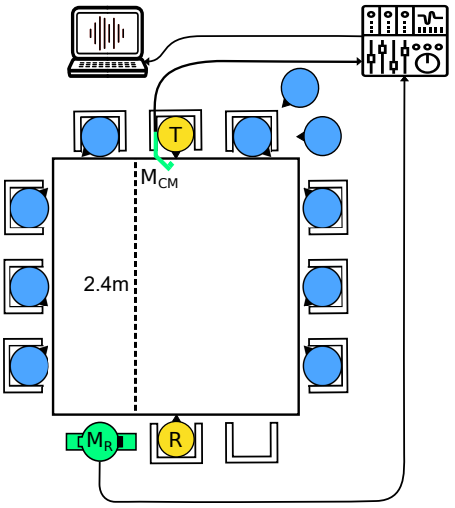
For both the single-channel and two-channel SNR analyses, the input recordings were divided into frames of 5 s with a 1-s shift between frames to obtain 294 SNR estimates within the 5-min-long recordings. These values were chosen to ensure a sufficient number of speech and noise samples within a frame and smooth transitions between frames, while maintaining the same average frame length that was used in the single-channel reference studies. Frames that contained only speech or only noise samples were excluded from the calculation. The speech and noise stimulus levels were calculated by computing digital RMS values and converted to SPLs.

Since the RIR transfer function H_{RIR} depends on the acoustic surroundings, it was measured in-situ in both sound environments. As illustrated in Fig. 2.5C, the RIR between T and R (red heads) was obtained by replacing the receiving talker with the KEMAR and recording 15-s-long exponential sinusoidal sweeps, from 20 Hz to 20 kHz, played by a two-way loudspeaker (KEF R3, KEF Audio, Maidstone, United Kingdom) placed in the target talker position (green rectangle). The sweep was played in a quiet background (interfering speakers and background were silent) at a level of 90 dB broadband SPL measured at R. Since the RIR was recorded between T and R, it had to be calibrated to account for the 0.5 m attenuation of S after convolution with H_{FFC}. During the calibration stage, the target talker was asked to speak at a conversational level to the receiver (in the same configuration as in Fig. 2.5C), in quiet. In the absence of noise ($N = 0$), the power of the recorded M_R signal, $P([S+N]_R)$ is equal to $P(S_R)$. A scaling factor α was applied to H_{RIR}, set such that the speech levels measured at the receiver ($P(S_R)$) and derived from the M_{CM} signal ($P(S \cdot H_{CM} \cdot H_{FFC} \cdot \alpha H_{RIR})$) were equal.

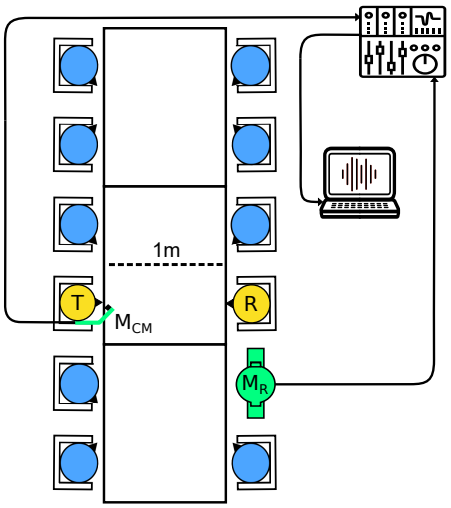
2.2.5 Simulated and real-world validation

To compare SNR_{2ch} with SNR_{1ch} and SNR_{True}, room acoustic simulations of the two real-world scenes were constructed (further denoted by the suffix "Sim" appended to a variable name). True SNR distributions around a desired median value were established by modeling the target speech with an anechoic source S, convolved with the H_{RIR} measured in the two real-world scenes to obtain S_R. This S_R signal was scaled and superimposed on an N signal, modeled by the noise-only M_R recordings made in the two real-world scenes, to obtain $[S+N]_R$. R and M_R were assumed to be in the same position. The target speech source consisted of 30 concatenated, anechoic sentences from the Danish Hearing in Noise Test (HINT) corpus. These male-spoken sentences were, on average, 1.5 s long and were separated by silence gaps set to 1 s, the average silence gap length in the real-world version of the target speech. A 5-second frame length and 1-second shift was used to process the signals. The two-channel method was simulated at a median SNR_{True} by using S and $[S+N]_R$ as inputs; the single-channel method only had access to $[S+N]_R$. The two-channel method's calibration procedure was simulated by setting the N signal in $[S+N]_R$ to 0.

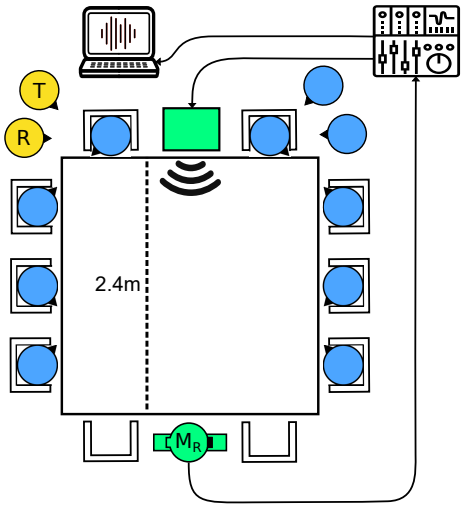
The simulations assumed S as recorded by M_{CM} to be anechoic (due to the use of the HINT corpus) and N to be isotropic (because of the assumption that M_R was in the same position as R). Since these assumptions may not entirely hold true in the real world, comparing simulation results to actual measurements is crucial. While SNR_{True}, by definition, could not be determined in the real-world scenes, differences between SNR_{2ch} and SNR_{1ch} were compared between the measurements and simulations. In addition, comparisons were made between the measured SNR_{2ch}, SNR_{1ch} and the simulated SNR_{2chSim}, SNR_{1chSim} and SNR_{True} by matching the measured SNR_{1ch} distributions to their simulated counterparts SNR_{1chSim} at their median.



(A) Office meeting recording



(B) Public lunch recording



(C) RIR measurement

Figure 2.5: Panel A: Conversation-in-noise recording setup in the office meeting scenario, including the target T and receiving talker R (red), M_{CM} and M_R (green) and other participants (blue). Panel B: Conversation-in-noise recording setup similar to (A), for the public lunch scenario. Panel C: Illustration of the RIR measurement setup in the office meeting scenario in the presence of all participants (blue and red), with the loudspeaker (top, green) and M_R (bottom, green) producing and capturing the excitation signal, respectively.

2.3 Results

The results described below reflect the outcome of the room acoustic simulations, evaluating the performance of the single-channel and two-channel estimation methods compared to the true SNR in the office meeting and the public lunch background noise. The in-situ measurement results relate the different methods to each other in a real-world application.

2.3.1 Room acoustic properties

Table 2.1 displays the main room acoustic parameters that characterize the office meeting and public lunch scenarios, based on the analysis (Hummersone, 2020) of the early decay characteristics of the measured RIRs: the reverberation time at 1 kHz (RT_{60}), the direct-to-reverberant ratio (DRR), the clarity (C_{50}) and early decay time at 1 kHz (EDT).

Table 2.1: Room acoustic parameters for the two real-world scenarios

	RT_{60} (s)	DRR (dB)	C_{50} (dB)	EDT (s)
Office meeting	0.4	6.6	16.9	0.2
Public lunch	3.5	16.6	23.5	0.4

The office meeting room had a dry response (low RT_{60}) of 0.4 s, with a considerable amount of early reflections (high EDT) and a relatively small direct sound contribution (low DRR) at the receiver position. In contrast, the large public lunch space contained considerable reverberation (high RT_{60}) and showed a relatively fast decay of early reflections and an increased DRR. These room acoustic parameters reflect the differences in the physical layout of the two scenarios. The office meeting space was a typical conference room with a carpeted floor, two glass walls and a suspended ceiling, all of which contribute to the low reverberation time. The public lunch took place in a large open-spaced canteen, with multiple highly reflective surfaces contributing to increased reverberation. The larger distance of 2.4 m between the target and receiver in the small office meeting room implied that multiple pronounced early reflections reached the receiver at different times after the direct sound, increasing the EDT and subsequently reducing the DRR and C_{50}. Conversely, the target-receiver distance of only 1 m in the public lunch space resulted in a much more prominent direct sound component, with sparse early reflections due to the size of the space, as evident through the low EDT and increased DRR and C_{50}.

2.3.2 Room acoustic SNR simulations

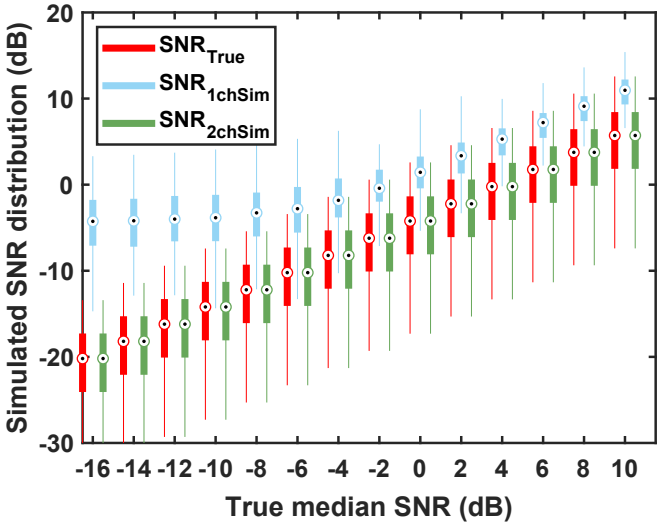
Figure 2.6A displays box plots of the true SNR distributions (SNR_{True}, red), simulated at specified median SNRs between -16 dB and 10 dB, in steps of 2 dB, as well as the corresponding SNR distributions obtained by simulating the single-channel (SNR_{1ch}, blue) and the two-channel (SNR_{2ch}, green) methods, for the office meeting scenario. Figure 2.6B shows the corresponding simulated distributions for the public lunch scenario. A one-way analysis-of-variance test showed a significant effect of the applied method in both scenes across all SNRs, with the single-channel method resulting in significantly increased SNRs compared to both the two-channel method and the true SNR ($p \leq .0001$ for all comparisons). The difference increased with decreasing SNRs, as the single-channel distributions flattened out around -10 dB SNR. The two-channel distributions were not significantly different from the true SNR distributions ($p = .77$ and $p = .87$ for the office and public lunch scenario, respectively) but slightly more spread out, especially for the public lunch scenario.

2.3.3 Real-world speech and background levels, SNR

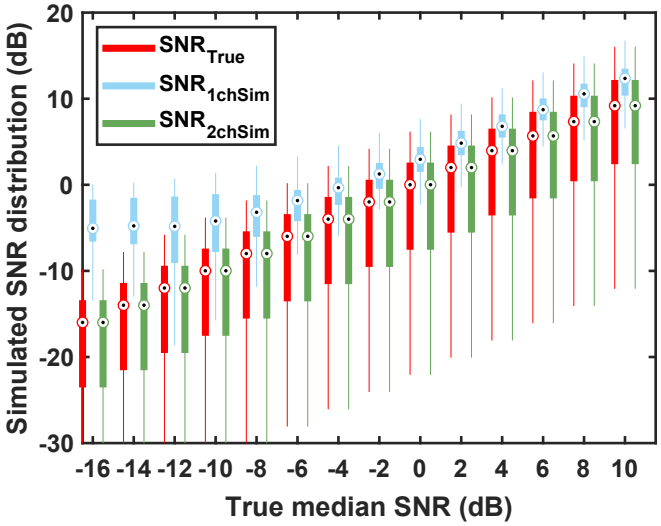
Figure 2.7A shows the S_R distributions obtained with the single-channel (S_R^{1ch}, blue) and the two-channel (S_R^{2ch}, green) methods as well as the common background noise level distribution (N, black) for the office meeting, using the left, dB SPL ordinate. The SNRs for the single-channel method (SNR_{1ch}, blue) and the two-channel method (SNR_{2ch}, green) are provided as well, alongside the simulated single-channel SNR distribution (SNR_{1chSim}, blue) matched at the median to SNR_{1ch} and the corresponding simulated two-channel distribution (SNR_{2chSim}, green), using the right, dB SNR ordinate. Finally, the corresponding simulated true SNR is shown (SNR_{True}, red). Figure 2.7B shows the corresponding results for the public lunch scenario. The left- and right-hand ordinates were aligned in both panels such that the median noise level in dB SPL corresponded to 0 dB SNR.

In the office meeting scenario, the median of S_R was 76.2 dB SPL for the single-channel method and 71.2 dB SPL for the two-channel method. The median of N was 73.5 dB SPL. The resulting median of S_R^{1ch} and S_R^{2ch} were -2.5 dB and 2.3 dB, respectively. SNR_{2chSim} had a median value of -3.1 dB at a corresponding median SNR_{True} of -3.4 dB. In the public lunch scenario, the median of S_R was 79.5 dB SPL in the case of the single-channel method and

75.4 dB SPL for the two-channel method, at a median of N of 75.5 dB SPL. The median SNR_{1ch} and SNR_{2ch} were 4.0 dB and -0.6 dB, respectively. SNR_{2chSim} had a median value of 1.2 dB for a median SNR_{True} of 1.5 dB.

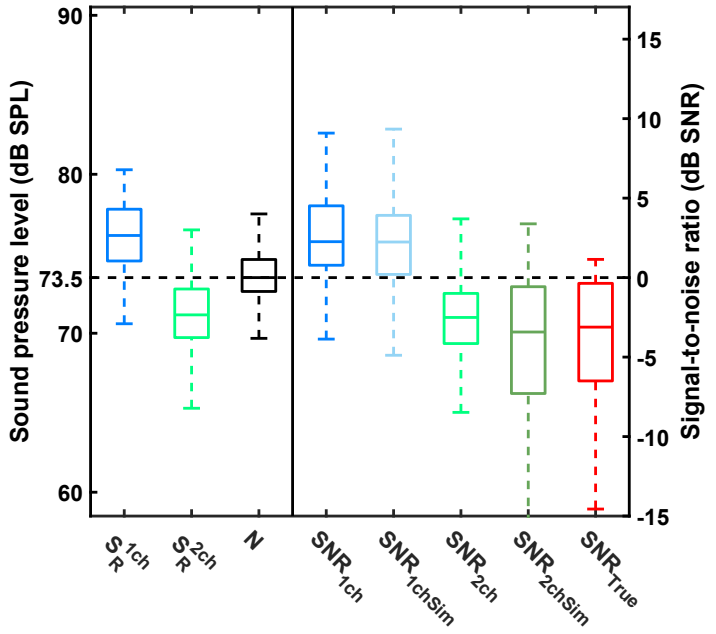


(A) Office meeting scenario SNR simulation results

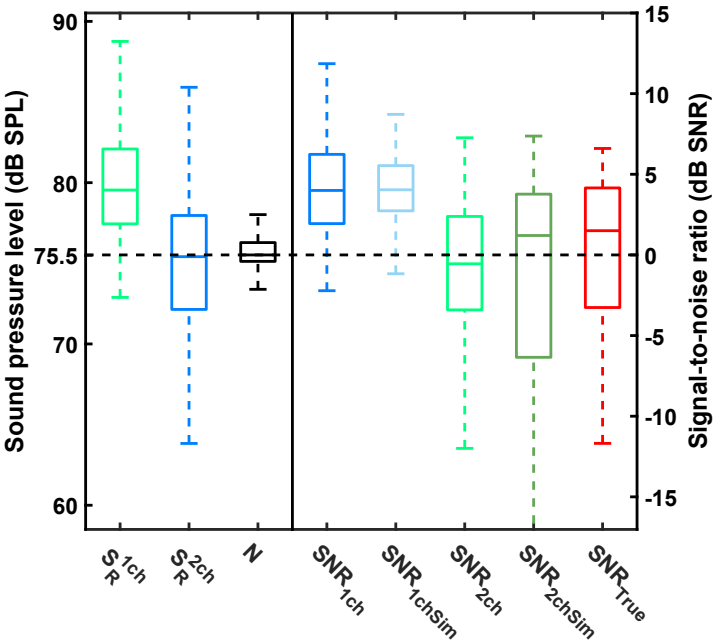


(B) Public lunch scenario SNR simulation results

Figure 2.6: Room acoustic SNR simulations for the office meeting scene (panel A) and the public lunch scene (panel B). The true SNR distributions (SNR_{True}, right, red) around the median, and the corresponding SNR distributions obtained with the single-channel (SNR_{1ch}, middle, blue) and two-channel (SNR_{2ch}, right, green) methods are shown.



(A) Office meeting SNR measurement results



(B) Public lunch SNR measurement results

Figure 2.7: For the office meeting (panel A) and public lunch scenario (panel B), speech level distributions obtained with the single-channel (S_R^{1ch}, blue) and two-channel (S_R^{2ch}, green) methods as well as the common background noise level distribution (N, black) are shown alongside SNR distributions for the single-channel method (SNR_{1ch}, blue), the two-channel method (SNR_{2ch}, green), the simulated single-channel method (SNR_{1chSim}, blue) matched at the median to SNR_{1ch}, the corresponding simulated two-channel method (SNR_{2chSim}, green) and simulated true SNR (SNR_{True}, red). The speech and noise level distributions use the left, dB SPL ordinate, while the SNR distributions use the right, dB SNR ordinate.

A one-way analysis-of-variance test showed that the speech level and SNR distributions were significantly higher for the single-channel method compared to the two-channel method, both in the office meeting and the public lunch scenario ($p \leq .0001$ when comparing S_R^{1ch} to S_R^{2ch} and SNR_{1ch} to SNR_{2ch}). Also in both scenarios, the SNR_{2chSim} distribution was significantly lower than the SNR_{1chSim} distribution, but not significantly different from either the SNR_{2ch} or the SNR_{True} distributions.

2.4 Discussion

The room acoustic simulation results clearly showed that the single-channel method consistently overestimated the true SNR, measured across a range of evaluated SNRs. The two-channel method approximated the true SNR very closely. Since the N signal was estimated in the same way for both methods, the difference was caused by the S_R signal estimations. The single-channel method assumes that speech and noise signals are uncorrelated, which is not the case for the multi-talker babble noise signal used here and therefore results in an overestimation of the clean speech power. This challenge did not arise in the two-channel method since $P(S_R)$ was derived directly from the M_{CM} signal.

In addition, the single-channel method suffered from saturation at SNRs below -10 dB, regardless of the true input SNR. This happens because, at low SNRs, $P(S_R)$ becomes small compared to the underlying $P(N)$, such that the SNR distribution essentially reflects the distribution of $P(N)$ during target speech relative to $P(N)$ during speech pauses (Kim and Stern, 2008). The two-channel method's use of the M_{CM} avoids such saturation. Lastly, while the implementation of the single-channel method in the present study avoided practical target-speech-segment labeling issues by reusing the two-channel method's VADs, the hand-labeled data in the reference studies may have been affected by the resulting under-representation of low speech levels in the SNR distributions.

Since the simulated two-channel method only differs conceptually from the true SNR in its approximation of $P(N)$ by $\tilde{P}(N)$, its slightly differing estimates occurred because the distribution of N during target speech and during speech pauses was not identical. This was more apparent in the public lunch scenario than in the office meeting, since the higher DRR and C_{50} values in the public lunch reflected a more fluctuating N. Nevertheless, the two-channel method approximated the true SNR far more closely than the single-channel method.

With regard to the real-world measurements, the potential effect of the target speech presence on the noise level as well as the likely violation of the assumptions of anechoic, noise-free target speech and the isotropic receiver noise need to be considered. The measured speech, noise and SNR distributions in the two real-world scenes indicated that, while the absolute S_R and N levels as well as the SNRs were higher for the public lunch scenario than for the office meeting scenario, the two-channel method provided about 4 dB lower median S_R levels and SNRs compared to the single-channel method in both scenes.

These differences were roughly consistent with the corresponding differences between the matched simulated single-channel SNR distributions and their two-channel counterparts, even though the widths of the measured two-channel SNR distributions were narrower than the simulated ones. This reduction in width was due to the more narrow distribution of the real-world recorded speech signal compared to the simulated one. The two-channel method estimated the median of SNR_{True} in the office meeting scenario slightly more accurately than in the public lunch. This is likely due to the lower DRR and C_{50} values in the office meeting scenario, indicating a more isotropic and stationary noise field compared to the public lunch, in line with the assumptions pertaining to the N signal. Nevertheless, the two-channel measured SNR distribution's inter-quartile range was lower than that of the simulated SNR distribution, for both scenarios.

The estimated median SNRs of the two-channel method of -2.5 dB and -0.5 dB are in line with SNRs obtained in other realistic scenarios (Culling, 2016) and consistent with the notion that conversational SNRs decrease with increasing talker distance (Weisser and Buchholz, 2019). The width of the S_R level distributions was found to be smaller in the office meeting than in the public lunch scenario for both methods. One explanation for this is that talkers maintained a reasonably constant talking level at a larger fixed distance - where communication is more difficult - compared to when they are close together. This, in turn, affects the widths of the corresponding SNR distributions as well. The distributions for the background noise level were found to be rather symmetric in both scenarios, and did not differ between the estimation methods since the noise contribution was calculated in exactly the same way.

While the two-channel method most likely characterizes conversational SNRs more accurately than the single-channel approach, it has several limitations. The necessity of the cheek microphone signal implies that existing single-channel recordings cannot be re-analyzed, such that that additional measurements are needed to acquire SNR distributions in scenes other than the two described here. The fact that the room impulse response needs to be recorded and calibrated at a predefined distance implies that the method is tailored to the fixed talker distance in a specific target-receiver configuration in the scene. Additionally, the free-field correction applied to the cheek microphone signal was only measured from the front, and thus did not account for potential head movements of the target talker. The two-channel method implements one specific way of estimating the acoustic path between the target and receiver, aiming to more accurately approximate the true SNR.

Nonetheless, the proposed SNR estimation method captures real-world SNR distributions with an increased degree of accuracy compared to the single-channel approach, while also allowing for the dynamical tracking of speech levels and SNRs in real-world scenarios. It can be applied in real-world scenes for both offline data collection, as implemented here, or real-time tracking. This enables applications beyond broadband level estimation, including precise frequency-specific target speech analysis and the accurate temporal characterization of speech rates, turn-taking and conversational behavior in a realistic way.

2.5 Conclusion

A two-channel method for the SNR estimation of a target talker in conversation was developed based on a room acoustical approximation to the true SNR. With the proper calibration and setup, the method was shown to result in significantly reduced speech levels and downward-shifted SNR distributions compared to a common single-channel reference method. Median values for the two-channel method were more than 4 dB lower than for the single-channel method, likely due to an overestimation of the level of a noise-correlated speech signal in the single-channel method. As such, the proposed method might provide interesting perspectives on how conversational real-world signal-to-noise ratios can be estimated.

3

Speech intelligibility in a realistic virtual sound environment[a]

Abstract

In the present study, speech intelligibility was evaluated in realistic, controlled conditions. "Critical sound scenarios" were defined as acoustic scenes that hearing aid users considered important, difficult and common through ecological momentary assessment. These sound scenarios were acquired in the real world using a spherical microphone array and reproduced inside a loudspeaker-based virtual sound environment (VSE) using Ambisonics. Speech reception thresholds (SRT) were measured for normal-hearing (NH) and hearing-impaired (HI) listeners, using sentences from the Danish Hearing In Noise Test, spatially embedded in the acoustic background of an office meeting sound scenario. In addition, speech recognition scores (SRS) were obtained at a fixed signal-to-noise ratio (SNR) of -2.5 dB, corresponding to the median conversational SNR in the office meeting. SRTs measured in the realistic VSE-reproduced background were significantly higher for NH and HI listeners than those obtained with artificial noise presented over headphones, presumably due to an increased amount of modulation masking and a larger cognitive effort required to separate the target speech from the intelligible interferers in the realistic background. SRSs obtained at the fixed SNR in the realistic background could be used to relate the listeners' SI to the potential challenges they experience in the real world.

[a] This chapter is based on Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (submitted); Speech intelligibility in a realistic virtual sound environment.

3.1 Introduction

Through their auditory perception, normal-hearing people are able to communicate nearly effortlessly even in challenging acoustic scenarios, such as at a social gathering or in a busy restaurant. In contrast, a person whose hearing is impaired often experiences diminished speech communication ability, hindering many social interactions (Moore, 1996). Hearing aids (HA) aim at compensating for hearing deficits by employing frequency- and level-dependent amplification to restore the wearer's sensitivity to soft sounds, compensate for loudness recruitment and increase overall sound quality. However, despite considerable technological advances in hearing aid technology, hearing aid benefit varies greatly among individual users, particularly in reverberant situations with multiple interfering sound sources.

To appreciate why this occurs, it is necessary to understand how human hearing is currently evaluated in the context of hearing aid applications. Typically considered paradigms, such as loudness perception or speech intelligibility, utilize well-defined, artificially created acoustic stimuli presented over headphones or small sets of loudspeakers. This approach, while having the advantage of being fully controlled and replicable, does not necessarily reflect conditions in the real world.

With respect to the acoustic stimuli employed in speech intelligibility (SI) para-digms, the often used speech-shaped stationary noise (SSN) maskers differ considerably from actual multi-talker acoustic interferers. SSN lacks the typical low-frequency modulations of multi-talker interferers and therefore does not provide the listener with the opportunity to utilize speech "glimpses" in the interferer (Dreschler et al., 2001). The Hearing In Noise Test (HINT), that has been widely used to evaluate speech intelligibility across many languages and in various acoustic conditions, typically uses SSN as its masker, or modulated noise lacking intelligible interferers. This test, along with other approaches like matrix-based sentence tests (Houben et al., 2014; Kelly et al., 2017; Wagener et al., 2003), results in normal-hearing (NH) speech reception thresholds (SRT) - corresponding to 50% speech intelligibility - that are well below a signal-to-noise ratio (SNR) of 0 dB (Nielsen and Dau, 2011; Soli and Wong, 2008; Wagener et al., 2003). In contrast, research trying to categorize real-world SNRs has consistently found a substantially higher range of values in the majority of sound scenarios (Smeds et al., 2015).

The presentation of the stimuli in an SI task is equally problematic in terms of realism. Headphones can "spatialize" the presented sounds using head-related-transfer functions (Wightman and Kistler, 1989), but this approach is limited in accuracy by many factors, including headphone placement and the limited spatial resolution imposed by the angle between the discrete functions. In addition, accounting for head movements and fitting a HA with headphones is cumbersome in practice. Quadraphonic loudspeaker setups, often used for the spatial evaluation of HA algorithms, physically separate the noise maskers to alleviate these problems, but they generally still do not faithfully reproduce spatially diffuse noise (ITU-T, 2018). In short, an SI task presenting artificial stimuli in a simplified spatial manner might misrepresent the difficulties that both NH and HI people experience when listening to speech in noise in their daily lives. Due to these discrepancies, it has remained unclear how SI performance scores in laboratory settings relate to these real-world difficulties (Culling, 2016).

To more precisely tailor the performance of a hearing aid to the needs of a user, it would be advantageous to utilize an SI testing paradigm that mimics conditions in the real world to the highest possible degree, thereby making it more ecologically valid (Reis and Judd, 2000). One option could be to actually conduct the SI task in the real world (e.g. through field tests). While perfectly realistic, real-world acoustic conditions are highly variable, resulting in outcome measures that would be difficult to interpret and reproduce. Attempting to bring the real world into the lab represents a trade-off between control and realism, both regarding stimulus choice and acoustic presentation. A proper balance of "controlled realism" would have the potential to result in consistent, yet ecologically valid, findings (Best et al., 2015).

A virtual sound environment (VSE) in the form of a spherical loudspeaker array is able to render complex three-dimensional sound fields at its center through Higher-Order Ambisonic (HOA) reproduction techniques (Bertet et al., 2006). Using such an array to present target speech sentences superimposed onto spatial recordings of realistic sound scenarios would ensure the reproduction of acoustic sound field properties within the limitations of the recording setup, allowing for head movements and providing a sense of spatial immersion. VSEs have been used extensively in combination with simulated spatialized maskers based on room-acoustic simulations (e.g. ODEON A/S, 2020) to study aspects of auditory spatial separation (Best et al., 2017a), informational masking (Westermann and Buchholz, 2015), hearing aid performance (Cubick and

Dau, 2016; Minnaar et al., 2010) and speech intelligibility (Ahrens et al., 2017; Best et al., 2015; Westermann and Buchholz, 2017). However, this approach is still limited by the number and complexity of sources that can be simulated and has been shown to correlate only poorly with real-world conditions (e.g. Ahrens et al., 2019). While real-world HOA recordings have become increasingly available (e.g. Weisser et al., 2019b), there exists, to the best of the authors' knowledge, no research that utilizes such spatially recorded maskers in VSE-based SI tasks. In addition, for these recorded sound scenarios (e.g. an office meeting or a restaurant visit) to become more ecological valid, they should be selected based on scenarios that users consider critical in their lives and captured in a real-world environment.

In the present study, a speech intelligibility task is presented that aims to increase ecological validity. A set of critical sound scenarios was selected based on a categorization of HA user ecological momentary assessment (EMA) data (Smeds et al., 2018). Out of these scenarios, an office meeting scenario was recorded in-situ with a spherical microphone array. This recording was subsequently reproduced as a VSE using Ambisonics over a 64-channel, fully spherical loudspeaker array inside an anechoic enclosure. Finally, the reproduced masker was combined with the spatialized speech corpus of the Danish HINT as part of a speech intelligibility task carried out by NH and HI listeners. Adaptive SRTs were captured as well as speech recognition scores (SRS) at a constant SNR of -2.5 dB, corresponding to the median conversational SNR between NH people in the office meeting scenario (Mansour et al., 2021). The hypotheses were that (i) a speech intelligibility paradigm employing realistic, spatialized stimuli would produce higher SRTs compared to those obtained with an artificial approach; and that (ii) SRSs at a real-world conversational SNR would reflect some of the difficulties HI people experience in the real world.

3.2 Methods

3.2.1 Sound scenario selection

To increase the relevance and ecological validity of the recorded critical sound scenarios as potential maskers in the speech intelligibility task, the scenario selection was based on EMA data from 281 field reports by HA users, collected by Smeds et al. (2018).

In EMA, user data are captured in real-time by subjects in everyday scenarios. The use of EMA data obtained in this way has become increasingly popular in attempts to describe and characterize which scenes HA users experience (Timmer et al., 2017).

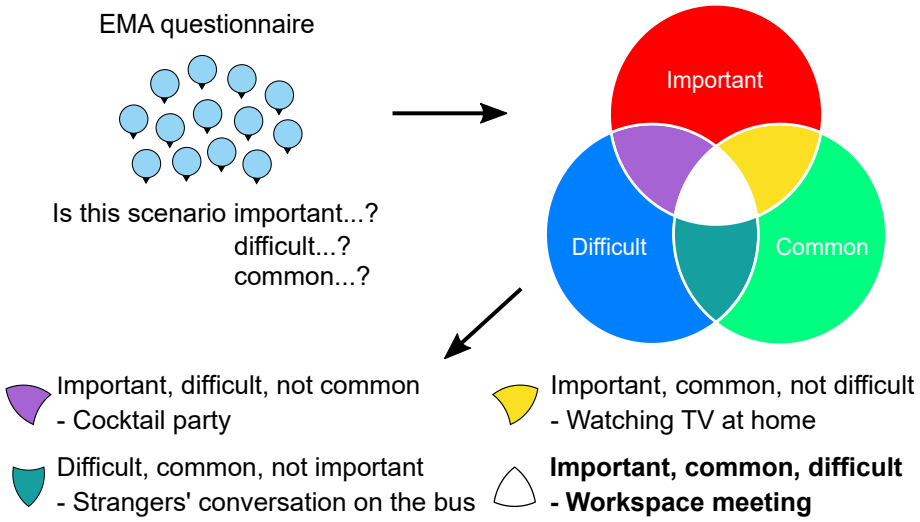


Figure 3.1: The critical sound scenario framework, developed to categorize HA users' ecological momentary assessment (EMA) field report data based on a binary combination of metrics of reported importance, difficulty and occurence of understanding speech in everyday scenarios. Examples of scenarios at the different intersections are given, with the area considered important, common and difficult displayed in bold.

Figure 3.1 shows the critical sound scenario framework that was developed in this study and which categorized a real-world scenario based on the binary combination of three EMA questionnaire metrics: the reported importance, difficulty and occurence of understanding speech in that scenario. Combinations of these three parameters have been shown to accurately separate different real-life situations (Wolters et al., 2016). The examples provided in Fig. 3.1 illustrate different combinations of these metrics, graphically depicted as a Venn diagram. For instance, while watching TV at home is a common and important scenario for HA users, it is generally not considered as difficult. Understanding speech in cocktail party scenarios is important and difficult even for normal-hearing people, but not very common for many HA users. Workspace meetings and having lunch in public are considered important as well as difficult and common. For the scenarios rated as important, the percentages displayed in the four sections of the corresponding circle in the Venn diagram denote their relative occurrences.

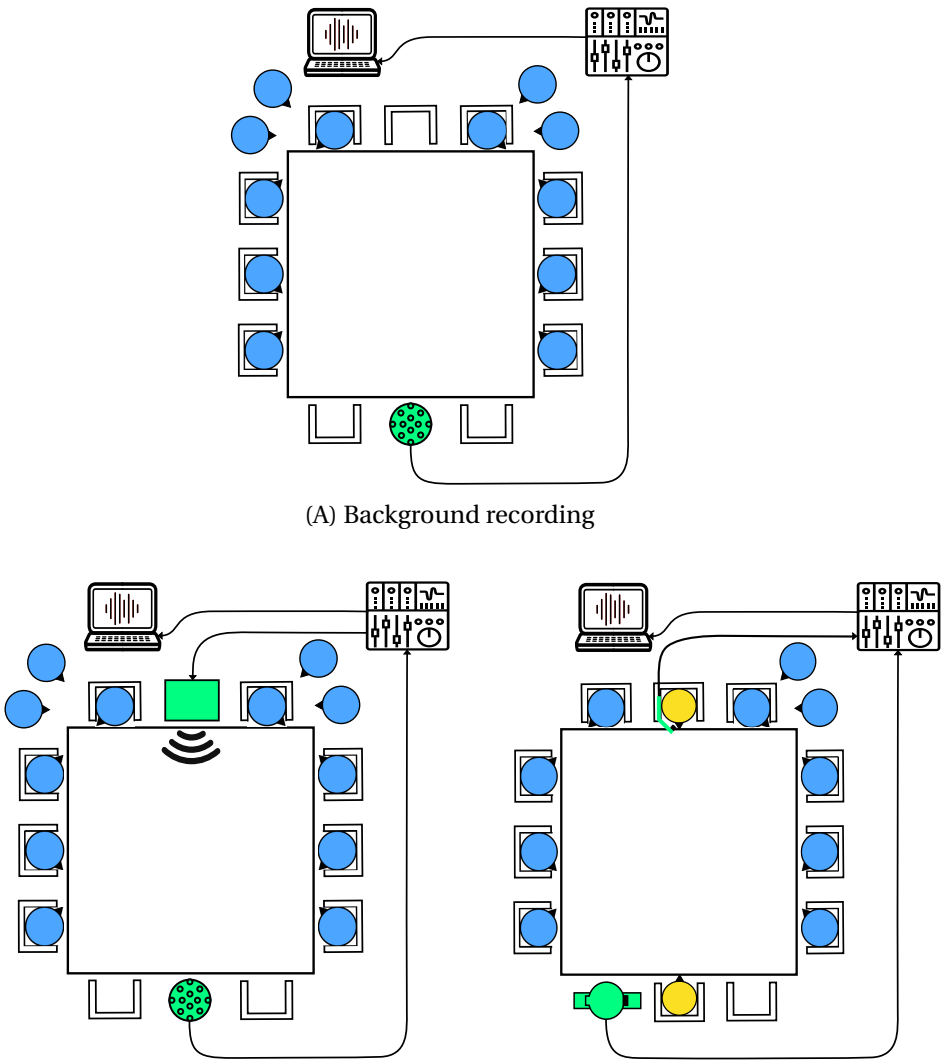
In the present study, the subset of scenarios that were simultaneously rated important, difficult and common were chosen for further analysis as they represent conditions in which HA users are challenged the most. This subset was cross-referenced with the EMA reports to reveal the three most prevalent critical sound scenarios: a public lunch, a small festive event (e.g. a family house party) and a workspace meeting. From these, the workspace meeting scenario was selected for further processing.

3.2.2 Sound scenario acquisition

Figure 3.2 illustrates the recording setup used to acquire the office meeting scenario. The scenario was captured during a staged office meeting with a spherical microphone array (em32 Eigenmike, mh acoustics LLC, USA) capable of 4th HOA recording (Bertet et al., 2006) with its spatial aliasing frequency at 9 kHz. In addition, a Knowles Electronic Manikin for Acoustic Research (GRAS A/S, 2018) with ear canals was used to capture binaural signals. While 12 participants conversed in pairs, seated at and standing around a conference table in the office meeting room (Fig. 3.2A), spatial scene recordings were obtained with the EigenMike and KEMAR in the listener position at the bottom center of the table. Room impulse responses (RIR) were captured from the target position (top center of the table) using a mounted loudspeaker producing a series of three repeated 15-second exponential sweeps (Müller and Massarani, 2001) between 20 Hz and 20 kHz, while all participants remained still and quiet to avoid altering the room reverberation (Fig. 3.2B).

To obtain conversational signal-to-noise ratios resulting from two interacting participants seated in the listener position and the target position, respectively (Fig. 3.2C), the method detailed in Chapter 2 (Mansour et al., 2021) was used. In this method, speech produced by the target speaker was recorded via a DPA 4066 cheek microphone. The recorded cheek microphone signal was free-field corrected and convolved with the captured impulse response to obtain an estimate of the target speech at the listener. The free-field correction was obtained as the transfer function between white noise recorded by the cheek microphone mounted on the KEMAR in an anechoic chamber and a reference microphone at 0.5m distance. The background noise was measured with the right ear of the KEMAR during gaps where neither the target nor the listener was speaking, and the resulting SNR was calculated as the ratio of the speech energy at the receiver and the background noise energy. Energy-based broadband

voice activity detection (VAD, Kinnunen and Li, 2010) was used to separate the target and listener speech from the background. The median value of the SNRs obtained in this manner, -2.5 dB, was used in the constant-SNR SI assessment.



(A) Background recording

(B) RIR recording

(C) SNR estimation

Figure 3.2: Illustrations of the office meeting scenario acquisition stages, consisting of the background scene recording (panel A), the room impulse response (RIR) recording stage (panel B) and the conversational SNR estimation stage (panel C). Each panel contains the human participants (blue heads) distributed around a large, square conference table, as well as the sound card and recording laptop. The spherical microphone array is depicted in the listener position at the bottom of panel A and panel B (green circle). Panel B additionally includes the loudspeaker in the target position at the top (green rectangle). The target talker, wearing the cheek microphone (green line), and the listener are depicted in the panel C in the target and listener positions, respectively (grey heads). The KEMAR is positioned to the left of the listener (green mannequin).

3.2.3 Sound scenario reproduction

The spherical microphone recordings made with the Eigenmike were encoded from the 32 raw input channels to a 25-channel Ambisonic 4th-order HOA format. These HOA signals were then rendered on the 64-channel spherical loudspeaker array in the AudioVisual Immersion Lab (AVIL) at DTU (Fig. 3.3A) using dual-band (basic, max-rE) decoding, with a crossover frequency of 2400 Hz. HOA auralization was chosen because of its physically faithful rendering of sound fields in the sweet spot at the center of the array (subject to the limitations imposed by the spatial aliasing frequency of the microphone array), ensuring their usability for HAs as well as human ears.

The level of the reproduced masker was required to not fluctuate strongly during its playback, to avoid largely varying SRTs in the SI task. To this end, specific subsections of the raw Eigenmike recording were extracted and concatenated before Ambisonic rendering based on level estimates derived from its front-facing microphone. The 10-minute-long recording was segmented into frames of 5 seconds (with 80% overlap) and level differences (in dB) were calculated between consecutive frames. The upper and lower boundaries for allowed level differences were set to the 5th and 95th percentile of the level difference distribution, respectively. The collection of consecutive frame segments within these boundaries was retained and concatenated in decreasing duration (from 25 seconds to about 7 seconds), resulting in a level-equalized recording of 2.5 minutes. The full 32-channel synchronized version of this reduced recording was rendered to the 64-channel reproduction, calibrated segment by segment to a fixed target sound pressure level (SPL) of 73.5 dB using a B&K 2669 free-field microphone, and cross-faded with a 1-s Hann-windowed overlap for smooth transitions between segments. The target SPL was selected as the median value of the noise level distribution measured during the conversational SNR estimation stage. Finally, the resulting 2-minute-long background reproduction was calibrated binaurally inside the loudspeaker array, using a B&K Type 4128 Head and Torso Simulator (HATS) with ear canals. This approach ensured that the background reproduction retained its intelligible properties despite having been dynamically stabilized in level.

3.2.4 Speech stimuli and interferers

To evaluate speech intelligibility, the Danish HINT (Nielsen and Dau, 2011) was used. The HINT uses brief, mundane, male-voiced, 5-word target sentences presented in speech-shaped stationary noise (SSN) to estimate SRTs using an adaptive, 1-up-1-down, sentence-based scoring procedure. During each trial, consisting of a sequence of 20 sentences, the procedure decreased the SNR of a sentence by 2 dB if the previous sentence was repeated back entirely correctly and increased it by 2 dB otherwise. The initial SNR was set to 0 dB and the first sentence was replayed at increasing SNRs until it was repeated back correctly, before continuing. The SRT of a trial was then calculated as the average of the SNRs across the last 15 sentences. In addition, a non-adaptive procedure, which presented the sequence of 20 sentences at a constant SNR, was implemented to estimate speech reception scores in % correct, testing the second hypothesis that SRSs at real-world SNRs reflect difficulties with speech intelligibility. Three conditions were evaluated for each procedure:

- (HP) The classical HINT reference condition, where anechoic target sentences were presented diotically in SSN over headphones, served as the control condition. In the following, this condition is referred to as the "headphone condition" (HP).

- (RE) The primary spatial condition used spatialized HINT target sentences that were integrated into the reproduced office meeting interferer and presented through the loudspeaker array. The spatialized sentences were obtained by convolving the 60 training and 200 test HINT sentences individually with the spherically recorded IR that was captured between the target and listener positions. Each sentence was calibrated individually to retain the same unique level as that of the single-channel version, measured in the sweet spot of the loudspeaker array. This condition is referred to in the following as the "realistic noise" condition (RE).

- (AR) The secondary spatial condition, with similarly spatialized HINT sentences presented in a decorrelated quadraphonic version of the HINT SSN playing from four loudspeakers in the array at 45°, 135°, 225° and 315° azimuth, 0° elevation. This "artificial noise" condition (AR) was included primarily to investigate the effect of changing only the type of spatialized background noise on speech intelligibility.

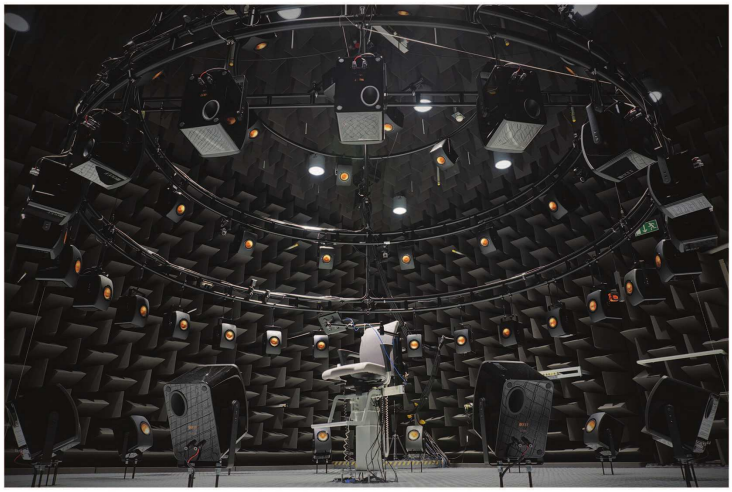
In each condition, speech-to-noise SNRs were established by varying speech sentence levels with respect to the continuously playing, looped background. The fixed SNR used in the non-adaptive procedure was set to the median SNR of -2.5 dB obtained from the conversational assessment (Sub. 3.2.2), considered as a representation of a realistic NH conversational SNR.

3.2.5 Listeners

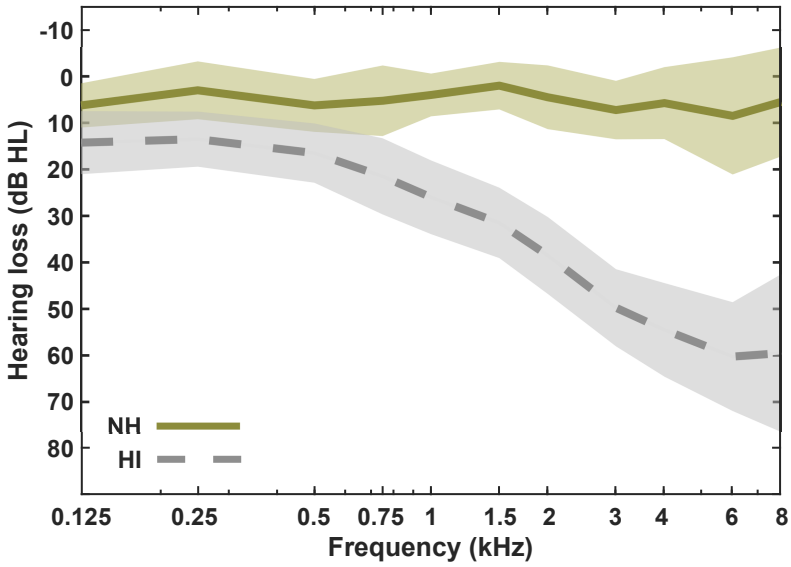
Ten NH and ten HI listeners participated in the experiment. The NH listeners were between 21-69 years old with a median age of 28, while the HI listeners were between 56-75 years old with a median age of 70. The NH listeners had a four-frequency average hearing loss (HL) of maximally 15 dB, while the HI listeners had an average sloping mild (N2) to moderate (N3) hearing loss (Bisgaard et al., 2010). Figure 3.3B shows the individual audiograms as well as their mean (with shaded standard deviations) for the NH (gold, solid) and HI (silver, dashed) listener groups. All HI listeners had a word discrimination score in quiet of at least 92% for both ears, and a left-right-ear HL difference of maximally 10 dB for all frequencies. All listeners provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

3.2.6 Speech intelligibility procedure

Three adaptive training rounds were carried out for target speech in quiet (to ascertain audibility), as well as for the target speech presented in the AR and RE conditions. Then, two evaluation rounds were conducted for all adaptive conditions, and one for all constant-SNR conditions. One HINT round contains a sequence of 20 predetermined sentences, presented in random order. The testing order was randomized over condition (HP-RE-AR) and test list number (1-9) through the use of two 9x9 Latin squares (Bradley, 1958) with two random completions. Within one condition, two adaptive test lists were always followed by one at the constant SNR. The two corresponding SRTs were averaged to obtain a final SRT, and an SRS was established as the percentage of correctly understood words in the constant-SNR list. The SNRs in the adaptive procedure were adjusted based on sentence scoring, where a 5-word target sentence is marked correct only when all 5 words were repeated accurately by the listener. This is the standard way of scoring the Danish HINT (Nielsen and Dau, 2011).



(A) AVIL



(B) NH and HI listener audiograms

Figure 3.3: The AudioVisual Immersion Lab (AVIL), serving as the reproduction laboratory, containing a spherical loudspeaker array in an anechoic enclosure (panel A) and the mean audiograms of the normal-hearing (NH, solid gold) and the hearing-impaired (HI, dashed silver) listener groups as well as their standard deviations (panel B).

The SRSs in the constant-SNR procedure were based on word scoring, where the number of correctly repeated words in every sentence is counted, summed over all sentences in a list and divided by 100. The decision to use word scoring for the constant-SNR procedure was taken to increase the scoring sensitivity of the SI task, thereby avoiding flooring effects, at an SNR were HI listeners were expected to struggle considerably.

All listeners provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The experiment lasted, on average, one hour and the HINT scoring was carried out by a native Danish speaking audiologist.

3.2.7 Questionnaire and statistical analysis

In addition to the objective SI assessment, all listeners were asked to fill out a questionnaire after completion of the experiment. Table 5.1 displays the questions that were asked, pertaining to the realism of the sound of the stimuli and the difficulty in understanding speech. The response scale was a 5-point Likert scale, asking the respondent to rate a percept from not at all present (1) to extremely present (5).

To check the data obtained in the different conditions (HP_{NH}, HP_{HI}, RE_{NH}, RE_{HI}, AR_{NH} and AR_{HI}), a two-way mixed-effects analysis of variance (MANOVA) statistical test was used where the condition HP/RE/AR represented a within-listener factor and the hearing status NH/HI represented a between-listener factor. The normality of each group was verified with the Anderson-Darling and Shapiro-Wilk tests and the similarity in variance between the compared groups required for the MANOVA was evaluated with a Bartlett test. A one-way analysis-of-variance (ANOVA) test was applied to investigate specific paired comparisons between NH and HI listeners, as well as a repeated-measurement ANOVA (RANOVA) to compare between HP/RE/AR conditions within a listener group. In all tests, the significance level was set at 5%.

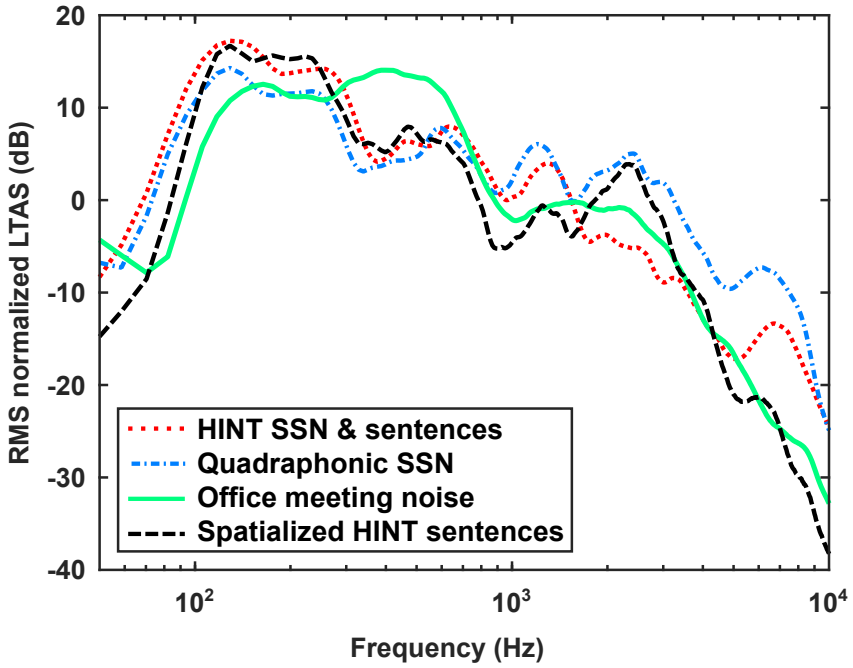
3.3 Results

3.3.1 Acoustic properties of the stimuli

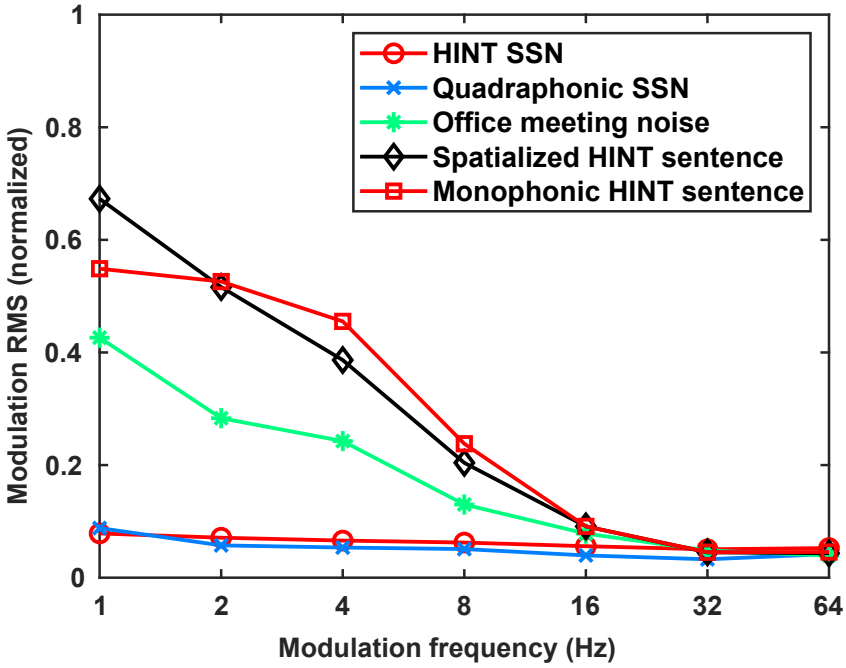
Figure 3.4 shows the long-term average spectra (LTAS, panel A) and the modulation power spectra (panel B) for the speech-shaped noise used by the HINT (red, dotted), the quadraphonic SSN (blue, dot-striped), the office meeting noise (green, solid), and a concatenation of all two hundred HINT test sentences, monophonic and spatialized with the office meeting room impulse response (black, dashed). The LTAS functions were normalized to have the same broadband RMS. The modulation power spectra were obtained by normalizing the calculated power within a modulation band by its respective bandwidth as well as the power in the DC component (Dreschler et al., 2001). The quadraphonic noise, the office meeting noise and the auralized HINT sentences were recorded binaurally with the HATS inside the loudspeaker array (left-ear spectra shown).

The LTAS of the HINT SSN (panel A) represents its speech-shaped spectral character. The averaged monophonic HINT sentence LTAS is identical to that of the HINT SSN, since this noise was originally constructed from the averaged power spectrum of one hundred HINT sentences. For the quadraphonic SSN, the LTAS is increased by up to 10 dB relative to the monophonic version at frequencies above 1100 Hz. The LTAS of the spatialized HINT sentences reflects the effect of the room on the speech-shaped stimuli. Similarly, the LTAS of the office meeting noise shows its speech-like nature, smoothed and altered by the room reverberation.

For the modulation spectra (Fig. 3.4B), the quasi-stationary nature of the HINT SSN is reflected by its low energy across the considered modulation frequencies. The modulation power remains low and roughly constant at all frequencies. The quadraphonic SSN exhibits the same modulation spectrum as the classical version. The monophonic HINT sentences contain a large amount of modulation power, typical for signals with speech-like envelope fluctuations. Spatializing these signals hardly alters these contributions. The modulation spectrum of the office meeting noise lies in between these two extreme patterns, since it contains various talkers speaking in a reverberant environment. As such, the HINT SSN reflects the averaged spectral characteristics of the speech stimuli but not their low-frequency modulations, while the office meeting noise is speech-like in both the spectral and the modulation spectral domains.



(A) Long-term average spectra



(B) Modulation frequency spectra

Figure 3.4: Root-mean-square (RMS) normalized long-term average spectra (panel A) and RMS normalized modulation frequency spectra (panel B) of the stimuli used in the speech intelligibility task, specifically the HINT monophonic target sentences and speech-shaped noise (SSN, red), the quadraphonic SSN (blue), the office meeting noise (green) and the spatialized HINT target sentences (black).

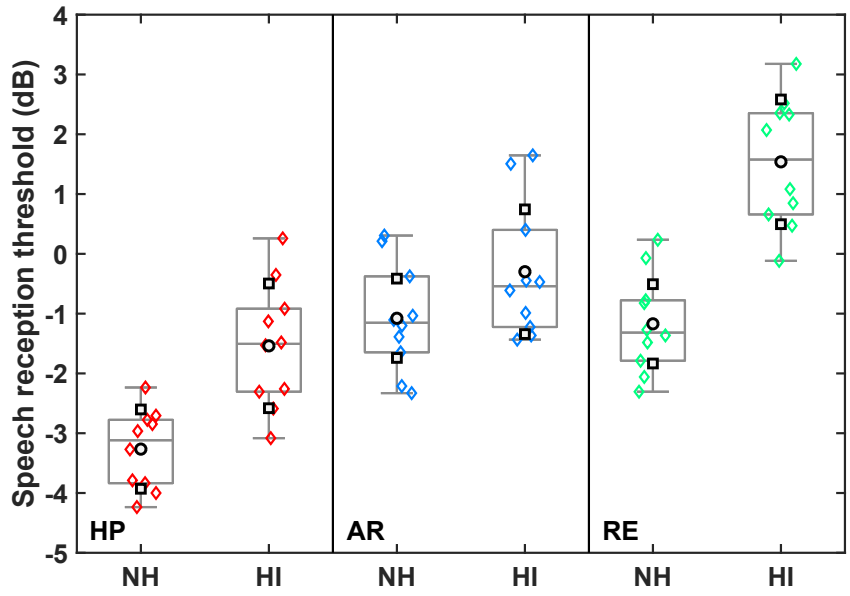
3.3.2 Speech reception thresholds

Figure 4.1A shows individual SRT results for the adaptive HINT procedure (diamonds), as well as their mean (black circle) and standard deviations (black squares), median and 25th/75th percentiles (box plot) for the NH and HI listeners in the three conditions HP (red, left), AR (blue, middle) and RE (green, right). Each individual SRT result represents the average of two SRT measurements. The mean HP SRTs were obtained at -3.3 dB for the NH listeners as opposed to -1.5 dB for the HI. The mean SRTs in the RE condition were -1.2 dB for the NH listeners and 1.5 dB for the HI listeners. For the AR condition, the mean SRT was -1.1 dB for the NH listeners, compared to -0.3 dB for the HI listeners.

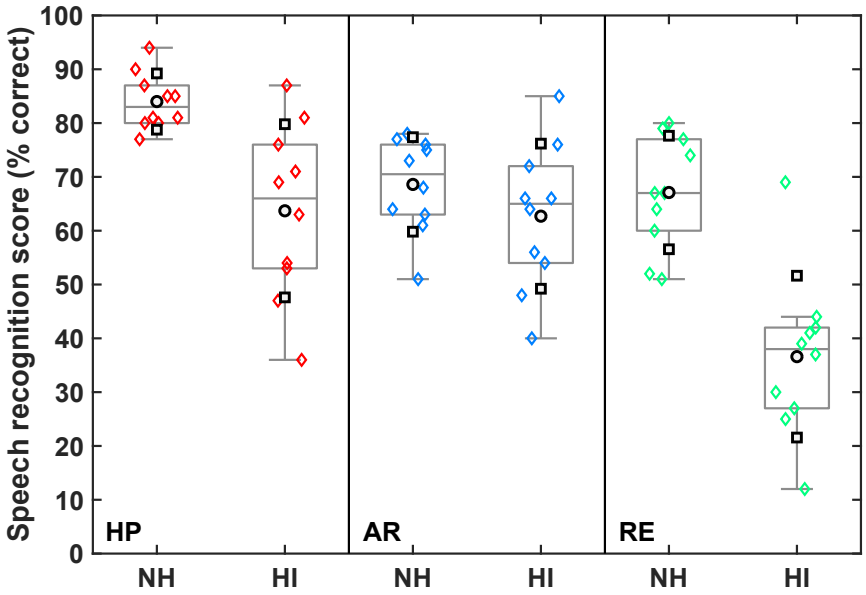
Comparing the HP to the RE condition, a significant effect of condition ($F(1,18) = 41.57, p \leq 0.0001$) and hearing status ($F(1,18) = 205.13, p \leq 0.0001$) was found, but no significant interaction between the two ($F(1,18) = 2.86, p = 0.1049$). Similarly, comparing the results obtained for the HP and AR conditions, significant effects for condition ($F(1,18) = 40.26, p \leq 0.0001$) and hearing status ($F(1,18) = 14.56, p = 0.0013$) were observed, but no interaction effect ($F(1,18) = 3.08, p = 0.0962$). Finally, a comparison of the results obtained in the AR and RE conditions revealed significant effects of condition ($F(1,18) = 8.80, p = 0.0083$) and hearing status ($F(1,18) = 27.57, p \leq 0.0001$), as well as a significant interaction ($F(1,18) = 10.78, p = 0.0041$).

Pair-wise comparisons showed that the SRTs for the HI listeners were significantly higher than for the NH listeners in the HP ($p = 0.0003$) and RE ($p \leq 0.0001$) conditions, but not in the AR condition ($p = 0.1056$). Within the NH listeners, SRTs were significantly higher in the RE ($p \leq 0.0001$) and AR ($p \leq 0.0001$) conditions compared to the HP condition, but similar between the RE and AR conditions ($p = 0.801$). For the HI listeners, SRTs were significantly increased in the RE condition compared to the AR ($p = 0.0034$) condition. Lastly, SRTs in the AR condition were significantly higher than those in the HP condition ($p = 0.034$).

These results demonstrate that speech intelligibility decreased (i.e. SRTs increased) in the RE condition compared to the HP condition in both listener groups, whereas the performance in the RE condition (compared to both the AR and HP conditions) was much more affected in the HI listener group than in the NH listener group.



(A) Speech reception thresholds



(B) Speech recognition scores at -2.5 dB SNR

Figure 3.5: Speech intelligibility results for the adaptive procedure (panel A) and the percentage correct procedure at the conversational SNR (panel B). Results are shown for the headphone condition (HP, red, left), the artificial speech-shaped noise condition (AR, blue, middle) and the office meeting noise condition (RE, green, right), for normal-hearing (NH) and hearing-impaired (HI) listeners. The box plots show median values and inter-quartile ranges. Individual data points are shown (diamonds) as well as the mean (black circle) and standard deviations (black squares).

3.3.3 Speech reception scores at the normal-hearing SNR

Figure 4.1B shows the speech reception scores obtained with the constant-SNR HINT procedure, summarizing the word scores of all listeners for the percentage correct evaluation at -2.5 dB SNR, the median SNR in the office meeting recording. The mean HP results corresponded to 84% correct for the NH listeners and 63.7% correct for the HI. The mean RE results corresponded to 67.1% correct for the NH and 36.6% correct for the HI listeners. For the AR condition, mean scores of 68.6% and 62.7% were obtained for the NH and HI listeners, respectively.

A significant effect of condition ($F(1,18) = 47.58, p \leq 0.0001$) and hearing status ($F(1,18) = 30.72, p \leq 0.0001$) was found with respect to the HP and the RE conditions, but no significant interaction ($F(1,18) = 2.56, p = 0.1272$). Significant effects for condition ($F(1,18) = 7.94, p = 0.0114$) and hearing status ($F(1,18) = 9.13, p = 0.0073$) as well as a significant interaction effect ($F(1,18) = 6.12, p = 0.0235$) were found between the HP and AR conditions. Comparing the AR to the RE condition revealed significant effects of condition ($F(1,18) = 14.52, p = 0.0013$) and hearing status ($F(1,18) = 19.85, p = 0.0003$), again with a significant interaction ($F(1,18) = 11.53, p = 0.0032$) [a].

Paired comparisons showed that percentage correct scores for the HI listeners were significantly lower than for the NH listeners in the HP ($p = 0.0013$) and RE ($p \leq 0.0001$) conditions, but not in the AR condition ($p = 0.2614$). Within the NH listeners, percent correct scores were significantly higher in the RE ($p \leq 0.0001$) and AR ($p = 0.0008$) conditions compared to the HP condition, but similar in the RE and AR conditions ($p = 0.7210$). For the HI listeners, percentage correct scores were significantly decreased in the RE condition compared to the HP ($p = 0.0003$) and AR ($p = 0.0018$) conditions, but were not significantly different between HP and AR conditions ($p = 0.8429$).

The SI scores showed the same trend as in the case of the adaptive SRT estimation procedure: 1) Speech intelligibility at -2.5 dB SNR was consistently poorer in the RE condition than in the HP condition for all listeners; 2) when comparing the RE to the AR condition, the HI listeners showed substantially poorer performance than the NH listeners. Overall, the spread of percentage correct responses for NH and HI listeners across conditions showed that neither

[a] Contrary to the SRT results, the Bartlett test rejected the null-hypothesis of equal variance between the groups ($p = 0.0337$), but given the balanced group size and the borderline significance, the MANOVA was still valid (Stevens, 2012).

ceiling nor flooring effects occurred, and that the RE condition resulted in the greatest separation between NH and HI performance.

3.3.4 Questionnaire results

Table 5.1 displays the results of the questionnaire given to all listeners. For both the NH and HI listeners, answers were accumulated per question and per response to produce the number ranges in the rightmost two columns. The highest frequency response within each response group is highlighted in bold. The results indicate that all listeners rated the background noise in the RE condition as mostly very realistic sounding, while the HI listeners experienced the speech in the RE condition as overall more realistic and difficult to understand.

Table 3.1: Content of the questionnaire given to all listeners after completing the experiment, as well as the 5-point Likert response scale. The frequency of responses of the normal-hearing (NH) and hearing-impaired (HI) listeners to each possible response are displayed in the two rightmost columns. The most often occurring response in each group is highlighted in bold.

Questions asked to the normal-hearing (NH) and hearing-impaired (HI) listeners	NH listener					HI listener				
Response: *Not at all (1) - Not that (2) Somewhat (3) - Very (4) - Extremely (5)*	1	2	3	4	5	1	2	3	4	5
How realistic did the office background noise in the experiment sound to you?	0	1	3	**4**	2	0	0	3	**6**	1
How difficult was it to understand the speech in the artificial background noise in the exp.?	1	1	**5**	2	0	0	2	**6**	2	1
How difficult was it to understand the speech in the office background noise in the exp.?	0	1	3	**5**	1	0	0	2	**7**	1
How realistic did the speech that you had to listen to in the exp. sound?	0	1	3	**5**	1	0	0	2	**7**	1

3.4 Discussion

3.4.1 Speech reception thresholds

SRTs for both listener groups were found to be 2-3 dB higher in the RE condition compared to the HP condition. This effect was likely caused by several factors. First, the HP condition used anechoic target sentences presented over headphones, as opposed to reverberant ones presented over loudspeakers in the RE condition. Thus, in the RE condition, the target speech direct-to-reverberant energy ratio (DRR) decreased considerably for the same broadband SNR, which is known to cause decreased speech intelligibility (Roman and Woodruff, 2013).

Second, the modulation spectra in Fig. 3.4B indicate the presence of modulation energy in the office meeting noise, in contrast to the stationary (and thus less modulated) speech-shaped HINT noise. These modulations were a consequence of the mixture of speech sources in the RE noise, but were less prominent than for the monophonic HINT target speech due to the room effect and the number of interfering talkers (Dreschler et al., 2001). Still, this specific type of speech-like noise can lead to energetic speech-on-speech masking of the target in both the spectral (Brungart et al., 2006) and the modulation spectral (Jørgensen and Dau, 2011) domains.

Third, the many interfering talkers in the RE noise were intelligible and distributed throughout the frontal plane of the listener position. This may have produced informational masking (Westermann and Buchholz, 2015), with a detrimental effect on SI, especially since the male gender of the target talker matched that of 10 out of 12 interfering talkers in the room (Helfer and Freyman, 2008). The overall higher variance in the obtained data for the HI listeners compared to the NH listeners was expected and was most likely caused by the differences in hearing loss across the HI listeners.

While the transition from the AR condition to the RE condition led to an increase in SRTs for the HI listeners, this was not the case for the NH listeners. This may have resulted from a combination of effects. Comparing the LTAS (Fig. 3.4A) for the quadraphonic SSN in the AR condition to the LTAS of the office noise in the RE condition, there is a considerable decrease in spectral energy above 1 kHz for the AR noise. This lower amount of high-frequency noise of the office noise might reduce its speech masking effect for the NH listeners, while the HI listeners would not benefit to the same extent due to their increasing hearing loss at higher frequencies. However, this effect could be

ruled out by testing a version of the AR noise that was spectrally matched to the office noise, rendering a similar NH and HI performance to the one reported here. Instead, one likely explanation is that the modulated maskers in the RE condition allowed for dip-listening, aiding the phonemic restoration of noisy speech (Warren, 1970) and leading to increased speech intelligibility (Peters et al., 1998). Such dip-listening ability is commonly reduced in HI listeners, negatively affecting their SI performance (e.g. Takahashi and Bacon, 1992). In addition, the effect of better-ear glimpsing on spatial release from energetic masking, a strategy used by NH listeners to increase their SI performance (Glyde et al., 2013), has been shown to be limited in HI listeners, potentially due to reduced audibility, even at increased target-to-masker ratios (Best et al., 2017b). Finally, the presence of realistic, meaningful speech in the RE condition may have increased the difficulty of the SI task to a greater extent in the HI listener group than in the NH listener group. This was evidenced by spontaneous and unanimous testimony by the HI listeners, who noted that the most challenging (and recognizable) aspect of performing the SI task in the RE condition was to not get distracted by the content of the background conversations. The NH listener group did not report these difficulties.

The increase in SRTs for both the NH and HI listeners between the HP condition and the AR condition occurred despite the fact that speech intelligibility typically increases when the target and the masker become spatially separated (Licklider, 1948), binaurally unmasking the speech from the noise (Durlach, 1963). However, the transition from diotic, anechoic target speech to spatialized, reverberant speech simultaneously decreased its spatial separation and its intelligibility. Spatial release from masking probably played a smaller role in the RE condition, since the background noise consisted of a large number of similar, interfering talkers (Freyman et al., 2001).

Besides the overall increases in SRTs and decreases in SRSs obtained in the office meeting condition (RE), the results for the HI listeners differed considerably from NH performance in this condition compared to the corresponding results in the HP and AR conditions. The virtual sound environment, combined with more realistic target and masker stimuli, therefore might reflect some of the described hearing deficits in HI listeners.

3.4.2 Speech reception scores at the normal-hearing SNR

While the adaptive SRTs results inform on changes in transitioning from an artificial to a more realistic SI task paradigm, SRSs at a constant SNR that represents normal-hearing conversation may provide insight into how the SI paradigm relates to the real world. For all conditions and both listener groups, the SRSs followed the same trend as the SRTs obtained with the adaptive procedure.

The word scoring procedure successfully avoided flooring effects, but the SRSs need to be corrected in order to compare them directly to the SRTs obtained with the adaptive, sentence-based scoring procedure. This is necessary because the word score of a HINT sentence between zero and four translates to a sentence score of zero, creating a non-linear negative bias of the sentence score versus the word score that increases with increasing word score. The distributions of the difference between the SRSs computed as word scores and those same SRSs computed as sentence scores are shown in Fig. 3.6. For all listeners, the word scores were consistently about 20-25% higher than the sentence scores across conditions, indicating that most listeners still repeated 2-3 words correctly in a sentence marked as incorrect by sentence scoring.

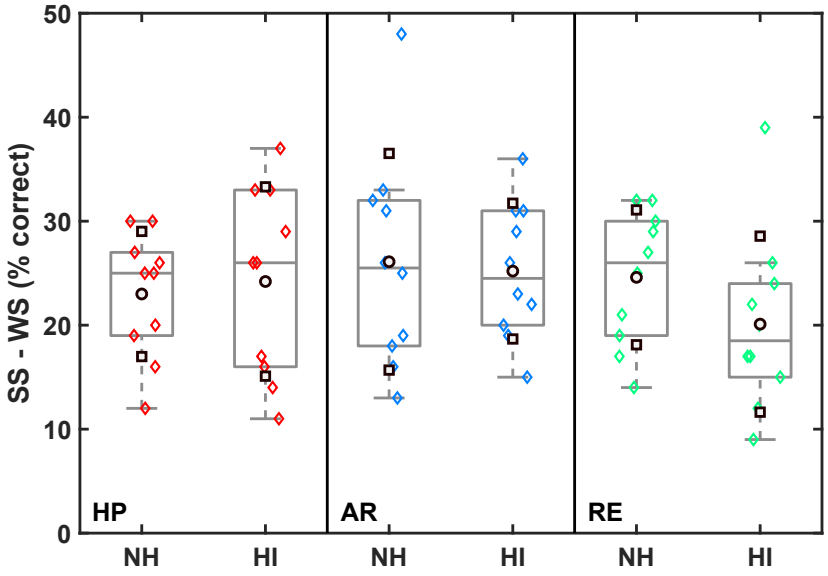


Figure 3.6: Distributions of the difference between the speech reception scores when calculated based on sentence scores (SS) and based on word scores (WS) in the percentage correct procedure at -2.5 dB SNR. Results are shown for the headphone condition (HP, red), the artificial speech-shaped noise condition (AR, blue) and the office meeting noise condition (RE, green), for normal-hearing (NH) and hearing-impaired (HI) listeners.

The RE condition shows that, on average, the HI listeners correctly received just over one word out of two, while the NH listeners correctly received two words out of three. Thus, the HI listeners understood about half as many words as the NH listeners did. While this relative comparison is irrespective of absolute performance, the SRSs for the NH listeners were only at a level of 67% for an SNR where their ability to communicate was close to 100% in the real office meeting scene. SRSs that reflect real-world SI might be used as target percentage correct scores for NH listeners when conducting an adaptive SI tasks in other VSE-based critical sound scenarios to relate the obtained SRTs back to real-world SNRs necessary for proper speech communication. A percentage correct task at these SNRs would then reveal the comparable HI performance. Once an appropriate SNR for real-world NH SI has been established, this method could be used in any SI task to relate the performance of NH listeners to that of HI listeners.

Lastly, Fig. 3.7 shows the psychometric functions of the HP condition (red, panel A) and the RE condition (green, panel B) for the NH listeners (solid line) and the HI listeners (dashed line), derived by fitting a cumulative normal distribution to the pooled percentage correct scores per discrete SNR data point for the adaptive SRT procedure. These functions thus represent performance based on sentence scoring. The mean SRTs, corresponding to the 50% correct point on the psychometric function, are indicated by straight dashed lines intersecting the black diamonds, and the aggregated NH and HI percentage scores are represented by diamonds and circles, respectively. The RE condition resulted in narrow, steeply sloped psychometric functions for both listener groups, comparable to those obtained with in the HINT HP condition. The realistic VSE therefore seems to provide sensitive as well as stable SI outcome measures.

The SI task implemented in the office meeting VSE still remained limited in realism in several ways. No visual stimuli were presented in the laboratory environment alongside the auditory signals. In the HP condition, the absence of visuals matched the HINT procedure it represented, since no visual stimuli were used there either. It has been shown that speech reception scores can increase by 20% or more when the face of the target speaker is visible to the listener (Neely, 1956), an effect which becomes especially important at negative SNRs (Sumby and Pollack, 1954) and high background noise levels (Hadley et al., 2019).

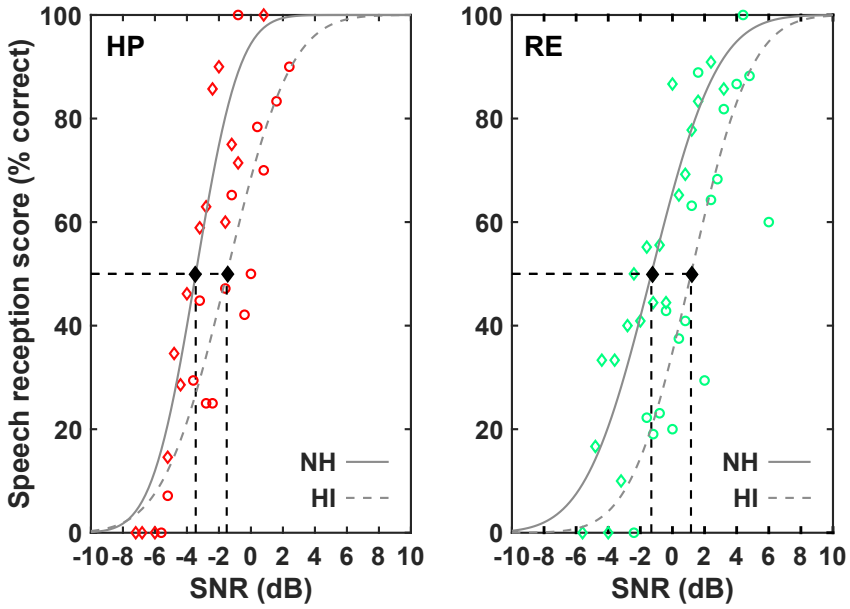


Figure 3.7: Psychometric functions for the HP condition (red, panel A) and RE condition (green, panel B) for the NH (solid line) and HI (dashed line) listeners. The NH and HI aggregated percentage scores are shown as diamonds and circles, respectively. The straight dashed lines that intersect the black diamonds relate the SNRs for both listener groups and conditions to the 50% correct point on the psychometric function.

With regard to the acoustical reproduction accuracy of the SI stimuli, the VSE condition remains limited by the applied HOA recording and reproduction methods. The spatial aliasing frequency of the microphone array reduces the acoustic reliability of the office meeting recording at frequencies beyond 10 kHz. The Ambisonic reproduction order of 4 used by the loudspeaker array guarantees a sufficiently large sweet spot for the listener, but might not supply enough spatial accuracy to accurately replicate narrow acoustic sources. However, this reproduction error is offset by the presence of reverberation in the reproduced environment (Oreinos and Buchholz, 2015).

Furthermore, only a limited number of conditions were considered in this study, due to limitations in the size of the SI speech corpus as well as time limitations in listener participation. It may be valuable to consider a condition with an unintelligible, phase-scrambled version of the office meeting background noise to assess the relative impact of informational masking and cognitive effort on SI performance, or to evaluate conditions with anechoic target speech in the spatialized maskers.

Lastly, while the experimental setup considered in this study was elaborate, it is not given that this level of sophistication is required to capture real-world SI performance. However, developing laboratory environments that approximate reality with increasing accuracy is a worthwhile endeavor, increasingly enabling the assessment of psychoacoustic phenomena beyond SI in an empirical way.

A more qualitative argument in support of increasing realism in SI paradigms is the juxtaposition of experimental realism to mundane realism, as defined in psychology. Mundane realism refers to experimental conditions that mimic those of the real world as closely as possible, whereas experimental realism indicates the extent to which listeners actually experience those conditions as realistic (Aronson et al., 1990). Therefore, to obtain meaningful results from a listener, his or her perception of realism may be just as important as its objective realization. Despite the mentioned limitations of the proposed SI paradigm, the questionnaire results from Table 5.1 confirmed that the experimental realism experienced by all listeners with respect to the sound of both the office meeting background noise as well as the speech stimuli was high. It was interesting to observe that, despite slight numerical differences, the overall distributions of difficulty and realism ratings were very similar for both listener groups. While the NH listeners achieved lower SRTs than the HI listeners, they rated the task in the realistic environments as similarly difficult as the HI listeners because the 50%-correct, adaptive HINT procedure presented both listeners groups with target speech sentences at similarly challenging SNRs.

3.5 Conclusion

A speech intelligibility task was designed and implemented, aiming to increase ecological validity and experimental realism with respect to the nature and presentation of the acoustic stimuli. It was shown that both NH and HI SRTs obtained in an HOA-reproduced office meeting critical sound scenario were, on average, 2-3 dB higher compared to the headphone-based HINT reference condition. These differences were found to be mainly due to the spatialization of the background noise (causing reverberation), the presence of speech-like modulations (causing speech-on-speech modulation masking) and the intelligibility of the interfering talkers (causing informational masking). Comparison with a spatialized artificial noise condition revealed that the HI listeners were more negatively affected by the realism in the VSE than the NH listeners, likely

due to their reduced ability to use better-ear listening and listening in the dips, as well as due to an increased cognitive effort to focus on the target speech in the presence of intelligible, interfering speech-like noise. SRSs provided a way to relate SI performance to potential difficulties experienced by HI listeners in the real world, by evaluating SI at a constant SNR at which NH communication ability was close to 100%. The approach presented in this study might be valuable for investigations into the effects of hearing loss and hearing aid benefit on SI in simulated real-world environments and could be extended by providing visual information to increase the realism of the simulated environment.

4

The effect of hearing aid dynamic range compression on speech intelligibility in a realistic virtual sound environment[a]

Abstract

Measures of "aided" speech intelligibility (SI) in listeners wearing hearing aids (HA) are commonly obtained using rather artificial acoustic stimuli and spatial configurations compared to those encountered in everyday complex listening scenarios. In the present study, the effect of hearing aid dynamic range compression (DRC) on SI was investigated in simulated real-world acoustic conditions. A spatialized version of the Danish Hearing In Noise Test (HINT) was employed inside a loudspeaker-based virtual sound environment (VSE) to present spatialized target speech in background noise consisting of either spatial recordings of two real-world sound scenarios or quadraphonic, artificial speech-shaped noise (SSN). Unaided performance was compared with results obtained with a basic HA simulator employing fast-acting DRC. Speech reception thresholds (SRTs) with and without DRC were found to be significantly higher in the conditions with real-world background noise than in the condition with artificial SSN. Improvements in SRTs caused by the HA were only significant in conditions with real-world background noise and were found to be related to differences in the output signal-to-noise ratio of the HA signal processing between the real-world versus artificial conditions. The results may be valuable for the design, development and evaluation of HA signal processing strategies in realistic, but controlled, acoustic settings.

[a] This chapter is based on Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (submitted); The effect of hearing aid dynamic range compression on speech intelligibility in a realistic virtual sound environment.

4.1 Introduction

Hearing aids (HA) attempt to restore hearing-impaired (HI) people's ability to reliably explore their auditory world. The usage of modern digital HAs has been shown to improve a wearer's hearing ability in complex real-world environments (Noble and Gatehouse, 2006). However, the introduction of more sophisticated HA signal processing algorithms over the past decades has not led to a substantial increase in HA user satisfaction (Kochkin, 2002). While a satisfaction rating depends on many factors, such as ease of use and wearing comfort, improving speech intelligibility (SI) in noise remains one of the core purposes of a HA. However, HAs have failed to provide a consistent SI benefit across users (Kochkin, 2002). This may be partially due to the focus of current HA fitting procedures on restoring audibility, rather than addressing supra-threshold distortions which HI listeners commonly experience when listening to speech in noisy situations (e.g. Sanchez-Lopez et al., 2019). In addition, signal processing algorithms in HAs have mostly been optimized for SI using speech-recognition-in-noise metrics, such as speech reception thresholds (SRTs), obtained with artificial acoustic stimuli, which may not correlate well with HA satisfaction in the listeners' real-world experience (Bentler et al., 1993; Cord et al., 2007; Wu, 2010). Therefore, it may be worthwhile to explore SI in more realistic, ecologically valid ways, both in unaided conditions as well as in conditions aided by the HA.

Various studies have investigated the impact of HA processing on SI, widely varying in scope and methodology. A common approach has been to combine speech-shaped noise (SSN) or some type of babble noise as a masker with anechoic speech sentences as the target, both presented over headphones, whereby the recordings were pre-processed to simulate the effect of HA amplification (Hunt et al., 2019; Jirsa and Norris, 1982; Saunders and Kates, 1997; Souza et al., 2015). Reverberant properties of both the background noise and the target speech, as well as effects of spatial source separation have often not been considered. In addition, head movements have largely been ignored in both the static playback of the stimuli and the HA processing. A few studies focused on the realism of the acoustic conditions and presented the noise and target speech stimuli over a spatially distributed set of loudspeakers, allowing for the use of physical HAs, either as a fitted commercial HA (Köbler and Rosenhall, 2002; Moore et al., 1985; Wouters et al., 1999), the participant's own HA

(Best et al., 2015; Oreinos and Buchholz, 2016) or a fully-controlled, real-time "master" HA (Hendrikse et al., 2020). In most of these studies, the small number of loudspeakers and the involved playback methods did not allow for a realistic reproduction of a real-world spatial sound field whereas Best et al. (2015) and Oreinos and Buchholz (2016) employed so-called virtual sound environments (VSEs), using a large number of loudspeakers to accurately reproduce real-world environments. Best et al. (2015) used a parametric room acoustic simulation technique to reproduce the spatialized sound field, while Oreinos and Buchholz (2016) employed recordings made in a room populated with simulated talkers, using higher-order Ambisonics (Daniel, 2000). Both studies still relied on simulated interferers in the noise masker, constructed using a number of anechoic speech samples placed in a room or room model, rather than in-situ recordings of real scenes.

In the present study, the effects of HA processing, specifically dynamic range compression (DRC), on speech intelligibility were evaluated in realistic acoustic conditions representing and based on recordings of real-world environments. Best et al. (2015) found that HAs provided a greater benefit to SI in their simulated VSEs compared to more artificial masker types. Here, the effect of HA processing was verified inside recorded VSE using Ambisonic auralizations of spatial recordings (Mansour et al., 2019). Two real-world scenes were recorded with a spherical microphone array using higher-order Ambisonics and reproduced inside a 64-loudspeaker loudspeaker array, providing the background noise masker for the VSE. An artificial, quadraphonic SSN stimulus matched to the long-term average spectrum of one of the real-world scenes was used as a reference condition. Target sentences from the Danish Hearing in Noise Test (Nielsen and Dau, 2011) were convolved with a room impulse response (RIR) recorded in the respective real-world scenes, obtaining spatialized target speech material for the SI task. A master HA was used to ensure consistency in HA processing across listeners and to relate the potential SI benefits caused by the HA to instrumental HA improvements. Ten listeners with symmetric mild-to-moderate hearing loss carried out the adaptive SI task wearing a HA shells controlled by the master HA, which implemented DRC based on the NAL-NL2 fitting rationale (Keidser et al., 2011). The listeners also completed the SI task using an "unaided" reference strategy, i.e. without wearing a HA. An instrumental HA analysis was conducted to relate the HA's input-output signal-to-noise ratio (SNR) performance to the SI results.

4.2 Methods

4.2.1 Virtual sound environment and spatial noise maskers

The Audiovisual Immersion Lab (AVIL) at the Technical University of Denmark, comprised of a fully spherical loudspeaker array mounted inside an anechoic chamber, provided the VSE which was used to play back the masker and target speech stimuli. The array consists of 64 loudspeakers (KEF LS50, KEF Audio, United Kingdom) at a distance of 2.4 m to a chair positioned in the center, with 24 loudspeakers separated by 15^o in the horizontal ring (at 0^o elevation).

Three spatial background noise conditions were considered in the experiment. The first condition was based on a one-minute-long spatial recording of a real-world office meeting, obtained with a 32-channel microphone array (em32 Eigenmike, mh acoustics LLC, USA). In the office meeting scenario, 12 normal-hearing participants conversed in pairs around a square conference table (2.4 m long) inside a conference room. The microphone array was placed at head level in one of the seats around the table. Fourth-order Ambisonic encoding and decoding steps were applied to the recording to match the geometry of the loudspeaker array for playback. The resulting signal was calibrated to the average broadband sound pressure level (SPL) of 73.5 dB observed in the original recording using the left ear of a head-and-torso simulator (HATS, B&K Type 4128, Brüel & Kjær A/S, Denmark) placed on the chair in the center, i.e. the sweet spot of the Ambisonic sound field.

The second background noise condition was constructed using a one-minute-long spatial recording of a public lunch. In this scenario, the 12 participants were seated around a rectangular lunch table (1 m in diameter) inside a large corporate restaurant. All recording and processing steps were the same as for the first masker, with the final signal calibrated to 75.5 dB SPL as the average level measured during recording.

The third background noise condition included SSN, matched to the long-term average spectrum (LTAS) of the public lunch recording. The SSN maskers were obtained by first recording the 64-channel public lunch masker at the left and right ear of the HATS placed in the center of the array. Then, the LTAS of the left- and the right-ear recordings were computed separately using frames of 64 ms, Hann-windowing with 50% overlap, and smoothed over 1/3rd octave bands using a normalized Gaussian kernel. A white noise signal was then filtered using a linear-phase finite impulse response filter (FIR), matched to the magnitude

spectrum of the LTAS. The resulting SSN signals were band-pass filtered between 20 Hz and 20 kHz using a 4th order Butterworth filter. Uncorrelated versions of the SSN masker derived from the left LTAS were played back over two loudspeakers in the horizontal ring at 45º and 135º azimuth, while the right, uncorrelated SSN maskers were played over the loudspeakers at 225º and 315º azimuth, creating a quadraphonic loudspeaker setup. The final, 4-channel SSN masker was calibrated in the center of the array to the same broadband level of 75.5 dB SPL as in case of the public lunch masker.

4.2.2 Listeners

Ten hearing-impaired (HI) listeners, nine male and one female, participated in the experiment. They were between 62 and 84 years old with a median age of 71.5. All listeners had a symmetric sensorineural hearing loss not exceeding an N3/S1 hearing loss category (Bisgaard et al., 2010) and showed word discrimination scores higher than 90%. The listeners were seated in a chair in the center of the AVIL loudspeaker array and their SI was evaluated for the unaided hearing strategy and the aided hearing strategy detailed below, in each of the three background noise conditions. The resulting 9 trials were randomized across listeners according to a balanced 9-by-9 Latin square design (Bradley, 1958) with one random completion. The SI scoring was carried out by a Danish audiologist, while the master HA was monitored to ensure its proper processing of the input signal without delay or feedback. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

4.2.3 Speech intelligibility task

Using the spatialized noise maskers, a speech intelligibility (SI) task was designed based on the Danish Hearing in Noise Test (HINT, Nielsen and Dau, 2011). The anechoic target speech sentences, on average 1.5 s long, were convolved with RIRs measured in the office meeting and the public lunch scenario between a loudspeaker and the microphone array placed on opposite sides of the table. The sentences were then calibrated at the left ear of the HATS placed in the center of the loudspeaker array. SRTs were determined using an adaptive, 1-up-1-down procedure, presenting the noise maskers at their constant broadband levels of 73.5 and 75.5 dB SPL for the office meeting and public

lunch recordings, and varying the level of the spatialized target sentences. The artificial SSN condition presented the noise masker at the same broadband level of 75.5 dB SPL as the public lunch masker from which it was derived, and used the target speech sentences that were convolved with the public lunch RIR.

4.2.4 Real-time hearing aid signal processing

To ensure that the HA signal processing operated consistently across listeners, a fully controlled, real-time HA system was implemented. Two HA shells (Signia, WSAudiology, Germany), each containing a microphone and a receiver, were connected via a custom preamplifier box and a sound card (RME Fireface 800, Audio AG, Germany) to a laptop. The HA signals were processed by the open-MasterHearingAid framework (openMHA, Herzke et al., 2017), encapsulated in a MATLAB control layer.

HA algorithms

The basic building blocks of the HA signal processing chain consisted of input- and output-level equalization with clipping protection, as well as a filter bank decomposition with windowing and gain application. The input signal was sampled at a rate of 44.1 kHz in time windows of 64 samples. An 65-tap finite impulse response (FIR) input equalization filter was applied to flatten the calibrated HA microphones' frequency response, while an output FIR filter was necessary to ensure that the frequency-dependent effect of the ear canal was compensated for in the calibrated HA receivers. The filter bank used an overlap-add strategy to process frames decomposed by a fast Fourier Transform at a length of 512 time-domain samples, windowed by a 256-sample Hanning window with 50% overlap. Each frame was decomposed in the frequency domain into 9 rectangular, $\frac{3}{4}$-octave-wide, non-overlapping bands with the lowest center frequency at 177 Hz (177, 297, 500, 841, 1414, 2378, 4000, 6727 and 11314 Hz).

The HA amplification employed DRC based on the listener's pure tone audiogram thresholds and following the gain prescription of the National Acoustics Laboratory-Non Linear 2 (NAL-NL2) fitting rationale (Keidser et al., 2011) (see Appendix A for more details). The attack and release time constants for the DRC were set to 5 ms and 100 ms, respectively. These values correspond to a fast-acting compression scheme, where the output gain is adjusted relatively quickly after a change in input level, both at the onset and end of the level change. Such

a configuration aims at restoring audibility on short time scales corresponding to syllables or phonemes.[b] Other features, such as feedback cancellation, noise reduction and beamforming, common in commercially available hearing aids, were not included in the HA processing. This was done to focus on the essential components in the processing and exclude potentially confounding adaptive algorithms. The omission of feedback cancellation implied that the HAs needed to be equipped with a fully closed dome and were limited in their gain prescription to at most a N3 or S1 hearing loss category (Bisgaard et al., 2010), or about a 40 dB hearing loss at 1 kHz.

HA implementation, fitting and evaluation

The HA processing chain was implemented in the openMHA framework. This plugin-based open source platform includes several basic HA features as well as HA calibration and validation tools and can interface in real-time with input-output sound card channels using a desired sample rate and frame size (see Appendix C for more details).

After the calibration of the HA microphones and receivers (see Appendix D for more details), a listener-specific validation routine was carried out, verifying that the HA receiver output levels matched the target gains of the NAL-NL2 rationale at the 65 dB SPL DRC knee point at each filterbank center frequency. The HAs were placed on the HATS, positioned on the chair in the center of the loudspeaker array and the HA receiver calibration values were fine-tuned until the measured gains at the HATS microphones deviated by less than 1 dB from the target gains. The processing delay between the sound card microphone input channels and receiver output channels was determined by feeding a test signal into each of the input channels and tracking the time it took for the signal to be processed by an openMHA instance and reach a receiver output channel. On average, the input-output delay amounted to 11 ms, which is more than the 5 ms delay commonly targeted in commercial HAs yet still considerable lower than the 20 ms limit tolerable to individuals with mild-to-moderate hearing loss (Stone and Moore, 1999).

[b] To limit the maximum output level of the receiver signal, an additional ultra-fast-acting compressor was applied to the output signal, with attack and release times of 2 ms and 5 ms respectively. This soft clipping protection mechanism compressed the signal at a slope of 0.5 when broadband levels of 0.8 or higher relative to the digital maximum were detected (Herzke et al., 2017).

4.2.5 Instrumental HA analysis

In addition to the SI assessment, an instrumental HA analysis was conducted to evaluate how the HA DRC processing affected the broadband SNR of target HINT sentences in the three noise conditions. With the HAs placed on the HATS in the center of the loudspeaker array, HA microphone recordings were made of 20-second excerpts of the two real-world background noises and the SSN as well as 10 randomly selected HINT target sentences. Each sentence was superimposed onto 10 different noise clips within each background noise excerpt, which was set to its respective broadband level. The sentences were scaled to achieve the desired SNR. Using the method proposed by Hagerman and Olofsson (2004), this process was then repeated while all noise clips were shifted in phase by 180°. For each speech-in-noise mixture, the in-phase and out-of-phase versions were then processed by an offline, file-based openMHA instance, configured to provide amplification with DRC according to the NAL-NL2 rationale fitted to the mean audiogram across all listeners. By adding or subtracting the output in-phase and out-of-phase mixtures, the separate speech and noise components can be recovered perfectly, assuming that the HA has a linear phase response. In this way, 100 speech-in-noise clips were analyzed per target SNR value, across a range from -12 dB to 12 dB (the approximate range of SNRs presented in the SI task), in 3 dB increments.

4.3 Results

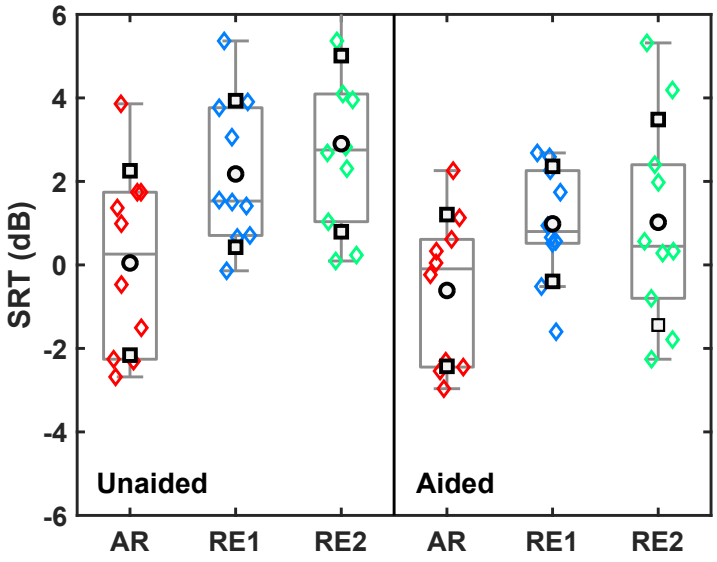
Figure 4.1A shows the measured SRTs across listeners for the unaided and the aided hearing strategies. Individual thresholds are plotted for the SSN (AR, red circles), the real-world office meeting noise (RE1, blue diamonds) and public lunch noise (RE2, green squares), as well as the median and 25th/75th percentiles of the SRT distributions (box plots). The mean values (black circles) and standard deviations (black squares) are also shown. A large variability of the SRTs across listeners can be observed in all conditions and for both processing strategies. For the unaided strategy, mean SRTs were obtained at 0.1 dB in the artificial SSN, at 2.2 dB in the office meeting noise and at 2.8 dB in the public lunch noise. Similarly for the aided strategy, the mean SRTs where -0.6 dB, 1.0 dB and 1.0 dB in the AR noise, RE1 noise and RE2 noise, respectively. The standard deviations for the unaided strategy amounted to 2.2 dB in the AR

condition, 1.8 dB in the RE1 condition and 2.1 dB in the RE2 condition. For the aided strategy, the standard deviations were 1.8 dB, 1.4 dB and 2.5 dB in the AR condition, RE1 condition and RE2 condition, respectively.

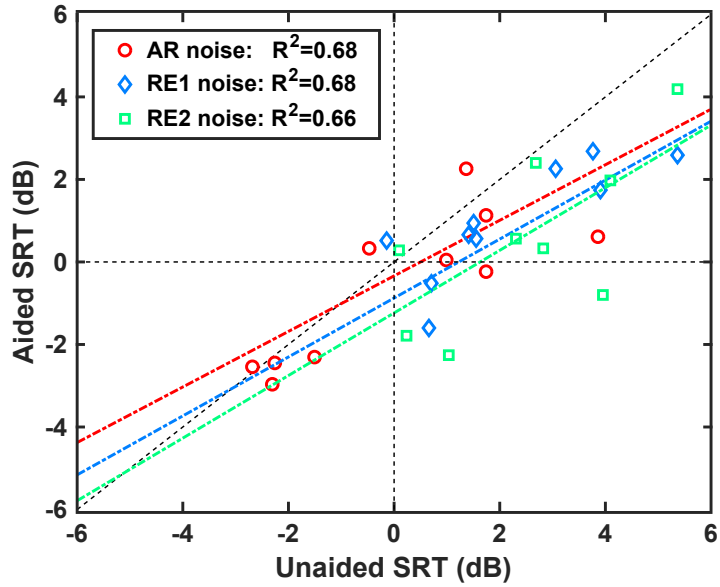
A two-way repeated-measures analysis of variance (RANOVA) revealed a significant effect of noise type ($F(1,9) = 23.2, p = 0.0001$) and hearing strategy ($F(1,9) = 47.7, p \leq 0.0001$). Pair-wise comparisons between noise conditions showed that SRTs were significantly higher for the RE1 and RE2 condition than for the AR condition, both for the unaided and the aided strategy. Comparisons between the RE1 and RE2 conditions did not reveal a significant difference. Conversely, there was a significant decrease in SRTs between the unaided and the aided hearing strategy for the RE1 and RE2 conditions, but no significant difference for the AR condition. SI performance within each noise condition was not significantly different between listeners. Thus, the realistic noises made it consistently more difficult to understand speech compared to the artificial noise, regardless of whether HAs were used or not. At the same time, the artificial noise did not show a significant effect of the HA processing, while both realistic noise types did.

Figure 4.1B illustrates the unaided versus aided SRTs for each of the three noise conditions AR, RE1 and RE2, as well as their linear least squares fits, with respective R^2 correlation factors of 0.68, 0.68 and 0.66. This indicates that the aided strategy was strongly correlated with the unaided strategy across all noise types, suggesting that the effect of the HA DRC processing depended on the unaided SRT. Most data points lie below the 45° line and all least squares fits have slopes below one, demonstrating the increasingly beneficial effect of the HA DRC processing on SI with increasing values for the SRT (i.e. decreasing SI) in the unaided condition. This effect was stronger for the RE1 and RE2 conditions, for which the HA started to provide a benefit at lower unaided SRTs compared to the AR condition.

Table 4.1 shows the R^2 correlation factors between the listeners' SRTs and their the four-frequency-average hearing loss (4FAHL). Correlations were weak across all noise types and weakest for the realistic noise types, and there was virtually no difference between the unaided and the aided hearing strategy. Thus, the 4FAHL was a poor predictor of supra-threshold SI performance in the conditions considered in the present study, especially in realistic VSEs, regardless of whether a HA DRC processing strategy was active or not.



(A) Speech reception thresholds



(B) Unaided vs. aided SRTs

Figure 4.1: Panel A: Measured speech reception thresholds (SRTs) for the unaided and the aided hearing strategies, in the SSN (AR), the real-world office meeting noise (RE1) and public lunch noise (RE2) conditions. The mean values (black circles) and standard deviations (black squares) are also shown. Panel B: Unaided vs. aided SRTs and R^2 correlation factors in the three noise conditions; AR, RE1 and RE2. The least-squares fits to each noise conditions are shown as dashed lines.

Table 4.1: Unaided and aided R^2 correlation factors between the average four-frequency-average hearing loss (4FAHL) across participants and their speech reception thresholds for the three noise types.

R^2	Unaided	Aided
AR	0.27	0.26
RE1	0.15	0.19
RE2	0.09	0.07

Figure 4.2 shows the output SNR distributions for the AR, RE1 and RE2 conditions at nine input SNRs between -12 dB and 12 dB, obtained from the instrumental HA analysis. The AR noise produced considerably lower SNR values (red symbols) than the RE1 (blue) and RE2 conditions (green), for all input SNRs. For the RE1 and RE2 conditions, the HA DRC processing increased the output SNR of the provided speech-in-noise mixture by up to 2 dB at negative input SNRs. The effect decreased with increasing input SNR and at input SNRs beyond around 5 dB, the output SNR became smaller than the input SNR. For the AR condition, the HA DRC processing decreased the output SNR regardless of the input SNR, but again most strongly at positive input SNRs. The HA processing in the AR condition started to provide a positive median output SNR at an input SNR of 4.1 dB, while this happened at input SNRs of -1 dB and -1.8 dB, respectively, for the RE1 and RE2 conditions. The spread in the SNR output distribution was markedly larger in the AR and RE2 conditions relative to that in the RE1 condition.

Figure 4.3A displays the histograms of the noise levels (in dB) estimated for a ten-second excerpt of the AR (red), RE1 (blue) and RE2 (green) interferers as recorded by the left front HA microphone and normalized to unit average power. Each level estimate was obtained by calculating the average power over a time window of 5 ms, corresponding to the analysis window of the HA DRC processing. The histogram bin width of 1 dB corresponds to the resolution of the compression lookup table inside the HA. These histograms show that the RE1 and RE2 noise types exhibit considerably greater amplitude fluctuations over time than the stationary AR noise. Similarly, Fig. 4.3B shows the corresponding histograms of the speech levels across the ten HINT sentences used in the instrumental SNR analysis, calculated in the same way as in the case of the noise histograms. The histogram for the anechoic speech is shown in black, together with the histogram for the RE1 (in blue) and the AR/RE2 (in green) speech, as recorded by the left front HA microphone and normalized to unit average power.

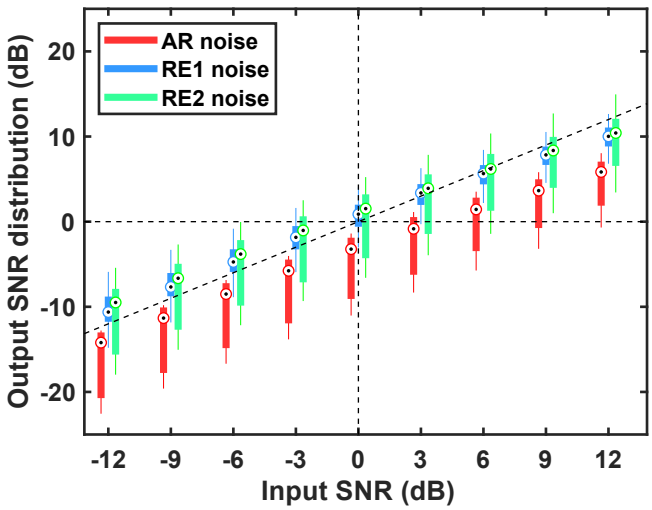
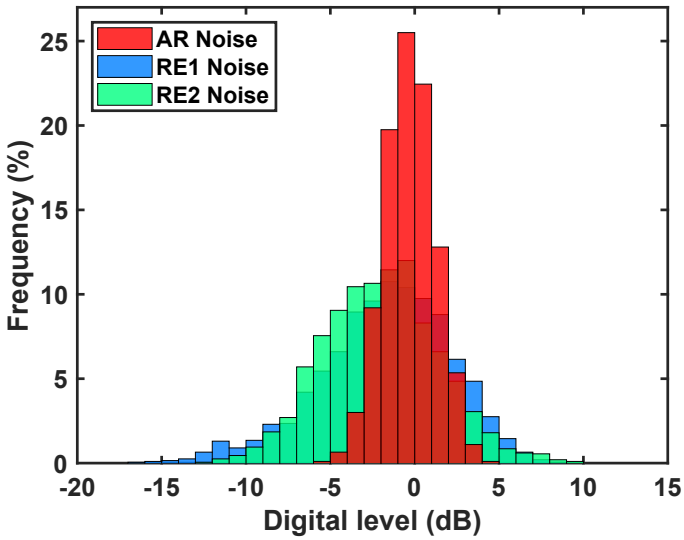
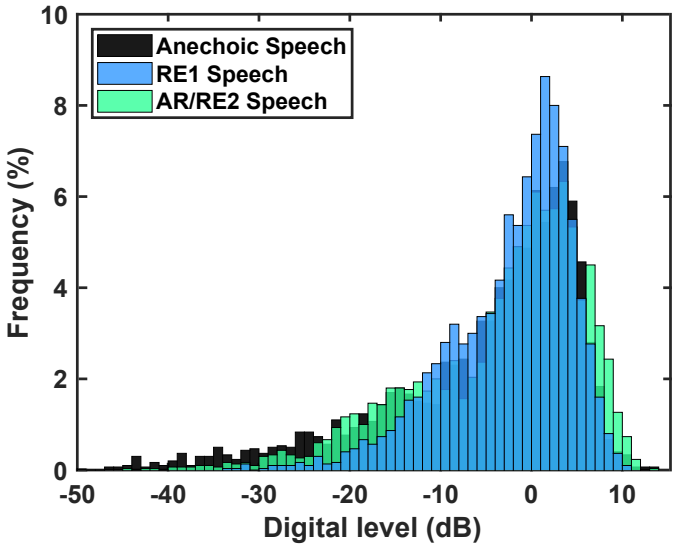


Figure 4.2: Output SNR distributions vs. input SNR values resulting from the instrumental HA analysis for 10 target speech sentences superimposed onto 10 noise fragments at nine input SNRs between -12 dB and 12 dB, for the artificial noise type (AR) and the two realistic noise types (RE1 and RE2).

The histograms obtained for the anechoic and the AR/RE2 speech were similar in width whereas the RE1 histogram showed a narrower distribution. This was expected due to the much lower direct-to-reverberant ratio (DRR) in the office meeting scenario (RE, 6.6 dB), introducing more reverberation into the anechoic HINT sentences than in the public lunch scenario (RE2, 16.6 dB). The more reverberant RIR of the office meeting reduced the level fluctuations in the target speech signals, thereby also reducing the width of the RE1 SNR distribution (see Fig. 4.2) compared to the AR and RE2 SNR distributions.



(A) Noise level histograms



(B) Speech level histograms

Figure 4.3: Panel A: Histograms of the noise levels (in dB) estimated for a ten-second segment of the AR (red), RE1 (blue) and RE2 (green) interferers, normalized to unit average power. Each level estimate was obtained by calculating the average power over a time window of 5 ms. Panel B: Corresponding histograms of the speech levels across the ten HINT sentences used in the instrumental SNR analysis, calculated in the same way as in the case of the noise histograms. The anechoic speech histogram is shown (black), alongside the RE1 (blue) and AR/RE2 (green) speech histograms, normalized to unit average power. The analysis window was the same as for the noise signals.

4.4 Discussion

The main aim of this study was to investigate the impact of applying in-situ background noise recordings as maskers in a speech intelligibility task conducted with and without hearing aids. Consistently with previous work (Best et al., 2015; Mansour et al., 2019), employing realistic noise maskers resulted in increased SRTs compared to artificial noise maskers. This reduction in SI in realistic backgrounds has been linked to energetic speech-on-speech masking in the spectral and modulation-spectral domains (Brungart et al., 2006; Jørgensen and Dau, 2011) as well as the presence of intelligible interferers in the realistic background (Westermann and Buchholz, 2015). In the present study, the LTAS of the AR noise was matched to that of the RE2 noise, such that the differences in the SRTs obtained in these conditions could not be caused by differences in the (long-term) spectral properties of these interferers.

At the same time, SRTs did not differ between the RE1 and RE2 conditions. Despite room acoustic and spectral differences between the two realistic scenes, the stimuli were presented at rather similar broadband SPLs (73.5 dB and 75.5 dB) and contained the same number of nearby interfering talkers (ten). This suggests that SI in realistic VSEs might be more strongly influenced by the overall background level and the number of interfering talkers, which have been found to be predictors of perceived scene complexity (Weisser et al., 2019a), than by the acoustic details of the background noise. The results obtained across noise types were similar between the aided and unaided strategies, indicating that the realistic VSEs were robust to the HA processing employed here.

The applied HA processing led to improved SI in the realistic conditions, but not in the AR condition. Several factors could have contributed to this finding. First, hearing-impaired listeners that have their audibility (partially) restored through HA amplification may regain the ability to employ dip listening, which may be more beneficial in modulated interferers (such as the realistic noises in the present study) than in stationary noise interferers (Peters et al., 1998).

Second, as revealed by the instrumental analysis, the HA DRC processing affected the different noise types differently. Due to the stationarity of the AR noise type, the applied compression resulted in a reduction in the output SNR of the HA across the entire input SNR range (see Fig. 4.2). In contrast, due to the wider dynamic range in the RE1 and RE2 noise types, the compressive processing was able to reduce the peak noise energy to a greater degree (at the

same overall SPL), resulting in an increase in the median output SNR at low input SNRs (up to about 3 dB). At highly positive SNRs, this effect disappeared (as can be seen in Fig. 2, above around 6 dB) because the speech signal, shared between all noise types, now determined the envelope of the mixture and thereby the impact of compression (Rhebergen et al., 2009). Unaided SRTs in the RE1 and RE2 conditions, however, fell in the range where the HA DRC processing still resulted in an SNR improvement for most segments (roughly between 0 dB and 5 dB, see the ordinate of Fig. 4.1B), which likely contributed to the HA benefit observed in these conditions. Previous work (Boike and Souza, 2000; Moore et al., 1999) also demonstrated the benefit of fast-acting DRC on the SNR of noise-dominated speech in modulated (artificial) maskers versus stationary maskers, as well as its beneficial effect on SI (Kowalewski et al., 2018; Rhebergen et al., 2009). Thus, the presence or lack of benefit of compressive HA processing on SI, both in terms of instrumental and perceptual measures, depends strongly on the range within which SRTs are obtained, which, in turn, depends on the properties of the speech and noise stimuli presented to the HA and the listener.

The results of the present study suggest that a realistic VSE can provide an effective setting for evaluating the impact of HA signal processing algorithms on a listener's SI performance and for relating that performance to instrumental HA performance metrics. The present approach, however, did not consider visual information, which is known to greatly improve speech intelligibility and subsequently shift unaided SRTs ranges downward, especially at low SNRs (Sumby and Pollack, 1954). It is likely that the lack of visual cues also affected the SRTs of the listeners in the current experiment, and by extension the aided benefit shown by the HA processing. Thus, an important next step would be to establish measurements of visually aided speech intelligibility in realistic VSEs.

The results of the instrumental analysis showed that, with a shift to lower SNRs, the HA DRC processing continued to improve the output SNR in the realistic scenes, implying that a visually aided SI assessment would continue to reveal an increased HA benefit for realistic versus artificial scenes. Furthermore, only a simple HA was implemented in this study, excluding commonly used features such as noise reduction and beamforming. The methodology proposed here can, however, be applied to HA simulators including any number of features. Lastly, it may be informative to consider psychoacoustic outcome measures beyond SI, such as localization or loudness perception in realistic VSEs.

Combining aided psychoacoustic measures with instrumental HA analyses may provide a better understanding of how various signal processing strategies perform in real-life environments.

4.5 Conclusion

In this study, the effect of HA DRC processing on SI was investigated in two realistic acoustic scenes, constructed using spatial background recordings and anechoic speech samples convolved with RIRs measured in-situ. It was found that both unaided and aided SRTs were significantly increased inside both realistic VSEs compared to a reference condition employing quadraphonic SSN as a masker. This is consistent with previous studies, which showed increased spectral and modulation-spectral energetic masking for realistic noises as well as a detrimental effect of intelligible speech in the maskers on SI (Best et al., 2015; Mansour et al., 2019). Despite the acoustic differences, SI was similar between the two realistic noise types, suggesting that precise acoustic details in the reproduced environment may be less important than its overall loudness and number of interfering sources.

HA DRC processing was found to provide a benefit to SI in the realistic scenes, but not in the artificial reference condition, which was consistent with output SNR differences observed between the realistic versus reference conditions. Results of an instrumental SNR analysis revealed that the stationary nature of the artificial noise led to consistently lower median SNRs at the output of the HA compared to the two realistic background noises.

The results of this study illustrate the relevance of evaluating the impact of HAs on SI in experimental conditions that are realistic, such that the SI task might produce SRTs that reflect real-life experience. By achieving this, HA processing strategies may then be tailored to SNRs that occur in real-world conditions and consequently become better matched to the user's every-day experiences.

5

Guided ecological momentary assessment in real and virtual sound environments[a]

Abstract

Ecological momentary assessment (EMA) outcome measures can relate people's subjective auditory experience to their objective acoustical reality. While highly realistic, EMA data often contain considerable variability, such that it can be difficult to interpret the results with respect to differences in people's hearing ability. To address this challenge, a method for "guided" EMA is proposed and evaluated. Accompanied and instructed by a guide, normal-hearing participants carried out specific passive and active listening tasks inside a real-world public lunch scenario and answered EMA questionnaires related to aspects of spatial hearing, listening ability, quality and effort. In-situ speech and background noise levels were tracked, allowing the guided EMA task to be repeated inside two acoustically matched, loudspeaker-based laboratory environments: a 64-channel virtual sound environment (VSE) and a three-channel audiology clinic setup. Results showed that guided EMA provided consistent passive listening assessments across participants and conditions. During active listening, the clinic setup was found to be less challenging than the real-world and the VSE conditions. The proposed guided EMA approach may provide more focused real-world assessments and can be applied in realistic laboratory settings to aid the development of ecologically valid hearing testing.

[a] This chapter is based on Mansour, N., Westermann, A., Marschall, M., May, T., Dau, T., and Buchholz, J. (submitted); Guided ecological momentary assessment in real and virtual sound environments.

5.1 Introduction

Relating the subjective auditory experience of an individual person in complex, real-world environments to objective measures of hearing ability has been of interest within hearing research for many years. Most studies have focused on evaluating psychoacoustic measures like loudness perception, spatial awareness, speech intelligibility and localization ability using well-controlled, yet artificial stimuli. Efforts have been made to increase the ecological validity (Reis and Judd, 2000) of these stimuli by reproducing real-world environments in the laboratory (e.g. Ahrens et al., 2017; Best et al., 2015; Mansour et al., 2019; Westermann and Buchholz, 2015). However, it is unclear how the task paradigms in such studies and their corresponding outcome measures, though reliable and reproducible, can be related to subjective hearing ability in the real world (Lutman, 1991; Timmer et al., 2015). Retrospective questionnaires, like the Speech, Spatial and Qualities of Hearing Scale (SSQ, Gatehouse and Noble, 2004) or the Glasgow Hearing Aid Benefit Profile (GHABP, Gatehouse, 1999), were developed specifically to quantify subjective hearing ability. While their responses can correlate well with objective measures (e.g. better-ear average hearing thresholds, Gatehouse and Noble, 2004), providing so-called construct validity, responses are often affected by recall bias (Moskowitz and Young, 2006), limiting their reliability.

To overcome these issues, the methodology of ecological momentary assessment (EMA) has emerged as an approach in which questionnaires are employed to capture people's subjective environmental impressions at frequent (chosen or triggered) intervals over an extended period of time in their every-day life (Shiffman et al., 2008). Effects of recall bias can largely be avoided since participants evaluate their surroundings while they are observing them and instrumental measures (e.g. background noise levels and frequency spectra) can be applied to acoustically characterize the in-situ environment. Several recent studies have applied the methodology of EMA to hearing research, establishing its reliability and construct validity (Galvez et al., 2012; Henry et al., 2012; Timmer et al., 2017; Wu et al., 2015).

However, drawbacks to the successful use of EMA remain. As reported in Timmer et al. (2017), there are potential issues of compliance (the participant's willingness to complete the assessments), feasibility (the extent to which the participant can fulfill the EMA task requirements), burden (the demand

placed on the participant) and data variability (the large inter-subject variability contained within EMA data sets). The variability in EMA data has two main sources: temporal, or intra-participant variability, i.e. the spread within a single participant's questionnaire responses over time due to the changing environment and circumstances, and inter-participant variability, i.e. the spread across different participants' responses due to differences in their every-day environment and their hearing ability. While EMA data in their traditional form are useful for characterizing experienced trends and differences across large groups of people and data sets, an approach with reduced intra- and inter-participant variability may be beneficial to explore measures of individual hearing ability.

In this exploratory study, a "guided" approach to EMA is proposed. During the experiment, which took place at a fixed time of the day and over a short time interval, each participant visited the same, predetermined real-world (RW) scene, accompanied by a human guide. The location and time-of-day constraints served to reduce the inter-participant variability in the EMA data, while the duration constraint aimed to reduce the intra-participant variability caused by environmental changes. The guide facilitated the participant's EMAs, attempting to improve compliance and feasibility and reduce burden, by providing a clearly structured passive listening task, a communication task and an active listening task. Each task was followed by a brief questionnaire on commonly addressed hearing topics in EMA research. The passive listening task required the participant to simply listen to their environment for one minute, while the conversation task consisted of one minute of natural conversation between the participant and the guide. During the active listening task, the participant was asked to listen to two monologues told by the guide, the first held at natural speech levels, established during the communication task, and the second at challenging speech levels, subjectively determined by the guide. These tasks were designed to mimic common aspects of unguided EMAs in a more controlled fashion, while the inclusion of the active listening task at challenging speech levels was intended to avoid potential ceiling effects.

Due to its rigid structure and short duration, the guided EMA method could be applied inside realistic laboratory environments, represented by a virtual sound environment (VSE) in the form of a spherical 64-loudspeaker array and a three-loudspeaker (front-left-right) clinic environment (CL), a simple yet common setup in audiology clinics. Both laboratory environments employed Ambisonic renderings of spatial audio recordings made in the RW environment

to represent the background noise, as well as prerecorded audiovisual (AV) clips to visually represent the guide's speech during the active listening tasks. The use of these prerecorded stimuli was intended to further reduce the inter- and intra-participant variability caused by acoustic changes in the speech and noise signals in the RW environment.

The background noise levels and the guide's speech levels occurring in the RW environment were matched inside the lab using an in-situ signal-to-noise ratio (SNR) estimation method. This matching technique aimed at reducing the intra-subject variability in the EMA data between the RW environment and the laboratory environments. Both the use of guided EMA and the level matching were undertaken to compare the participants' laboratory EMA data with their real-world EMA data and to investigate the consistency of the results across participants and conditions. It was hypothesized that the EMA results for both the passive and the active listening tasks would be consistent (i.e. have a low variability) between participants in the RW condition, and that the RW EMA results would be similar to the corresponding results in the VSE and CL conditions.

5.2 Methods

5.2.1 Real-world assessment

The RW assessment phase of the experiment took place inside a canteen on a university campus over lunch time. Such a public lunch scenario is known to occur commonly in people's lives and is generally rated as important and challenging to hear in (Mansour et al., 2019; Wolters et al., 2016). The public lunch scenario was characterized acoustically by a reverberation time (RT_{60}) of 2.5 s, an early decay time (EDT) of 0.2 s and a direct-to-reverberant ratio (DRR) of 7.4 dB. As shown in Fig. 5.1, a participant and the guide were seated across from each other at a table without other occupants, with a distance of 1 m between them. Several similar tables were placed in the immediate surrounding area, populated by people having lunch. As outlined in the left column of Fig. 5.2, the participant, following the instructions of the guide, completed a passive listening task, a communication task and an active listening task.

In the passive listening task, the participant listened to their surroundings for one minute without talking and then completed a 4-part EMA questionnaire using a proprietary smartphone app. The communication task consisted of

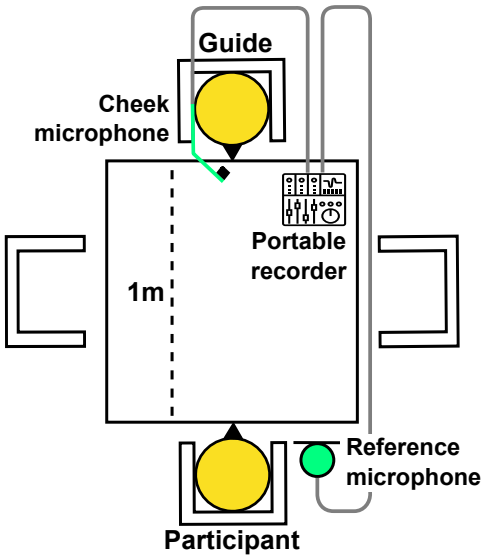


Figure 5.1: Schematic overview of the physical setup of the real-world phase in the guided EMA experiment, which took place inside a crowded university canteen over lunchtime. The participant and guide are seated across from each other at a table (yellow heads). The mobile microphone recording system is also shown (green icons).

a one-minute-long unscripted conversation between the participant and the guide, in order to determine the natural, conversational speech levels in the scenario. Finally, the participant listened to two one-minute monologues held by the guide, the first at the established conversational speech levels and the second at challenging levels that were subjectively determined by the guide. After each monologue, the participant answered another 4-part EMA questionnaire on their listening experience. The monologues were the same in content for each participant but were not scripted.

During the one-minute assessments of the passive listening, communication and active listening tasks, audio recordings (sampled at 48 kHz, 24 bit) were made with a custom-built, mobile microphone system, to obtain measurements of the broadband background levels as well as the guide's speech levels at the position of the participant. The system contained a cheek-mounted microphone and a reference microphone (both DPA 4066, DPA, Denmark), both connected to a portable audio recorder (Zoom H6 Handy Recorder, Zoom Corp, USA). The reference microphone was mounted vertically on a stand and placed at ear-height next to the participant, the same distance away from the guide. The cheek microphone was worn by the guide, who also operated the recording system.

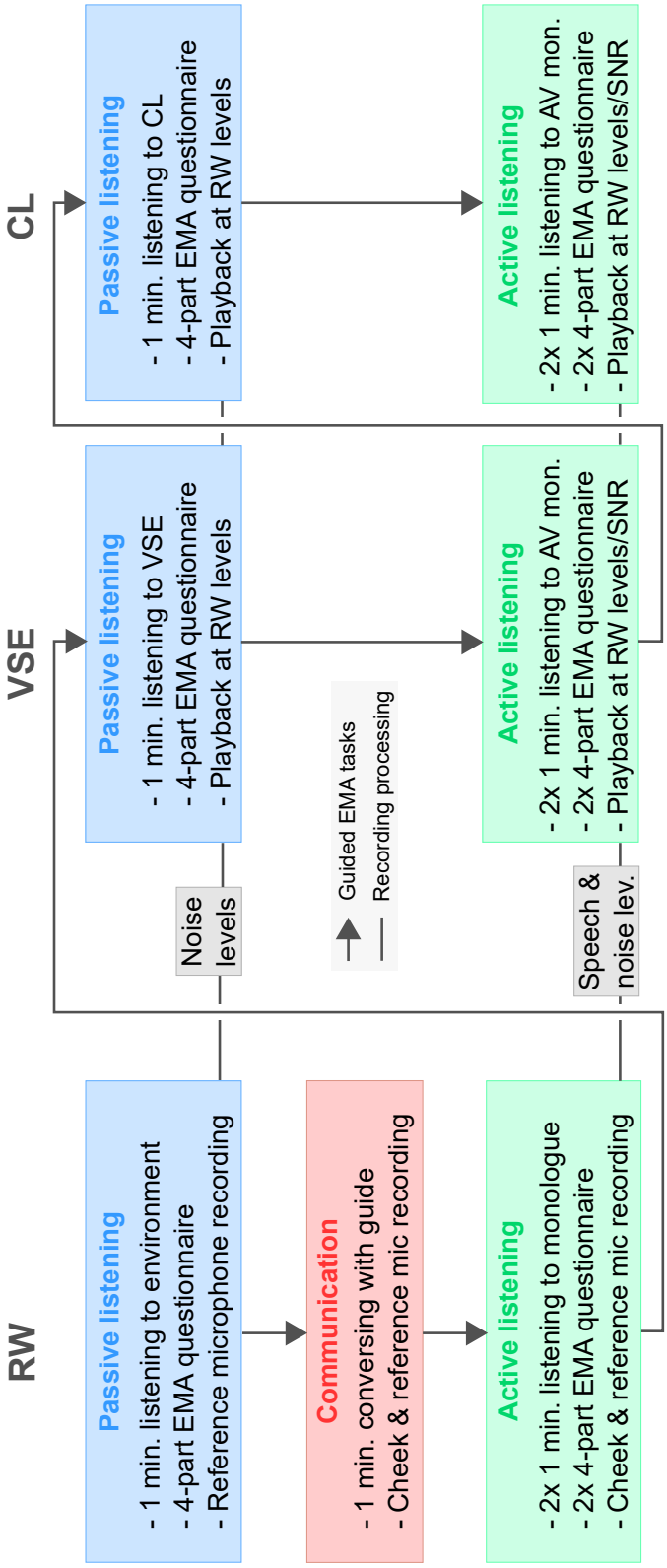


Figure 5.2: Diagram depicting the sequence of the RW, VSE and CL phases in the guided EMA experiment, as well as the passive listening, communication and active listening stages within each phase. The grey arrows indicate the succession of guided EMA tasks in the experiment. The grey lines signify the recording processing and use in the VSE and CL phases.

The system was calibrated before each experiment by recording the digital level measured in the reference microphone with a 1-kHz pure tone calibrator at a sound pressure level (SPL) of 94 dB (B&K Type 4231, Brüel & Kjaer, Denmark). To calibrate the cheek microphone, the guide spoke a brief, scripted message while wearing the microphone inside an anechoic chamber, at a distance of 1 m to the reference microphone. This allowed for the derivation of a fixed scaling factor representing the broadband, free-field decay in speech level from the mouth of the guide to the reference microphone position at 1 m.

The microphone recordings were analyzed to derive broadband-level estimates of the background noise and the speech of the guide. The overall background noise level during the passive listening tasks was determined by calculating the average power in the reference microphone's recorded signal in dB SPL. To derive the guide's speech level at the participant's position during the communication and active listening tasks, the calibrated scaling factor was applied to the speech segments in the cheek microphone recording, which had been extracted using an adaptive, energy-based voice activity detector (VAD, Kinnunen and Li, 2010). The background noise level was then computed from the reference microphone signal, using segments where the guide was not speaking (using the VAD derived from the synchronized cheek microphone signal). To obtain the background noise level during the communication task, the voice of the participant was also removed from the reference microphone signal using an additional energetic VAD. The details of the used VADs are described in Chapter 2 (Mansour et al., 2021).

5.2.2 VSE assessment

Following the RW assessment, each participant repeated the passive and active listening tasks while seated inside an anechoically enclosed, 64-channel spherical loudspeaker array. This environment represented the fully spatialized VSE condition. The loudspeaker array, with a 2.4 m radius, used a spatial reproduction of a pre-recorded background noise signal to simulate the canteen environment. The background noise signal was recorded by a 32-channel spherical microphone array (em32 Eigenmike, MH Acoustics, USA), placed at head height inside the canteen, in the position where the participant was seated during the RW assessment. This 2-minute-long recording was then encoded to a 4th order Ambisonic signal and subsequently decoded to the geometry of the loudspeaker array for the VSE condition. The participants also carried out the

tasks in the CL condition. This condition employed only two loudspeakers in the array to present the background noise, positioned at +/-90 degrees azimuth and 0 degrees elevation, mimicking a simple audiology clinic setup. The two loudspeaker signals were derived from the same Ambisonic signal by decoding them to a binaural reproduction (Weisser et al., 2019b) and subsequently applying a diffuse-field equalization step to account for the path between the loudspeakers and the ears of a head-and-torso simulator head-and-torso simulator (B&K Type 4128, Brüel & Kjær A/S, Denmark) located in the sweet spot. Both noise conditions were calibrated using the reference microphone, positioned in the acoustic sweet spot of the loudspeaker array.

The passive listening task was carried out by the participant in both the VSE and CL conditions, identical in structure to the corresponding tasks in the RW condition, as is summarized in the center and right columns of Fig. 5.2, respectively. To simulate the background noise, one-minute-long excerpts of the decoded noise signals described above were played back at the same participant-dependent broadband level as was measured in the real world. For the active listening task, the guide's monologue was simulated using pre-recorded, 1-minute-long AV recordings spoken by the guide. For each participant, four monologues on four different topics were selected randomly out of a total of 16. The monologues were filmed inside an anechoic chamber on a neutral background using a digital camera (Sony a6000, Sony, Japan), and simultaneously recorded with the reference microphone, both at a distance of 1m. The microphone recordings were calibrated and stored in both an unprocessed, single-channel format as well as a 64-channel, spatialized format obtained by convolving the single-channel signal with a spatialized room impulse response (RIR). This RIR was constructed by deriving the 64-channel Ambisonic loudpeaker signals from a spatial RIR recording, captured in quiet with the microphone array positioned at head height on the participant's chair inside the canteen. The resulting speech signals were then scaled to match the levels for the corresponding conditions in the real world and superimposed onto the background noise signal. The spatialized speech was used for the VSE condition, while the anechoic speech was presented from the frontal array loudspeaker (0° azimuth and elevation) in the CL condition. All speech signals were processed to provide 1 s of silence in the beginning and at the end.

The video recordings were played back on an 11-inch iPad positioned 1 m in front of the participant at eye level. The recordings were synchronized to the speech playback by using the video camera's own recorded speech signals to derive the delay with those of the reference microphone via cross-correlation. Since the video signals were played back using VLC media player, and not MATLAB, the additional processing delay between both signal paths was taken into account by recording the video camera audio signal played back over the iPad at the same time as the loudspeaker array audio signal and using cross-correlation to derive the constant offset between both signals.

5.2.3 EMA questionnaire design

Table 5.1 describes the EMA questionnaires for the passive and active listening stages, designed specifically for this experiment, detailing individual questions as well as their title in the smartphone app.

Table 5.1: EMA questionnaires for the passive listening and the active listening stages and their respective title in the smartphone app, as well as the 5-point Likert response scale (for Q1-Q7). Q8 was rated on a continuous scale between 0 and 100% (in 1% increments).

Passive listening stage		
No.	Title	Question
Q1	Difficulty to focus	*Is it difficult for you to focus on specific sounds in this environment?*
Q2	Pleasantness of sound	*Does this environment sound pleasant to you?*
Q3	Annoyance with sound	*Are you annoyed with certain sounds in this environment?*
Q4	Effort to relax	*Is it effortful for you to relax in this environment?*
Active listening stage		
Q5	Loudness of speech	*How loudly did you feel that the person the person talking to you was speaking?*
Q6	Listening effort	*How effortful was it to listen to the person talking to you?*
Q7	Naturalness of speech	*How naturally did you think the person was talking to you?*
Q8	Understanding of speech	*How well did you understand what the person talking to you was saying?*
Response scale for Q1-Q7		
Not at all (1) - *Not that* (2) - *Somewhat* (3) *Very* (4) - *Extremely* (5)		

The passive-stage questions focused on spatial sound perception (Q1), sound quality (Q2-Q3) and ability to relax (Q4), while the active-stage questions probed loudness perception of speech (Q5), listening effort (Q6) and quality of speech (Q7) as well as self-assessed speech understanding (Q8). These topics were chosen for their commonality in existing EMA research as well as in retrospective questionnaires like the SSQ. The response scale at the bottom of Tab. 5.1 indicates the possible responses on a 5-point Likert scale (Likert, 1932) for all questions except the final one (Q8), which was rated on a continuous scale from 0 to 100% (in 1% increments). Questions Q1 through Q7 were phrased to fit the 5-point Likert response scale in order to provide a rigid and consistent structure to the questionnaires. The continuous answer scale of Q8 was chosen to allow the derivation of a self-assessed speech understanding score, similar to speech intelligibility (SI) scores produced by SI paradigms. Because of the known RW location, no questions were needed to establish the nature of a participant's surroundings. To reduce the variability in the RW environment between the assessment stages, the number of questions in both the passive and active listening questionnaires was limited to four, ensuring that the entire RW session could be completed within 30 minutes.

5.2.4 Participants

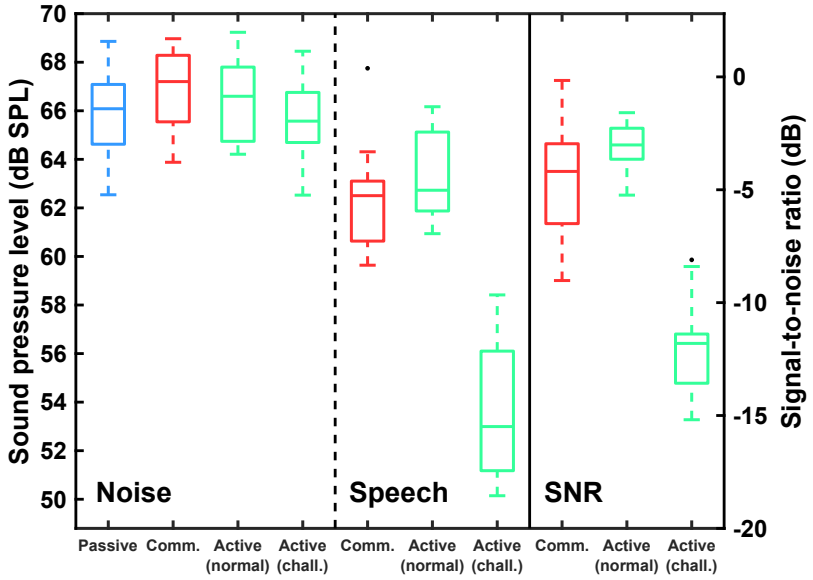
Fifteen participants with self-reported normal hearing carried out the guided EMA experiment. The participants were between 21 and 39 years old, with a median age of 25, and all had English as their native language. The participants were recruited from the general public using a web advertisement and were financially compensated for their time. All regulations and guidelines with regard to hygiene and social distancing, brought about by the COVID-19 pandemic, were adhered to. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

5.3 Results

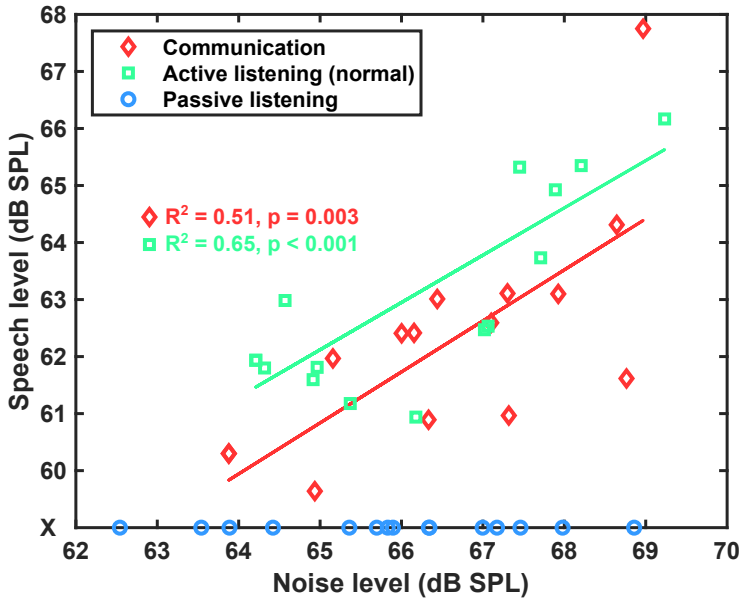
5.3.1 Real-world noise and speech levels

Figure 5.3A shows the distributions of broadband noise SPLs (left), measured inside the canteen during the RW passive (blue), communication (red) and active listening stages (green) as well as the guide's speech SPLs (middle) and SNRs (right) during the communication (red) and active listening stages (green). The left SPL ordinate indicates the noise data, while the right SNR ordinate represents the SNR data. The means (black circles) and standard deviations (black squares) of the distributions are indicated as well. The normality of each group was verified with the Anderson-Darling and Shapiro-Wilk tests. Mean noise levels were 65.8 dB SPL during the passive listening stage, 66.8 dB SPL during the communication stage and 66.4/65.6 dB SPL during the active listening stage at normal and challenging speech levels, respectively. As verified with a repeated-measures analysis-of-variance test (RANOVA), the noise distributions, though ranging from 62.5 dB SPL to 69.5 dB SPL, were not significantly different from each other ($F(3,42) = 1.15, p = 0.34$). The mean speech level during the communication stage was 62.4 dB SPL, resulting in an SNR distribution with a mean of -4.5 dB. For the active listening stages at normal and challenging speech levels, the mean speech levels were 63.3 dB SPL and 53.6 dB SPL, respectively, yielding SNR distributions with mean values of -3.1 dB and -11.9 dB. Similarly, there was no significant difference between communication stage speech levels and normal speech levels in the active stage ($F(1,14) = 2.70, p = 0.13$), nor between communication SNRs and normal, active-stage speech SNRs ($F(1,14) = 3.37, p = 0.09$). However, there were significant effects of participant on the difference between the noise level distributions ($F(14,42) = 11.1, p < 0.001$), as well as on the difference between the speech level distributions ($F(14,14) = 9.41, p < 0.001$) and the corresponding SNR distributions ($F(14,14) = 9.41, p < 0.001$).

These results confirm that the background noise levels in the canteen stayed constant over the course of the RW assessment phase, which was intended to limit the intra-subject variability in the guided EMAs. In addition, the speech levels and SNRs remained constant between the communication task and the normal active listening task, implying that normal speech levels in the active listening task could be established using the communication task.



(A) Broadband level and SNR distributions



(B) Noise levels vs. speech levels

Figure 5.3: Panel A: Distributions of broadband noise SPLs (left), measured inside the canteen during the real-world passive (blue), communication (red) and active listening stages (green) as well as the guide's speech SPLs (middle) and SNRs (right) during the communication (red) and active listening stages (green). The noise and speech data use the left ordinate (in dB SPL), while the SNR data follow the right ordinate (in dB). The means (black circles) and standard deviations (black squares) of the distributions are indicated as well. Panel B: Noise levels vs. speech levels for the communication task and the active listening task at normal speech levels, along with least-squares fits to the data and its R^2 correlation value and significance p. The passive noise levels are shown on the bottom x-axis and do not have a corresponding speech level (marked by X). Least-squares fits between corresponding speech and noise levels are shown with their R^2 correlation factor and goodness-of-fit p-value.

For each of the participants, the maximum difference between the four background noise levels, as well as the maximum difference between the communication speech level and the normal active listening level, never exceeded 3 dB. Both the noise levels and the guide's speech levels did, however, vary significantly across different participants, despite the imposed time-of-day and location restrictions. The subjectively determined, challenging speech level for the normal-hearing participants was found to occur at a level of more than 10 dB below the background noise level.

Similarly, Fig. 5.3B shows the individual speech levels occurring during the communication task and the active listening task at normal speech levels as a function of their respective noise levels, as well as the noise levels during the passive listening task (shown on the bottom x-axis, without a corresponding speech level as marked by X). Least-squares fits between corresponding speech and noise levels are shown with their R^2 correlation factor and goodness-of-fit p-value. The results highlight the varying background noise levels across participants during the passive listening tasks and indicate that the speech and noise levels were significantly positively correlated across a similar range during the communication task and the active listening task at normal speech levels. This correlation is in agreement with the established increase in speech effort as well as level with increasing background noise level (known as the Lombard effect, Lombard, 1911) and explains the observed effects of participant in the statistical analysis. Particularly for the active listening task at normal speech levels, the least-squares fit matches the data very well and is evident from the low variance observed in the SNR distribution of the active listening task at normal speech levels.

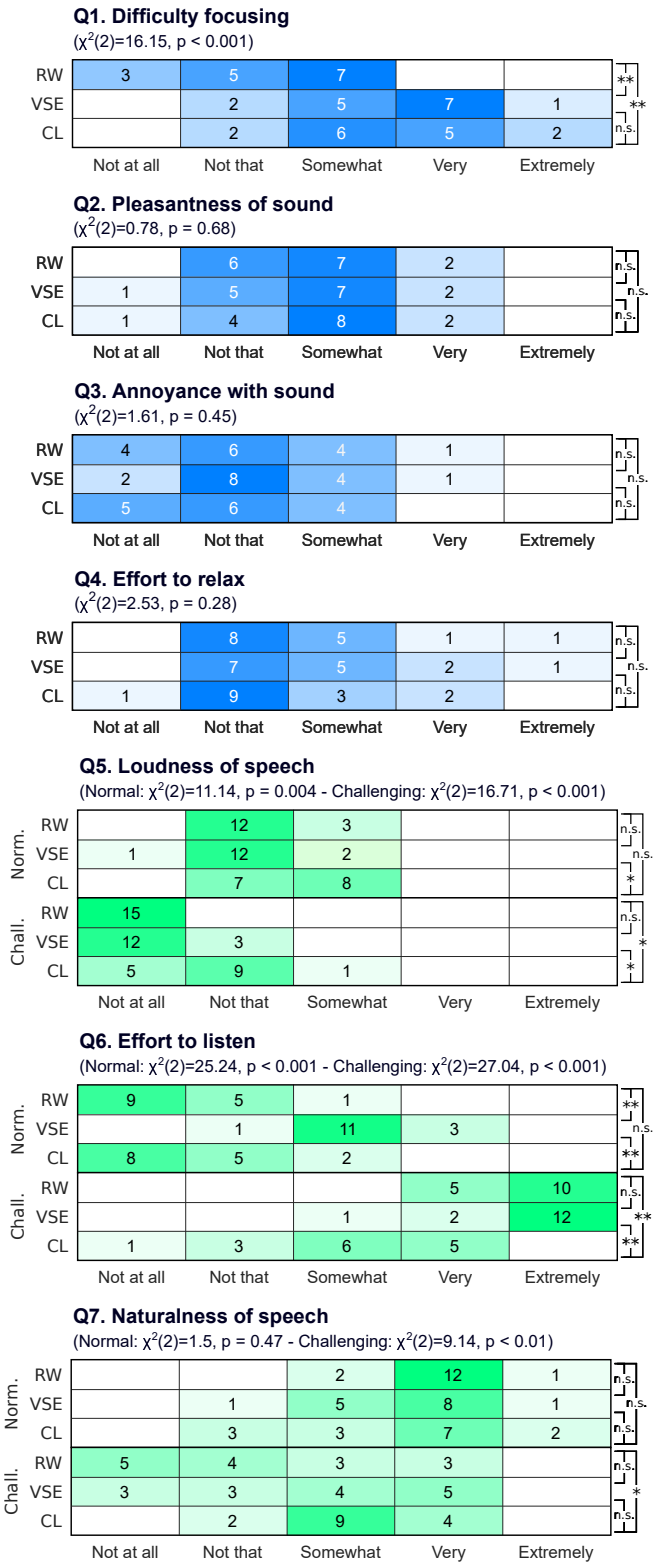


Figure 5.4: EMA questionnaire responses for the passive listening (left, blue) and active listening (right, green) stages. For each question, the number of responses for the real-world (RW), VSE and clinic (CL) condition are given for each possible response on the 5-point Likert scale. The saturation of each response box corresponds to the relative frequency of the response. The summary statistics are displayed underneath each question title and the significances of the post-hoc results are indicated to the right of each table. Each active listening stage question is divided into responses at the normal (Norm.) and challenging (Chall.) speech level.

5.3.2 EMA responses

Figure 5.4 displays the participants' responses to the questions in the passive (left, blue) and active (right, green) listening stages of the guided EMA. For each question, the number of responses for the RW, VSE and CL conditions are given for each possible response on the 5-point Likert scale. The saturation of each response box corresponds to the relative frequency of the response. Each active listening stage question is divided into responses at normal (Norm.) speech levels and at challenging (Chall.) speech levels. Due to the non-parametric nature of the categorical output data, a Friedman test was applied, combined with Wilcoxon-rank post-hoc tests with Bonferroni correction, to investigate differences between the conditions. The summary statistics are displayed underneath each question title and the significances of the post-hoc results are indicated on the right side of each table.

Of the passive stage questions (Q1 through Q4), only Q1 revealed significant differences between any of the conditions, specifically between the RW condition and both VSE and CL conditions. This implies that it was more difficult for participants to focus on specific sounds in the laboratory environments than in the real world (Q1). There were no significant differences between any of the conditions with regard to the experienced pleasantness of sound (Q2), the annoyance with sound (Q3) or the effort it took to relax (Q4). Participants generally agreed that the public lunch environment sounded mainly "somewhat pleasant" and "not that annoying", and was "not that effortful" to relax in. Except for Q1, these results support the absence of significant deviations in EMA results between the real world and the laboratory environments.

When actively listening to the guide's speech, its loudness was perceived as significantly higher in the CL condition than in the VSE condition, both at normal and challenging speech levels (Q5). At challenging speech levels, the RW condition was perceived significantly softer than the CL condition, but equally soft as the VSE condition. There were no significant differences between the RW and VSE conditions, which were perceived as mainly "not that loud" and "not at all loud" when judged at normal and challenging speech levels, respectively. These results indicate that while the VSE provided the expected consistency in loudness perception of speech to the real world, the clinic environment did not.

With regard to the participants' listening effort (Q6), the VSE condition was perceived as significantly more effortful than both the RW and CL conditions at

normal speech levels, which were both perceived as mainly "not at all effortful". This difference disappeared at challenging speech levels, where the listening effort was now lower in the CL condition than in the RW and VSE conditions, where listening was perceived as mainly "extremely effortful". In contrast to what was expected, the listening effort at normal levels was thus higher in the VSE than in the real world, despite its equal perceived loudness. While the clinic environment reflected real-world listening effort at normal speech levels, it resulted in an "underestimation" of listening effort at challenging speech levels, unlike the VSE which was now similar to the real world. However, the clinic environment's similarity in listening effort to the real world at normal speech levels may have been caused by a flooring effect on the response scale.

No significant differences in the naturalness of the speech (Q7) were observed between the RW, VSE and CL conditions at normal speech levels, perceived everywhere as mainly "very natural". At challenging speech levels, there was only a significant difference between the RW and CL conditions, whereby the RW condition was considered as mainly "not at all natural" compared to the CL condition being perceived as mainly "somewhat natural". As expected, the real-world naturalness of speech was thus preserved inside the laboratory environments at normal speech levels. At challenging speech levels, the naturalness of speech in the real world as well as the VSE was reduced, potentially due to the level of the speech stimulus being considered unnaturally low. This effect was partially mitigated in the CL condition due to the increased perceived loudness of speech (Q6).

Interestingly, there were no significant differences between participants for any of the passive listening questions and for any of the active listening questions at normal speech levels, despite the indicated fluctuations in speech and noise levels. All active listening questions at challenging speech levels contained a significant effect of participant. These observations support the guided EMA method's ability to produce, under normal listening circumstances, consistent assessments across participants. At challenging speech levels, the variability between participants increased, potentially due to the greater variance in SNRs and the absence of ceiling effects.

Figure 5.5 shows the distributions of percentage speech understanding for the final active listening question about self-assessed speech understanding (Q8) in each condition (RW, VSE, CL) at normal (Norm.) and challenging (Chall.) speech levels. A one-way repeated-measurement ANOVA (RANOVA) showed a

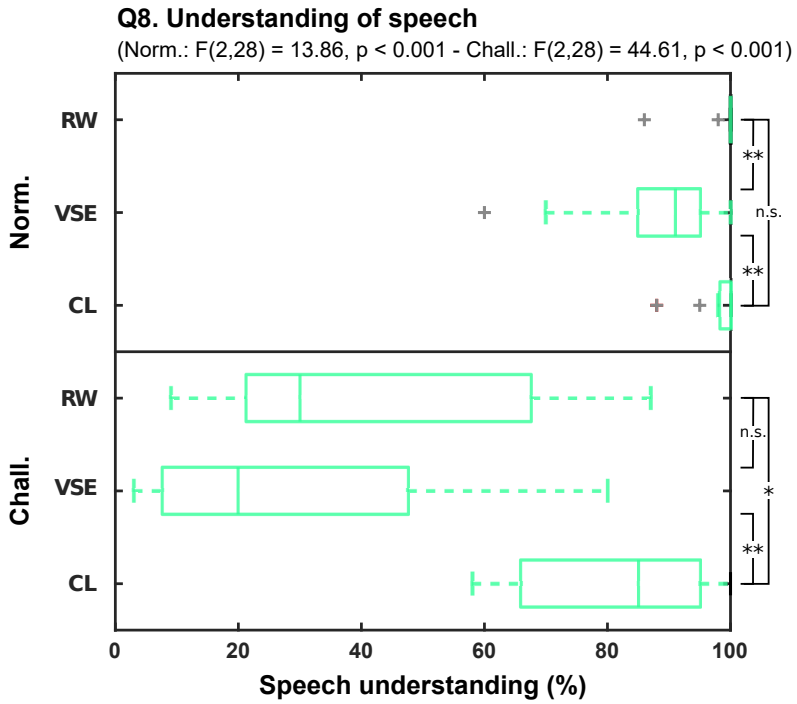


Figure 5.5: EMA questionnaire responses for question 8 on self-assessed speech understanding. Panel A shows the ratings for the real-world (RW), VSE and clinic (CL) condition, at the normal (Norm.) and challenging (Chall.) speech levels. The summary statistics (one-way repeated-measurement ANOVA) are displayed underneath the question title and the post-hoc results (paired-samples t-test) are indicated to the right of the table.

significant effect of condition, both at normal and challenging speech levels. As verified by paired-samples t-tests, speech at normal levels was understood significantly less well in the VSE condition than in both the RW and CL conditions, which were not significantly different from each other. This difference disappeared at challenging speech levels, where the RW and VSE conditions were no longer significantly different even though the VSE condition remained somewhat more challenging than the RW condition. Here, speech understanding in the RW and VSE conditions was significantly lower than in the CL condition. Similar to Q6, the clinic environment seems to have been affected by a ceiling effect at normal speech levels, whereas the VSE reflected the real world most accurately at challenging speech levels.

Figure 5.6 shows psychometric functions of the participants' self-assessed speech understanding scores, representing each score as a function of the corresponding speech SNR established in the real-world and laboratory environments (see Fig. 5.3). The psychometric functions for the RW and VSE conditions are

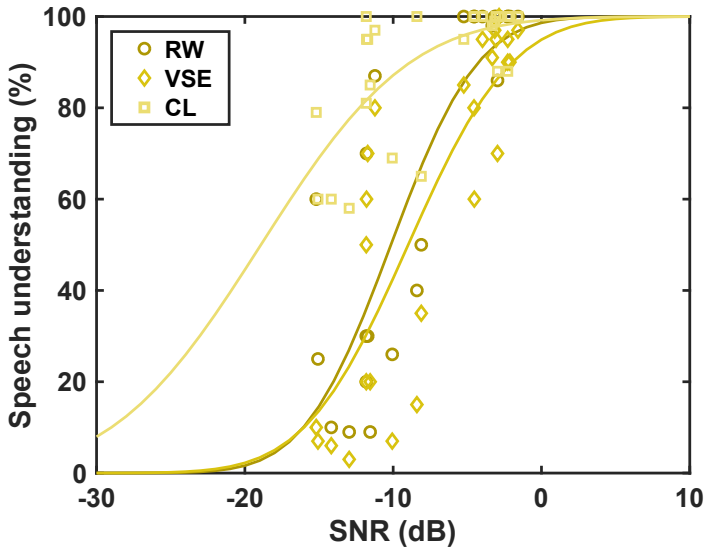


Figure 5.6: Psychometric curves (solid lines) fitted to the participants' self-assessed speech understanding scores in question 8 as a function of the corresponding guide's speech SNRs established in the real-world (RW, circles), VSE (diamonds) and clinic (CL, squares) environments.

very similar in their overall range and slope, with a 50% correct SNR of -10.2 dB and -9 dB, respectively. The CL psychometric function has a shallower decay towards far more negative SNRs, with a 50% correct SNR at -19 dB. R^2 goodness-of-fit values for the psychometric functions are 0.69 for the RW condition, 0.60 for the VSE condition and 0.53 for the CL condition. This indicates that self-rated speech understanding in the VSE resembled real-world values very closely, while the clinic environment resulted in both substantially overestimated speech understanding ratings as well as a poorer fit. The poorer quality of the clinic environment psychometric function is exacerbated by the absence of data points below 60% correct understanding.

5.4 Discussion

5.4.1 Real-world noise and speech levels

Due to the highly structured design of the guided EMA experiment and the presence of a guide, participants were able to carry out the EMAs without fail, indicating the full compliance of the participants in the proposed guided EMA method. The participant burden was reduced by the help of the guide and

the limited assessments required by the participant. By using the same RW environment across all participants, selected for its importance, common occurrence and difficulty in people's lives, and limitating the experiment in time, the method also aimed to reduce inter- and intra-participant data variability. The observations from Fig. 5.3 showed that the background noise levels in the RW environment were consistent across the different assessment stages and that the guide's speech levels during the communication stage were similar to their normal speech levels during the active listening stage, as intended. Thus, the RW environment was acoustically stable in terms of sound levels over the course of the RW assessment stage, despite modest fluctuations across different participants caused by the changing distribution and number of interferers in the RW environment. Interestingly, the normal-speech-level SNRs were consistently negative around -4 dB, implying that conversational SNRs between normal-hearing interlocutors reached values below 0 dB even at noise levels below 70 dB SPL. Noise levels necessary to produce negative SNRs were reported to be over 5 dB higher in other studies (Pearsons et al., 1977; Weisser and Buchholz, 2019). This may partially have been caused by the method with which the speech and noise levels were derived from the microphone recordings, which has been shown to result in lower, yet more accurate, SNR estimates (see Chapter 2, Mansour et al., 2021).

5.4.2 EMA responses

With regard to data variability and laboratory applicability, the passive stage questionnaire results (Q1-Q4) seem to have provided focused and consistent responses across participants (in favor of the first hypothesis) and, with the exception of Q1, across environmental conditions. The highly similar perception of pleasantness of sound (Q2), annoyance with sound (Q3) and effort to relax (Q4) between the RW, VSE and CL conditions further indicates that both laboratory environments could reproduce these sensations realistically (in favor of the second hypothesis). The increased difficulty of focusing on specific sounds (Q1) in the laboratory environments was likely due to the absence of visual stimuli, which are known to aid sound source localization (Shelton and Searle, 1980), as well as (to a lesser extent) resulting from the limited spatial resolution the VSE could provide (Huisman et al., 2020).

Similarly, the active stage assessments yielded consistent RW responses between participants at normal speech levels (in favor of the first hypothesis), with

the variance increasing somewhat at challenging levels. However, there were significant differences in the performance of the two laboratory environments relative to the real-world environment (in contrast to the second hypothesis).

First, despite the presence of an AV speech stimulus, the VSE caused an increased listening effort and speech understanding difficulty compared to the real world. This was likely a consequence of the imperfect sound field reconstruction of the target speech by the 4th order Ambisonics system which has been shown to negatively affect speech intelligibility (Favrot and Buchholz, 2009). The absence of the differences in listening effort and speech understanding at challenging levels may have been caused by the overall perceived difficulty of the task. Nevertheless, the similarity between the psychometric functions of the VSE and the real world (see Fig. 5.5B) suggests that the VSE-based guided EMA task could discriminate self-rated speech understanding in a similar way to the real world. Second, the greater perceived loudness of speech (Q5), reduced listening effort (Q6) and shallower psychometric function resulting from the CL condition compared to the VSE and RW conditions suggests that the clinic environment allowed for an overall easier perception of speech in noise than the VSE and the real world.

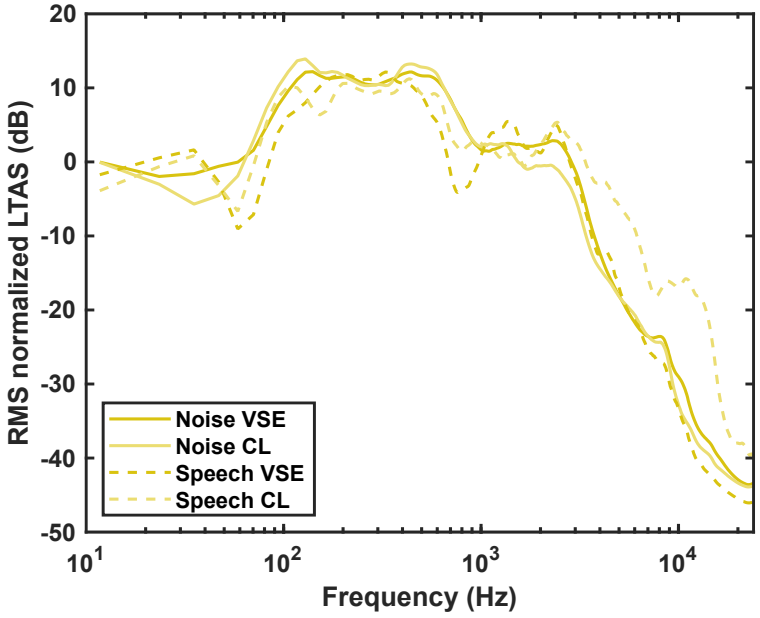


Figure 5.7: Long-term average spectra of the VSE and CL noise (solid lines) and speech (dashed lines), binaurally recorded inside the loudspeaker array, averaged over both ears and normalized relative to their average power.

The perceptual differences between the VSE and CL conditions might have originated from the acoustic differences between the single-loudspeaker anechoic CL speech source and the 64-loudspeaker reverberant VSE speech source. The absence of reverberation in the single-loudspeaker CL speech increased its intelligibility compared to the more reverberant VSE speech (Duquesnoy and Plomp, 1980), consistent with the higher speech transmission index (Steeneken and Houtgast, 1980) of 0.98 for the CL stimulus compared to 0.89 for the VSE speech. In addition, the spectral dissimilarities between the VSE and CL speech stimuli may have further contributed to the differences in perception. Figure 5.7 displays the long-term average speech spectra (LTAS) of the VSE and CL background noise (solid lines) and the VSE and CL speech (dashed lines), binaurally recorded inside the loudspeaker array, averaged over both ears and normalized relative to their average power. While the LTAS for both noise types was very similar (average power difference of less than 1 dB), the CL speech contained more speech power than the VSE speech, particularly in the region between 700 Hz and 1 kHz and anywhere above 1.5 kHz (average power difference of 3 dB).

5.4.3 Limitations and outlook

The high naturalness with which speech was perceived by participants across conditions at normal speech levels showed that the guided EMA methodology could elicit natural listening experiences, in the real world as well as in the lab. Nevertheless, more efforts can still be made to improve the realism of the laboratory environments. Particularly the inclusion of more realistic visuals of the surroundings in addition to the target talker video would bolster the validity of EMA inside laboratory environments even further. The fact that participants always assessed the real-world environment first may have biased some of the EMAs due to the prior knowledge of what the environment was supposed to look and sound like. This was a necessary constraint, since the real-world speech and noise signals needed to be captured to inform the reproduction inside the laboratory environments. Finally, an important next step would be to apply the guided EMA methodology to hearing-impaired individuals, a participant group which was not included in the current study due to restrictions imposed by the COVID-19 pandemic. By evaluating hearing-impaired participants, unaided as well as aided by a hearing device, differences in subjective hearing ability between participants, as well as the effects of wearing a hearing device, could be captured. As such, the method of guided EMA could be further validated

and potentially used to relate subjective hearing ability to objective measures of hearing loss and hearing device processing, both in the real world as well as in the lab.

5.5 Conclusion

This study explored a guided approach to EMA, which was designed to represent high feasibility, high participant compliance and low burden, as well as low inter- and intra-subject variability. The method was applied in a RW canteen environment and inside a VSE and a clinic laboratory environment, which were both acoustically matched with respect to the guide's speech and background noise levels. The results showed that the guided EMA methodology produced consistent passive listening EMAs within and across participants and environments. During active listening, the VSE generally resulted in EMAs most similar to the real-world environment, an observation which was supported by their highly similar psychometric functions. The clinic environment was perceived as less challenging, likely due to the increased intelligibility of its target speech source. The method of guided EMA may provide a new way of assessing subjective hearing ability in the real world as well as in the lab that can capture differences between participants and relate them to objective acoustic or psychoacoustic outcome measures.

6

General discussion

In this thesis, the evaluation of hearing and hearing devices was investigated in a more ecologically valid way by making use of realistic VSEs. A method was developed to accurately estimate real-world conversational SNRs in-situ (*Chapter 2*). Knowledge of these SNRs was used in a VSE-based SI experiment which explored differences in NH and HI SI performance caused by the increased ecological validity of the stimuli and reproduction method (*Chapter 3*). The same paradigm was subsequently employed to assess the impact of HA DRC processing in realistic VSEs (*Chapter 4*). Finally, a guided approach to EMA was proposed that attempted to reduce the data variability inherent in traditional EMA methodologies and was evaluated inside VSE-based laboratory environments (*Chapter 5*).

6.1 Summary of main findings

The two-channel method for realistic, conversational SNR estimation was shown to yield more accurate SNR estimates in a room acoustic simulation compared to those produced by a single-channel reference method. The accuracy advantage of the two-channel approach increased with decreasing SNR due to the saturation of the single-channel method's estimate at negative SNRs caused by its inability to reliably estimate the speech power in dominating background noise. A similar pattern was observed when applying both methods to in-situ recordings made in two real-world scenes.

Overall, the VSE-based SI tasks revealed an increased difficulty of understanding speech in noise (higher SRTs) compared to a headphone-based reference condition. This was the case both for NH listeners and, in particular, for HI listeners. An intermediary condition using artificial spatialized noise revealed that the presence of envelope modulations and intelligible interferers in the realistic background noise contributed to the increased SI difficulty experienced by the HI listeners. SRSs of around 67% correct at the median NH conversational

SNR of -2.5 dB could be used to relate SI performance to communication ability in a real-world scene.

When aided by a simple HA with DRC processing, HI listeners continued to have a reduced SI performance in the realistic VSE-based condition compared to a more artificial spatialized noise condition. In addition, the HA processing provided a greater benefit in the realistic VSE than in the artificial spatialized noise. By instrumentally analyzing the HA processing, the difference in HA performance could be related to modulation properties of the background noise and their effect on the SNR at the output of the compressive HA processing.

Lastly, the guided EMA method showed mostly consistent assessments of subjective passive and active hearing ability across participants and across the real-world and the acoustically matched VSE conditions. A simplified, three-loudspeaker clinic condition over-estimated active hearing ability, i.e. speech-in-noise performance, both at normal and challenging speech levels, likely due to the absence of reverberant target speech. Guided EMA may provide an outcome measure that can relate subjective hearing ability to objective differences in hearing status and the environmental acoustics.

6.2 The importance of accurate real-world SNRs

Since conversational SNRs have been shown to directly affect a person's SI performance (Bradley et al., 1999) as well as the performance of HA algorithms, it is important to capture their real-world values accurately. Differences in conversational SNRs can be influenced by many factors, including the considered real-world acoustics and the nature of the communication task. However, the observations in *Chapter 2* indicated that the traditional single-channel estimation method introduced a fundamental positive bias in its SNR estimates. This bias was most severe at negative SNRs, which was also where the in-situ measured NH SNR distributions were situated. Thus, real-world conversations may occur at considerably lower SNRs for commonly occurring background noise levels than what has previously been reported. Since those negative real-world SNRs are especially challenging for HI listeners, the previous estimates may have underestimated HI people's struggle in everyday life.

6.3 Speech intelligibility in the VSE versus the real world

The finding that a more realistic, VSE-based SI task lead to higher overall SRTs, i.e. lower SI performance, compared to those obtained with artificial SI paradigms is in line with results obtained in previous research (Best et al., 2015; Culling, 2016). However, the greater detrimental effect of the realistic VSE on HI listeners' SI performance, compared to NH listeners, demonstrated that such increased realism was required to more accurately characterize the effect of hearing loss on SI performance. In addition, by increasing the ecological validity of an SI task, SRSs obtained at real-world SNRs corresponded more closely to SI performance in the real world. It is important that the values of such SRSs, which reflect the difficulty of understanding speech in noise, are established for the "right" reasons. For example, a headphone-based SI paradigm that results in increased SRSs, at a constant offset with respect to a realistic SI paradigm, may not be an ecologically valid replacement for that realistic paradigm, even when compensating for such an offset. This is because the precise real-world mixture of (psycho)acoustic phenomena that influence the realistic SRSs is not present in the headphone-based paradigm. However, since contextual visual stimuli were not present in the VSE-based SI tasks considered in *Chapter 3*, these tasks probably underestimated SI performance at a given SNR. If visual information was included, this would cause SRSs to be higher for the real-world SNRs than the 67% found here. Generally, the more realistic a VSE-based SI task becomes, the closer its objective SRSs will correspond to the "true" SRS values occurring in the real world.

6.4 Realistic hearing aid testing

The instrumental HA analysis in *Chapter 4* revealed that the potential benefit of compressive HA processing on SI strongly depended on the type of background noise as well as the input SNR of the target speech. This implies that SI results obtained with paradigms that do not make use of realistic stimuli or reproduction methods may not reflect the (lack of) benefit provided by HAs in the real world. Furthermore, the SNR range within which SRTs converged in an adaptive scoring procedure critically affected the performance of the HA used by the HI listeners, emphasizing again the importance of knowing real-world conversational SNRs and constructing SI tasks that capture real-world

performance at those SNRs. Even though current HAs do not (yet) make use of visual information, the decrease in SRTs caused by adding visual stimuli to the SI task would affect the performance of the HA, particularly at challenging SNRs.

Beyond influencing instrumental HA performance, the realism of the VSE-based SI task may also influence the perceptual benefit experienced by aided listeners, e.g. through restoring dip-listening ability and amplifying high-frequency speech portions. Disentangling instrumental and perceptual effects of HA processing in a purely psychoacoustic experiment is difficult, and it would therefore be important to conduct purely acoustic HA tests (e.g. using a HATS) inside realistic VSEs as well. Nevertheless, the HA user's individual hearing ability is inextricably linked to any potential benefit of their HA, such that ecologically valid psychoacoustic evaluations are at least as important as objective, acoustic ones. *Chapter 4* only considered relatively simple HA processing, and it is likely that more advanced configurations, e.g. including beamforming and noise reduction, would render an SNR-dependent benefit that is different from that shown in this thesis. The principle, however, remains the same.

6.5 Real-world hearing ability and EMA

The most ecologically valid measure of real-world hearing ability and the potential benefit of a HA processing strategy is arguably the user's subjective opinion. The main takeaway from *Chapter 5* is that the proposed guided approach to EMA was able to provide concise, questionnaire-driven measures of self-rated, real-world hearing ability which could be reproduced inside acoustically matched, realistic laboratory environments. This may provide opportunities for using guided EMA as a tool to relate subjective hearing experiences to objective measures like SI or environmental acoustic properties. Moreover, by applying the methodology to (un)aided HI listeners, effects of HA processing could be directly related to a person's real-world experience. The applicability of guided EMA in the lab allows for these comparisons to happen in a controlled, yet more ecologically valid way.

From a clinical perspective, guided EMA may be useful as a realistic speech-in-noise task in the context of hearing evaluation and HA fitting, using knowledge of predetermined correlations between HA parameter settings and EMA responses. While further research beyond this exploratory work is needed to

better match EMAs obtained inside a simplified clinic laboratory environment to those obtained in the real world, VSE-based guided EMA tasks provide yet another application for VSEs in hearing research, bridging the gap between the laboratory and the real world.

6.6 The future of ecologically valid hearing research

The way people use their auditory system to communicate in complex, real-world acoustic scenes is characterized by a multitude of interacting physiological, psychoacoustic and behavioral mechanisms. Highly controlled experimental paradigms are valuable and necessary for isolating individual components of this system, signifying a bottom-up approach. Highly realistic paradigms, signifying a top-down approach, remain indispensable to empirically characterize the way in which those individual components work together to produce a subjectively perceived reality. Even though the focus in psychoaoustic research has been mainly on the former approach, both are symbiotic elements of a reality where the whole may be greater than the sum of its parts.

In this thesis, approaches to hearing and hearing device testing were explored that attempted to increase their ecological validity by trading off control and realism. However, as discussed, several limitations to the present work remain, providing possibilities for future investigations. Further improvements to the proposed real-world SNR estimation method, such as accounting for a moving target talker or making the setup more portable, may render it practically applicable in a wide variety of everyday sound scenarios, providing estimates of conversational SNRs for use in fully integrated, audiovisual speech-in-noise experiments. Those experiments could then be used to evaluate more advanced HA processing strategies with regard to both objective and subjective measures of hearing ability and support HA fitting procedures.

Lastly, an often overlooked advantage of constructing speech-in-noise experiments that are maximally ecologically valid is that potentially unknown phenomena may be reflected in their outcome measures. Paradigms that trade off control and realism in an optimal way may thus lead into uncharted auditory territory. By integrating ever more realistic sensory modalities into experimental designs, future ecologically valid hearing research may one day be able to obtain maximally controlled and fully realistic measures of hearing ability.

A

Appendix

A.1 NAL-NL2 fitting rationale

The NAL-NL2 rationale attempts to restore speech intelligibility of a HA user while preserving comfortable normal-hearing loudness levels. This is achieved using a level- and frequency-dependent fitting formula, that specifies real-ear insertion gain (REIG) factors to be applied to the input signal at 50 dB, 65 dB and 80 dB across 19 frequencies between 125 Hz and 8 kHz, spaced 1/3rd of an octave apart. Linear level interpolation is applied in the 50-65 dB and 65-80 dB ranges, while linear gain is applied below 50 dB and above 80 dB. Further linear frequency interpolation is applied to match the center frequencies of the HA filter bank. The fitting formula is defined for a range of 10 standard audiograms (Bisgaard et al., 2010), from which the one is selected that minimizes the sum of absolute differences with the listener's available audiogram frequencies.

A.2 openMHA programming

The openMHA framework uses a custom configuration language allowing for line-by-line human-readable text commands to be inserted in a configuration file in order to build a HA signal processing chain. A basic code sample is shown in Fig A.1. Elements of the microphone and receiver level equalization stages are shown, as well as excerpts from the algorithm chain, the filter bank settings and the input-output channels.
All of the used processing algorithms, including the DRC and beamforming, were sourced from available openMHA plugins. To control the basic operations of toggling the HA processing and loading appropriate configuration files, openMHA uses Java interface libraries that can invoked from encapsulating MATLAB functions. This allowed for a homogeneous implementation of the HA processing alongside the MATLAB-based spatial sound processing programming.

```
mha.plugin_name = mhachain
...
mha.calib_in.peaklevel = [111.3537 112.166 111.0489 112.1515]
mha.calib_in.fir   = [[-8.2451e-05 -0.00013246 -0.00018752 ...]
...
mha.calib_out.peaklevel = [112.5364 110.1078]
mha.calib_out.fir = [[-0.00019513 -0.00018781 -0.00025777 ...]
...
mha.mhachain.overlapadd.mhachain.algos = ...
                          [route:left_in acSteer:mvdr steerbf:left ...
                           route:right_in steerbf:right route:out ...
                           fftfilterbank dc combinechannels]
...
mha.mhachain.overlapadd.mhachain.fftfilterbank.unit = Hz
mha.mhachain.overlapadd.mhachain.fftfilterbank.f = ....
              [177 297 500 841 1414 2378 4000 6727 11314]
...
io.con_in = [system:capture_1 system:capture_2 ...
                system:capture_3 system:capture_4]
io.con_out = [system:playback_3 system:playback_4]
```

Figure A.1: Excerpts from an openMHA configuration file

In addition, the openMHA framework provides a graphical user interface for deriving appropriate configuration file values based on an input audiogram and a desired fitting rationale. The NAL-NL2 rationale was implemented specifically for this experiment.

A.3 Hearing aid calibration

The microphones and receivers in the master HA were calibrated to ensure proper operation with the listeners. To calibrate the 4 HA microphones for a equal level across frequencies, the HAs were placed in an portable anechoic enclosure inside of which 3rd-octave band white noise bursts around the HA filter bank center frequencies were played at a level of 80 dB SPL. The obtained digital levels for each of the microphones were then transformed to the appropriate openMHA configuration peak values and FIR filters using openMHA-provided scripts. The validation of the microphone calibration consisted of rerunning the calibration procedure and verifying the correct 80 dB SPL values across the considered frequencies.

The HA receiver calibration ensured that the prescribed gain of the fitting rationale accounted for the natural frequency dependency of the ear canal, resulting in the real-ear insertion gain (REIG). By definition, the REIG is equal to the difference between the real-ear aided gain (REAG) and the real-ear unaided gain (REUG), referring to the gain at the eardrum compared to a reference point at the canal entrance,with the HA inserted into the ear canal and turned on or not inserted, respectively.
Here, a single set of general REUG values was measured for use across all listeners using the ear canals of the HATS, which was placed in the center of the loudspeaker array. Specifically, 1/3rd octave band noise around the HA filter bank center frequencies was played from the frontal horizontal loudspeaker in the array, calibrated to 80 dB SPL by a calibrated reference microphone placed at the entrance to the HATS' left and right ear canals. Simultaneously and without the HAs inserted, the levels at the calibrated ear drum HATS microphones were recorded. Subtraction of the ear canal entrance levels from the ear drum levels resulted in the left and right REUGs. The HA receivers were calibrated by repeating this procedure with the HAs inserted into the ear canals, configured to unit gain. The measured levels at the ear drum microphones then had to match the desired level of 80 dB SPL after subtraction of the REUG.
This procedure ensured that potential additional corrections to the receiver equalization filter, e.g. caused by the fact that the closed HA ear tip attenuated sound entering the ear canal in a frequency-dependent way, were taken into account in the calibration.

Bibliography

ANSI (1997). *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America.

Ahrens, A., M. Marschall, and T. Dau (2017). "Measuring speech intelligibility with speech and noise interferers in a loudspeaker-based virtual sound environment". In: *The Journal of the Acoustical Society of America* 141.5, pp. 3510–3510.

Ahrens, A., M. Marschall, and T. Dau (2019). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments". In: *Hearing research* 377, pp. 307–317.

Alexander, H. (1998). "Hearing Aids: Smaller and Smarter". In: *New York Times*.

Arons, B. (1992). "A review of the cocktail party effect". In: *Journal of the American Voice I/O Society* 12.7, pp. 35–50.

Aronson, E., J. M. Carlsmith, and P. C. Ellsworth (1990). *Methods of research in social psychology*. McGraw-Hill New York.

Astolfi, A and M Filippi (2004). "Good acoustical quality in restaurants: a compromise between speech intelligibility and privacy". In: *Proc. of ICA*, pp. 1201–1204.

Bentler, R. A., D. P. Niebuhr, J. P. Getta, and C. V. Anderson (1993). "Longitudinal study of hearing aid effectiveness. II: Subjective measures". In: *Journal of Speech, Language, and Hearing Research* 36.4, pp. 820–831.

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* 93.5, pp. 2764–2778.

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

Best, V., J. M. Buchholz, and T. Weller (2017a). “Measuring auditory spatial perception in realistic environments”. In: *The Journal of the Acoustical Society of America* 141.5, pp. 3692–3692.

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. 682–690.

Best, V., C. R. Mason, J. Swaminathan, E. Roverud, and G. Kidd Jr (2017b). “Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures”. In: *The Journal of the Acoustical Society of America* 141.1, pp. 81–91.

Bisgaard, N., M. S. Vlaming, and M. Dahlquist (2010). “Standard audiograms for the IEC 60118-15 measurement procedure”. In: *Trends in amplification* 14.2, pp. 113–120.

Boike, K. and P. Souza (2000). “Effect of compression ratio on speech recognition in temporally complex background noise”. In: *International Hearing Aid Conference, Lake Tahoe, CA*.

Bradley, J., R. D. Reich, and S. Norcross (1999). “On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility”. In: *The Journal of the Acoustical Society of America* 106.4, pp. 1820–1828.

Bradley, J. V. (1958). “Complete counterbalancing of immediate sequential effects in a Latin square design”. In: *Journal of the American Statistical Association* 53.282, pp. 525–528.

Brungart, D. S., M. E. Barrett, J. I. Cohen, C. Fodor, C. M. Yancey, and S. Gordon-Salant (2020). “Objective assessment of speech intelligibility in crowded public spaces”. In: *Ear and hearing* 41.Suppl 1, 68S.

Brungart, D. S., P. S. Chang, B. D. Simpson, and D. Wang (2006). “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation”. In: *The Journal of the Acoustical Society of America* 120.6, pp. 4007–4018.

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.

Compton-Conley, C. L., A. C. Neuman, M. C. Killion, and H. Levitt (2004). “Performance of directional microphones for hearing aids: real-world versus simulation”. In: *Journal of the American Academy of Audiology* 15.6, pp. 440–455.

Cord, M., D. Baskent, S. Kalluri, and B. Moore (2007). "Disparity between clinical assessment and real-world performance of hearing aids". In: *Hearing Review* 14.6, p. 22.

Cubick, J. and T. Dau (2016). "Validation of a virtual sound environment system for testing hearing aids". In: *Acta Acustica united with Acustica* 102.3, pp. 547–557.

Culling, J. F. (2016). "Speech intelligibility in virtual restaurants". In: *The Journal of the Acoustical Society of America* 140.4, pp. 2418–2426.

Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Ph.D. Thesis, University of Paris VI, France.

Daniel, J., S. Moreau, and R. Nicol (2003). "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging". In: *Audio Engineering Society Convention 114*. Audio Engineering Society.

Davis, A. C. and H. J. Hoffman (2019). "Hearing loss: rising prevalence and impact". In: *Bulletin of the World Health Organization* 97.10, p. 646.

Demol, M., W. Verhelst, and P. Verhoeve (2007). "The duration of speech pauses in a multilingual environment". In: *Eighth Annual Conference of the International Speech Communication Association*.

Dirks, D. D. and R. H. Wilson (1969). "The effect of spatially separated sound sources on speech intelligibility". In: *Journal of Speech and Hearing Research* 12.1, pp. 5–38.

Dreschler, W. A., H. Verschuure, C. Ludvigsen, and S. Westermann (2001). "ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment". In: *Audiology* 40.3, pp. 148–157.

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

Durlach, N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences". In: *The Journal of the Acoustical Society of America* 35.8, pp. 1206–1218.

Elberling, C., C Ludvigsen, and P. Lyregaard (1989). "DANTALE: A new Danish speech material". In: *Scandinavian Audiology* 18.3, pp. 169–175.

Elko, G. (2018). *em32 Eigenmike Version 18 Release Notes*. mh Acoustics.

Favrot, S. and J. M. Buchholz (2009). “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.

Fogerty, D., A. Alghamdi, and W.-Y. Chan (2020). “The effect of simulated room acoustic parameters on the intelligibility and perceived reverberation of monosyllabic words and sentences”. In: *The Journal of the Acoustical Society of America* 147.5, pp. 396–402.

Fogerty, D. and L. E. Humes (2010). “Perceptual contributions to monosyllabic word intelligibility: Segmental, lexical, and noise replacement factors”. In: *The Journal of the Acoustical Society of America* 128.5, pp. 3114–3125.

Freyman, R. L., U. Balakrishnan, and K. S. Helfer (2001). “Spatial release from informational masking in speech recognition”. In: *The Journal of the Acoustical Society of America* 109.5, pp. 2112–2122.

GRAS A/S (2018). *GRAS 45BC KEMAR Head & Torso with Mouth Simulator*. GRAS Sound & Vibration A/S.

Galvez, G., M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry (2012). “Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users”. In: *Ear and hearing* 33.4, p. 497.

Gatehouse, S. (1999). “Glasgow hearing aid benefit profile: derivation and validation of”. In: *Journal of the American Academy of Audiology* 10.80, p. 103.

Gatehouse, S. and W. Noble (2004). “The speech, spatial and qualities of hearing scale (SSQ)”. In: *International journal of audiology* 43.2, pp. 85–99.

Gerzon, M. A. (1973). “Periphony: With-height sound reproduction”. In: *Journal of the audio engineering society* 21.1, pp. 2–10.

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.

Grange, J. A. and J. F. Culling (2016). “The benefit of head orientation to speech intelligibility in noise”. In: *The Journal of the Acoustical Society of America* 139.2, pp. 703–712.

Hadley, L. V., W. O. Brimijoin, and W. M. Whitmer (2019). “Speech, movement, and gaze behaviours during dyadic conversation in noise”. In: *Scientific reports* 9.1, pp. 1–8.

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.

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.

Helfer, K. S. and R. L. Freyman (2008). “Aging and speech-on-speech masking”. In: *Ear and hearing* 29.1, p. 87.

Hendrikse, M. M., G. Grimm, and V. Hohmann (2020). “Evaluation of the Influence of Head Movement on Hearing Aid Algorithm Performance Using Acoustic Simulations”. In: *Trends in Hearing* 24, p. 2331216520916682.

Henry, J. A., G. Galvez, M. B. Turbin, E. J. Thielman, G. P. McMillan, and J. A. Istvan (2012). “Pilot study to evaluate ecological momentary assessment of tinnitus”. In: *Ear and hearing* 32.2, p. 179.

Herzke, T., H. Kayser, F. Loshaj, G. Grimm, and V. Hohmann (2017). “Open signal processing software platform for hearing aid research (openMHA)”. In: *Proceedings of the Linux Audio Conference*, pp. 35–42.

Hopkins, K. and B. C. Moore (2009). “The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise”. In: *The Journal of the Acoustical Society of America* 125.1, pp. 442–446.

Houben, R. et al. (2014). “Development of a Dutch matrix sentence test to assess speech intelligibility in noise”. In: *International Journal of Audiology* 53.10, pp. 760–763.

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.

Huisman, T., A. Ahrens, and E. MacDonald (2020). “Sound source localization with various ambisonics orders in virtual reality”. In: *The Journal of the Acoustical Society of America* 148.4, pp. 2786–2786.

Hummersone, C. (2020). *Impulse response acoustic information calculator*. Github.

Hunt, R., S. Bell, and D. Simpson (2019). “Predicting the impact of hearing aid processing on speech intelligibility”. In: *The Journal of the Acoustical Society of America* 146.4, pp. 2920–2920.

ITU-T (2018). *Recommendation ITU-T P.570: Artificial noise fields under laboratory conditions*. International Telecommunication Union.

Jirsa, R. E. and T. W. Norris (1982). "Effects of intermodulation distortion on speech intelligibility." In: *Ear and hearing* 3.5, pp. 251–256.

Jørgensen, S. and T. Dau (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing". In: *The Journal of the Acoustical Society of America* 130.3, pp. 1475–1487.

Keidser, G., H. Dillon, M. Flax, T. Ching, and S. Brewer (2011). "The NAL-NL2 prescription procedure". In: *Audiology research* 1.1, pp. 88–90.

Kelly, H., G. Lin, N. Sankaran, J. Xia, S. Kalluri, and S. Carlile (2017). "Development and evaluation of a mixed gender, multi-talker matrix sentence test in Australian English". In: *International journal of audiology* 56.2, pp. 85–91.

Kim, C. and R. M. Stern (2008). "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis". In: *Ninth Annual Conference of the International Speech Communication Association*.

Kinnunen, T. and H. Li (2010). "An overview of text-independent speaker recognition: From features to supervectors". In: *Speech communication* 52.1, pp. 12–40.

Köbler, S and U Rosenhall (2002). "Horizontal localization and speech intelligibility with bilateral and unilateral hearing aid amplification". In: *International journal of audiology* 41.7, pp. 395–400.

Kochkin, S. (2002). "10-year customer satisfaction trends in the US hearing instrument market". In: *Hearing Review* 9.10, pp. 14–25.

Kowalewski, B. et al. (2018). "Effects of slow-and fast-acting compression on hearing-impaired listeners' consonant–vowel identification in interrupted noise". In: *Trends in hearing* 22.

Licklider, J. (1948). "The influence of interaural phase relations upon the masking of speech by white noise". In: *The Journal of the Acoustical Society of America* 20.2, pp. 150–159.

Likert, R. (1932). "A technique for the measurement of attitudes." In: *Archives of psychology*.

Litovsky, R. Y. (2005). "Speech intelligibility and spatial release from masking in young children". In: *The Journal of the Acoustical Society of America* 117.5, pp. 3091–3099.

Lombard, E. (1911). “Le signe de l’elevation de la voix (The sign of the elevation of the voice)”. In: *Ann. Mal. de L’Oreille et du Larynx (Annals of Diseases of the Ear and the Larynx)*, pp. 101–119.

Lutman, M. E. (1991). “Hearing disability in the elderly”. In: *Acta Oto-Laryngologica* 111.476, pp. 239–248.

Mansour, N., M. Marschall, T. May, A. Westermann, and T. Dau (2021). “A method for realistic, conversational signal-to-noise ratio estimation”. In: *The Journal of the Acoustical Society of America* 149.3, pp. 1559–1566.

Mansour, N., M. Marschall, A. Westermann, T. May, and T. Dau (2019). “Speech intelligibility in a realistic virtual sound environment”. In: *23rd International Congress on Acoustics*. Deutsche Gesellschaft für Akustik eV, pp. 7658–7665.

May, T., B. Kowalewski, and T. Dau (2018). “Signal-to-noise-ratio-aware dynamic range compression in hearing aids”. In: *Trends in hearing* 22.

Miles, K. M., G. Keidser, K. Freeston, T. Beechey, V. Best, and J. M. Buchholz (2020). “Development of the Everyday Conversational Sentences in Noise test”. In: *The Journal of the Acoustical Society of America* 147.3, pp. 1562–1576.

Mills, M. (2011). “Hearing aids and the history of electronics miniaturization”. In: *IEEE Annals of the History of Computing* 33.2, pp. 24–45.

Minnaar, P., S. Favrot, and J. M. Buchholz (2010). “Improving hearing aids through listening tests in a virtual sound environment”. In: *The Hearing Journal* 63.10, pp. 40–42.

Moore, B., J. Alcántara, M. Stone, and B. Glasberg (1999). “Use of a loudness model for hearing aid fitting: II. Hearing aids with multi-channel compression”. In: *British journal of audiology* 33.3, pp. 157–170.

Moore, B. and B. Glasberg (1997). “A model of loudness perception applied to cochlear hearing loss”. In: *Auditory neuroscience* 3.3, pp. 289–311.

Moore, B. C. (1996). “Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids”. In: *Ear and hearing* 17.2, pp. 133–161.

Moore, B. C. and B. R. Glasberg (2004). “A revised model of loudness perception applied to cochlear hearing loss”. In: *Hearing research* 188.1-2, pp. 70–88.

Moore, B. C., R. F. Laurence, and D. Wright (1985). “Improvements in speech intelligibility in quiet and in noise produced by two-channel compression hearing aids”. In: *British Journal of Audiology* 19.3, pp. 175–187.

Moskowitz, D. S. and S. N. Young (2006). “Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology”. In: *Journal of Psychiatry and Neuroscience* 31.1, p. 13.

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

National Research Council (2004). *Hearing loss: Determining eligibility for social security benefits*. National Academies Press.

Naylor, G. and R. B. Johannesson (2009). “Long-term signal-to-noise ratio at the input and output of amplitude-compression systems”. In: *Journal of the American Academy of Audiology* 20.3, pp. 161–171.

Neely, K. K. (1956). “Effect of visual factors on the intelligibility of speech”. In: *The Journal of the Acoustical Society of America* 28.6, pp. 1275–1277.

Nielsen, J. B. and T. Dau (2009). “Development of a Danish speech intelligibility test”. In: *International journal of audiology* 48.10, pp. 729–741.

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

Nielsen, N. O., S. Santurette, and C.-H. Jeong (2016). “Subjective evaluation of restaurant acoustics in a virtual sound environment”. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 253. 2. Institute of Noise Control Engineering, pp. 5990–5999.

Noble, W. and S. Gatehouse (2006). “Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing scale (SSQ)”. In: *International Journal of Audiology* 45.3, pp. 172–181.

ODEON A/S (2020). *Version 16 User’s Manual*. ODEON Room Acoustics Software.

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.

Oreinos, C. and J. M. Buchholz (2016). “Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids”. In: *Journal of the American Academy of Audiology* 27.7, pp. 541–556.

Pearsons, K. S., R. L. Bennett, and S. Fidell (1977). *Speech levels in various noise environments*. Office of Health, Ecological Effects, Office of Research, and Development, US EPA.

Peters, R. W., B. C. Moore, and T. Baer (1998). “Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people”. In: *The Journal of the Acoustical Society of America* 103.1, pp. 577–587.

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

Reis, H. T. and C. M. Judd (2000). *Handbook of research methods in social and personality psychology*. Cambridge University Press.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler (2009). “The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise”. In: *The Journal of the Acoustical Society of America* 126.6, pp. 3236–3245.

Roman, N. and J. Woodruff (2013). “Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold”. In: *The Journal of the Acoustical Society of America* 133.3, pp. 1707–1717.

Sanchez-Lopez, R. H., T. Dau, and M. L. Jepsen (2019). “Hearing-aid settings in connection to supra-threshold auditory processing deficits”. In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 7, pp. 221–228.

Saunders, G. H. and J. M. Kates (1997). “Speech intelligibility enhancement using hearing-aid array processing”. In: *The Journal of the Acoustical Society of America* 102.3, pp. 1827–1837.

Scollie, S. et al. (2005). “The desired sensation level multistage input/output algorithm”. In: *Trends in amplification* 9.4, pp. 159–197.

Seewald, R. C., S. P. Bornstein, and K. J. Randolph (1981). “Speech intelligibility as a function of hearing aid microphone location”. In: *The Journal of the Acoustical Society of America* 70.S1, S108–S108.

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

Shiffman, S., A. A. Stone, and M. R. Hufford (2008). “Ecological momentary assessment”. In: *Annual Review of Clinical Psychology* 4, pp. 1–32.

Smeds, K., S. Gotowiec, F. Wolters, P. Herrlin, J. Larsson, and M. Dahlquist (2020). “Selecting scenarios for hearing-related laboratory testing”. In: *Ear and Hearing* 41, 20S–30S.

Smeds, K., F. Wolters, J. Larsson, P. Herrlin, and M. Dahlquist (2018). “Ecological momentary assessments for evaluation of hearing-aid preference”. In: *The Journal of the Acoustical Society of America* 143.3, pp. 1742–1742.

Smeds, K., F. Wolters, and M. Rung (2015). “Estimation of signal-to-noise ratios in realistic sound scenarios”. In: *Journal of the American Academy of Audiology* 26.2, pp. 183–196.

Soli, S. D. and L. L. Wong (2008). “Assessment of speech intelligibility in noise with the Hearing in Noise Test”. In: *International Journal of Audiology* 47.6, pp. 356–361.

Sørensen, A. J. M., E. N. MacDonald, and T. Lunner (2019). “Timing of turn taking between normal-hearing and hearing-impaired interlocutors”. In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 7, pp. 37–44.

Souza, P. E., K. H. Arehart, J. Shen, M. Anderson, and J. M. Kates (2015). “Working memory and intelligibility of hearing-aid processed speech”. In: *Frontiers in Psychology* 6, p. 526.

Steeneken, H. J. and T. Houtgast (1980). “A physical method for measuring speech-transmission quality”. In: *The Journal of the Acoustical Society of America* 67.1, pp. 318–326.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge, p. 249.

Stoica, P., R. L. Moses, et al. (2005). “Spectral analysis of signals”. In: *Pearson Prentice Hall Upper Saddle River, NJ*.

Stone, M. A. and B. C. Moore (1999). “Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses”. In: *Ear and Hearing* 20.3, pp. 182–192.

Studebaker, G. A., R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney (1999). “Monosyllabic word recognition at higher-than-normal speech and noise levels”. In: *The Journal of the Acoustical Society of America* 105.4, pp. 2431–2444.

Sumby, W. H. and I. Pollack (1954). “Visual contribution to speech intelligibility in noise”. In: *The journal of the acoustical society of america* 26.2, pp. 212–215.

Takahashi, G. A. and S. P. Bacon (1992). “Modulation detection, modulation masking, and speech understanding in noise in the elderly”. In: *Journal of Speech, Language, and Hearing Research* 35.6, pp. 1410–1421.

Timmer, B. H., L. Hickson, and S. Launer (2015). "Adults with mild hearing impairment: Are we meeting the challenge?" In: *International journal of audiology* 54.11, pp. 786–795.

Timmer, B. H., L. Hickson, and S. Launer (2017). "Ecological momentary assessment: Feasibility, construct validity, and future applications". In: *American Journal of Audiology* 26.3S, pp. 436–442.

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

Wagener Carola, K., M. Hansen, and C. Ludvigsen (2008). "Recording and classification of the acoustic environment of hearing aid users". In: *Journal of the American Academy of Audiology* 19.4, pp. 348–370.

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

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds". In: *Science* 167.3917, pp. 392–393.

Weisser, A. and J. M. Buchholz (2019). "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions". In: *The Journal of the Acoustical Society of America* 145.1, pp. 349–360.

Weisser, A., J. M. Buchholz, and G. Keidser (2019a). "Complex Acoustic Environments: Review, Framework, and Subjective Model". In: *Trends in hearing* 23, p. 2331216519881346.

Weisser, A. et al. (2019b). "The ambisonic recordings of typical environments (ARTE) database". In: *Acta Acustica United With Acustica* 105.4, pp. 695–713.

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

Westermann, A. and J. M. Buchholz (2017). "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.

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

Wolters, F., K. Smeds, E. Schmidt, E. K. Christensen, and C. Norup (2016). "Common sound scenarios: A context-driven categorization of everyday sound

environments for application in hearing-device research". In: *Journal of the American Academy of Audiology* 27.7, pp. 527–540.

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

Wu, Y.-H. (2010). "Effect of age on directional microphone hearing aid benefit and preference". In: *Journal of the American Academy of Audiology* 21.2, pp. 78–89.

Wu, Y.-H., E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, and J. Oleson (2018). "Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults With Mild to Moderate Hearing Loss." In: *Ear and Hearing* 39.2, pp. 293–304.

Wu, Y.-H., E. Stangl, X. Zhang, and R. A. Bentler (2015). "Construct validity of the ecological momentary assessment in audiology research". In: *Journal of the American Academy of Audiology* 26.10, pp. 872–884.

Yang, W and J. Bradley (2009). "Effects of room acoustics on the intelligibility of speech in classrooms for young children". In: *The Journal of the Acoustical Society of America* 125.2, pp. 922–933.

Contributions to Hearing Research

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vol. 40: *Niclas A. Janssen,* Binaural streaming in cochlear implant patients, 2019.
External examiners: Tim Jürgens, Hamish Innes-Brown

Vol. 41: *Wiebke Lamping*, Improving cochlear implant performance through psychophysical measures, 2019.
External examiners: Dan Gnasia, David Landsberger

Vol. 42: *Antoine Favre-Félix*, Controlling a hearing aid with electrically assessed eye gaze, 2020.
External examiners: Jürgen Tchorz, Graham Naylor

Vol. 43: *Raul Sanchez-Lopez*, Clinical auditory profiling and profile-based hearing-aid fitting, 2020.
External examiners: Judy R. Dubno, Pamela E. Souza

Vol. 44: *Juan Camilo Gil Carvajal*, Modeling audiovisual speech perception , 2020.
External examiners: Salvador Soto-Faraco, Kaisa Maria Tiippana

Vol. 45 *Charlotte Amalie Emdal Navntoft*, Improving cochlear implant performance with new pulse shapes: a multidisciplinary approach, 2020.
External examiners: Andrej Kral, Johannes Frijns

Vol. 46 *Naim Mansour*, Assessing hearing device benefit using virtual sound environments, 2021.
External examiners: Virginia Best, Pavel Zahorik

The end.

To be continued...

Hearing well in noisy everyday situations can be challenging, especially for people affected by hearing loss. Hearing devices try to restore a hearing-impaired person's ability to accurately perceive sounds and understand speech. However, many psychoacoustic tests currently in use to evaluate speech intelligibility and hearing device performance do not take the acoustic properties of complex real-world sound scenes into account, typically relying on artificial target speech and background noise signals presented over headphones or small sets of loudspeakers. While these laboratory settings provide highly controlled and reliable results, they may not entirely reflect how people experience their real-world auditory reality.

The aim of this thesis was to increase the realism in psychoacoustic listening tasks inside controlled laboratory environments by employing "virtual sound environments". A virtual sound environment, or VSE, consists of a spherical array of many loudspeakers and is capable of rendering physically accurate, spatial sound fields to a listener positioned in the center. By using VSEs in combination with spatially recorded real-world noise signals and spatialized target speech, realistic speech intelligibility tasks were designed and implemented. This included the development of a new, realistic method for measuring conversational speech levels. The tasks were shown to increase the difficulty of understanding speech compared to more artificial conditions, especially for hearing-impaired listeners. Hearing aids benefited speech intelligibility most in the realistic conditions, which could be related to properties of the speech and noise signals and their effect on the hearing aid processing. A newly devised method for evaluating subjective, listener-reported hearing ability in a more controlled way was shown to be applicable inside realistic VSEs.

Overall, this thesis showed the ability of VSE-based laboratory environments to provide increased realism in psychoacoustic listening tasks, as well as render more ecologically valid results for both normal-hearing and hearing-impaired listeners. The development of increasingly realistic hearing and hearing device evaluation tests, using these environments, has the potential to increase the benefit these devices provide to users in their everyday life.

DTU Health Tech
Department of Health Technology

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