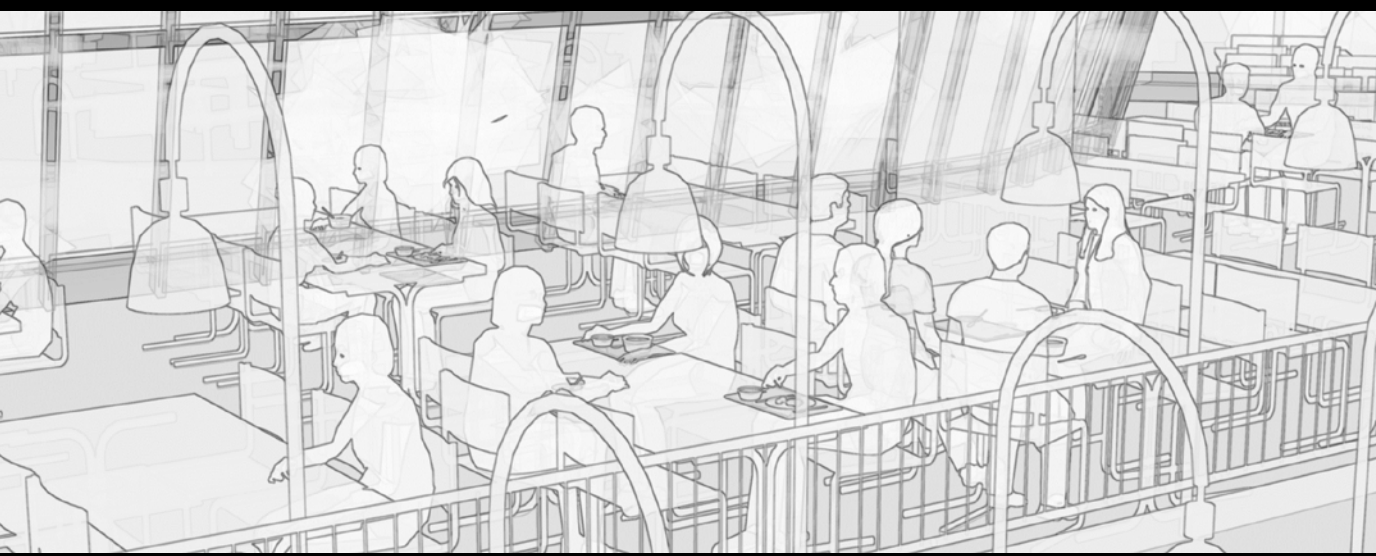
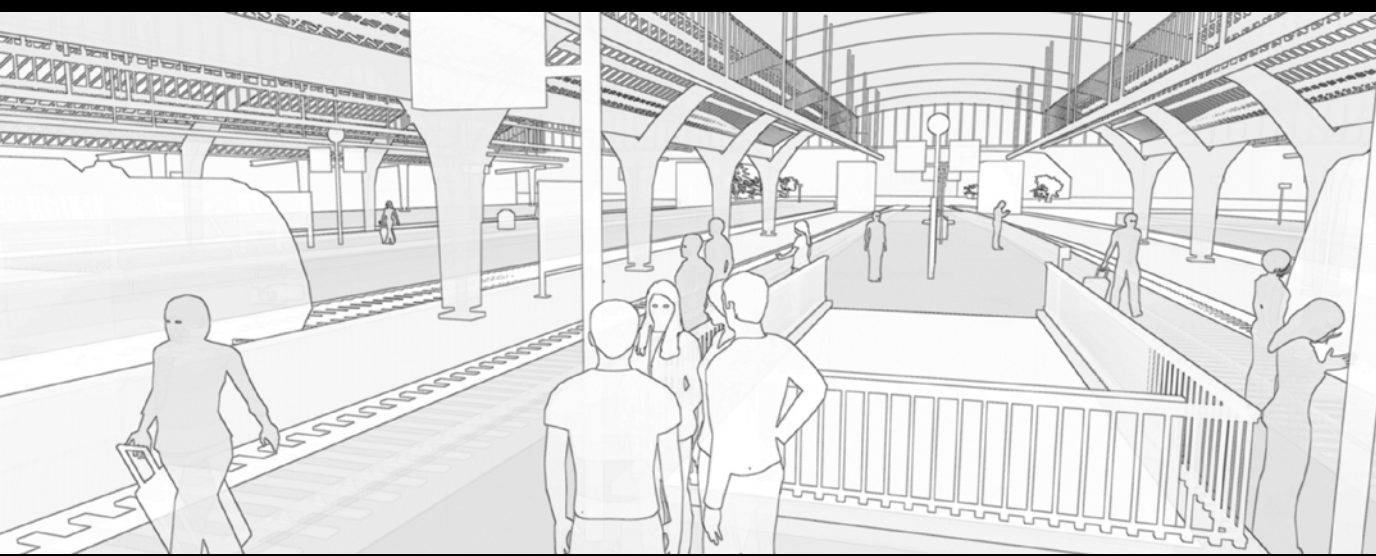
Carl von Ossietzky
Universität
Oldenburg

Audiovisual perception and psychoacoustics:
speech intelligibility, loudness perception, and technology preference

Gerard Llorach Tó





Audiovisual perception and psychoacoustics: speech intelligibility, loudness perception, and technology preference

Von der Fakultät für Medizin und Gesundheitswissenschaften

der Carl-von-Ossietzky Universität Oldenburg

zur Erlangung des Grades und Titels eines

Doktoren der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation

von

Gerard Llorach Tó

geboren am 22. Februar 1991

in Barcelona, Spanien

Gutachter: Volker Hohmann

Weitere Gutachter: Brian C.J. Moore, Hartmut Meister

Tag der Disputation: 28.08.2023

Summary

According to the World Health Organization (WHO), 5% of the population suffers from hearing impairment. It is predicted that these numbers will rise in the future due to increasing noise pollution and exposure. Hearing aids provide a solution to rehabilitate hearing disability and enrich spoken communication. Current hearing aid development and fitting are mostly based on properties of the hearing system, such as the loss of hearing in specific frequency bands. In realistic situations, hearing is influenced by many other factors, which are rarely considered in the clinic. Moreover, hearing aids usually perform well during clinical testing but not necessarily in real-life experiences. Unsatisfactory experiences by new hearing aid users lead to rejecting the technology, and thus to a poorer quality of life. Therefore, increasing the realism in clinical evaluations might improve hearing aid fitting, provide a fuller diagnosis of hearing loss, and thus contribute to better hearing and quality of life for the hearing impaired.

In this work, it is investigated what are the effects of visual cues on psychoacoustic experiments, if audiovisual simulations reflect real-life hearing-related activities, and if virtual reality technologies are ready for research and clinical procedures. The effects of visual cues are investigated in two areas: speech perception (Chapter 2), and loudness perception (Chapter 3). Chapter 4 investigates the preference and acceptance of audiovisual technologies. The methods used in this investigation range from simple clinical setups, such as headphones and a computer monitor, to complex audiovisual simulations with surrounding audio and immersive displays. These technologies aim at more

ecological validity, i.e., that the results gathered in the laboratory reflect real-life situations and hearing function. Virtual reality technologies are tested with young, older, and hearing-impaired participants to evaluate technology acceptance and that these technologies are not a deterrent factor for audiological procedures.

Speech perception (Chapter 2) is most important for human communication. Not being able to communicate with others is one of the reasons for reaching hearing acousticians and audiologists for counseling and treatment. The Matrix Sentence Test (MST) is an established test for measuring speech intelligibility. A method for adding synchronous visual speech to existing audio-only speech material is presented in this work. With this method, the audiovisual version of the German MST is developed and validated. It is found that visual speech contributes to speech understanding and that the test can be used for clinical evaluations. The visual speech recordings and the method for adding synchronous visual speech are published and accessible online.

Loudness perception of vehicles is one of the main causes of noise pollution and acoustic discomfort in cities. Loudness perception has been studied in the past, showing that even the color of a vehicle can influence how loud we perceive its sound. In this work, loudness perception of vehicles is evaluated in the field and in the laboratory (Chapter 3). Different vehicles are driven in a controlled outdoors environment, while loudness perception is evaluated. The vehicle driving actions are recorded and then played in the laboratory to the same participants to compare the perception in the laboratory and in the field. The laboratory conditions range from immersive visual cues and stereo audio to a single loudspeaker setup. The experiment shows that audiovisual setups with immersive cues induce participants to rate loudness as in the field, whereas

simplistic setups induce participants to rate sounds louder than in the field. The recordings of the vehicle driving actions are published and accessible online.

The third experiment of this work compares the preference for audiovisual technologies for clinical procedures (Chapter 4). The laboratory experiment uses different audiovisual conditions and technologies: a curved screen, a head-mounted display, video recordings and virtual characters. It is suggested that curved screens or other non-intrusive displays are the preferred option for clinical setups, but head-mounted displays can be used if needed. Video recordings are clearly chosen over virtual characters and no visual cues (audio only).

This work investigates how visual cues and laboratory setups influence audiovisual perception and preference. When evaluating speech perception, loudness perception and technology preference, visual cues were found to be relevant. Such results are particularly relevant when designing experiments, as in some cases the realism of the laboratory setup is crucial to obtain results that are closer to real-life situations. The materials used in this work are published and open for others to use.

Zusammenfassung

Nach Angaben der Weltgesundheitsorganisation (WHO) leiden 5 % der Bevölkerung an einer Hörbehinderung. Es wird prognostiziert, dass diese Zahlen in Zukunft aufgrund der zunehmenden Lärmbelastung und -belastung steigen werden. Hörgeräte bieten eine Lösung zur Rehabilitation von Hörbehinderungen und bereichern die gesprochene Kommunikation. Die derzeitige Entwicklung und Anpassung von Hörgeräten basiert hauptsächlich hinsichtlich der Eigenschaften des Hörsystems, wie beispielsweise dem Hörverlust in bestimmten Frequenzbändern. In Realsituationen wird das Gehör jedoch von weiteren, in Kliniken nur selten berücksichtigten Faktoren, beeinflusst. Daher lassen sich vorwiegend gut abgeschnittene klinische Test nicht nicht pauschal auf reale Erfahrungen von Hörgeräteträgern übertragen. Unbefriedigende Erfahrungen neuer Hörgeräteträger führen oft zur Ablehnung der Technologie von Hörhilfen und damit zu einer fortbestehenden schlechteren Lebensqualität. Daher könnten realitätsnähere klinische Bewertungsmethoden die Anpassung von Hörgeräten durch eine vollständige Diagnose des Hörverlustes allgemein verbessern, und somit durch im Resultat besseres Hören die Lebensqualität von Hörgeschädigten deutlich steigern.

In dieser Arbeit wird untersucht, welche Auswirkungen visuelle Hinweise auf psychoakustische Experimente haben. Hierbei werde erforscht ob audiovisuelle Simulationen reale hörbezogene Aktivitäten widerspiegeln und ob Virtual-Reality-Technologien für die Forschung und klinische Verfahren bereit sind. Die Auswirkungen von visuellen Hinweisen werden in zwei Bereichen untersucht:

Sprachwahrnehmung (Kapitel 2) und Lautstärkewahrnehmung (Kapitel 3). Kapitel 4 untersucht die Präferenz und Akzeptanz audiovisueller Technologien. Die in dieser Untersuchung verwendeten Methoden reichen von einfachen klinischen Aufbauten wie Kopfhörern und einem Computermonitor bis hin zu komplexen audiovisuellen Simulationen mit Umgebungsgeräuschen und immersiven Displays. Diese Technologien zielen auf mehr ökologische Gültigkeit ab, d. h. dass die im Labor gesammelten Ergebnisse reale Situationen und die Hörfunktion widerspiegeln. Virtual-Reality-Technologien werden mit jungen, älteren und hörgeschädigten Teilnehmern getestet, um die Technologieakzeptanz zu bewerten und festzustellen, dass diese Technologien kein abschreckender Faktor für audiologische Verfahren sind.

Die Sprachwahrnehmung (Kapitel 2) ist essenziell für die menschliche Kommunikation. Nicht in der Lage zu sein, mit anderen zu kommunizieren, ist einer der Gründe, Hörakustiker und Audiologen zur Beratung und Behandlung aufzusuchen. Der Matrixsatztest (MST) ist ein etablierter Test zur Messung der Sprachverständlichkeit. Es wird ein Verfahren zum Hinzufügen von synchroner visueller Sprache zu bestehendem Nur-Audio-Sprachmaterial entwickelt. Mit dieser Methode wird die audiovisuelle Version des deutschen MST entwickelt und validiert. Es wurde festgestellt, dass visuelle Sprache zum Sprachverständnis beiträgt und dass der Test für klinische Bewertungen verwendet werden kann. Die visuellen Sprachaufnahmen und das Verfahren zum Hinzufügen von synchroner visueller Sprache sind veröffentlicht und online zugänglich.

Die Lautstärkewahrnehmung von Fahrzeugen ist eine der Hauptursachen für Lärmbelästigung und akustische Beschwerden in Städten. Die Lautstärkewahrnehmung wurde in der Vergangenheit untersucht und es stellte

sich heraus, dass die Klangwahrnehmung sogar durch die Farbe eines Fahrzeugs beeinflusst werden kann. In dieser Arbeit wird die Lautstärkewahrnehmung von Fahrzeugen im Feld und unter verschiedenen Laborbedingungen evaluiert (Kapitel 3). Verschiedene Fahrzeuge werden in einer kontrollierten Außenumgebung gefahren, während die Lautstärkewahrnehmung bewertet wird. Die Fahraktionen des Fahrzeugs werden aufgezeichnet und dann im Labor denselben Teilnehmern vorgespielt, um die Wahrnehmung im Labor und im Feld zu vergleichen. Die Laborbedingungen reichen von immersiven visuellen Hinweisen und Stereo-Audio bis hin zu einem einzelnen Lautsprecher-Setup. Das Experiment zeigt, dass audiovisuelle Setups mit immersiven Hinweisen die Teilnehmer dazu veranlassen, die Lautstärke wie im Feld zu bewerten, während vereinfachte Setups die Teilnehmer dazu veranlassen, Geräusche lauter als im Feld zu bewerten. Die Aufzeichnungen der Fahrzeugfahraktionen werden online veröffentlicht und zugänglich gemacht.

Das dritte Experiment dieser Arbeit vergleicht die Präferenz für audiovisuelle Technologien für klinische Verfahren (Kapitel 4). Das Laborexperiment nutzt verschiedene audiovisuelle Bedingungen und Technologien: einen gekrümmten Bildschirm, ein Head-Mounted-Display, Videoaufnahmen und virtuelle Charaktere. Es wird vorgeschlagen, dass gekrümmte Bildschirme oder andere nicht störende Displays die bevorzugte Option für klinische Einrichtungen sind, aber bei Bedarf können am Kopf montierte Displays verwendet werden. Videoaufnahmen wurden eindeutig virtuellen Charakteren und keinen visuellen Hinweisen (nur Audio) vorgezogen.

Diese Arbeit untersucht, wie visuelle Hinweise und Laboreinstellungen die audiovisuelle Wahrnehmung und Präferenz beeinflussen. Bei der Bewertung

von Sprachwahrnehmung, Lautstärkewahrnehmung und Technologiepräferenz erwiesen sich visuelle Hinweise als relevant. Solche Ergebnisse sind besonders relevant bei der Gestaltung von Experimenten, da in einigen Fällen der Realismus des Laboraufbaus entscheidend ist, um Ergebnisse zu erhalten, die näher an realen Situationen liegen. Die in dieser Arbeit verwendeten Materialien werden veröffentlicht und stehen anderen zur Verwendung offen.

Acknowledgements

Thea and Ruth, without you I don't know what I would have done in Oldenburg and probably I would not have been able to finish this work, thanks a lot! Graham and Xana, always there and present, my solid and caring friends, thank you so much for your time. Milan and Max, so many years letting our music expressions mix and communicate, those sessions where memorable. Alex, the summer guy, you were a pilar in my stay in Oldenburg. Peter, you made it to Hamburg, I am proud of you. Pablo, my skateboarding mentor, together with Ricardo, Alberto and Peter, beautiful nights we had wrapped with the barbecue smell.

Volker, thank you for betting on me and supporting me throughout the process. Giso, Maartje, Resa, Jürgen, Steffan, Christoph, Mattes, Micha, Marcos, Matthias for all the shared office hours.

ENRICH candidates and tutors, thank you for all the memories and trips shared together.

Frank, for setting me on the right path.

Joanna, no words are enough, you just make a better person.

To my family and my friends, always there and together.

Table of Contents

Chapter 1

General Introduction

Hearing disability is more and more present in today's society. According to WHO, "it is estimated that by 2050 over 700 million people will have disabling hearing loss"[1]. If left untreated, hearing loss may lead to social exclusion, isolation, and a shorter life expectancy (Rutherford et al., 2018; Tareque et al., 2019).

Hearing loss is measured and diagnosed with standard audiological tests, e.g., the audiogram, which examines the physical properties of the hearing system and quantifies the hearing loss in decibels. A hearing disability is diagnosed with an audiogram when an individual has a hearing loss of more than 20 dB HL and it is considered "disabling" when the hearing loss is higher than 35 dB HL in the better hearing ear. Nevertheless, standard clinical tests do not capture the full complexity of the hearing capabilities of an individual. Hearing is a much more complex process, it involves tasks such as separating target and background sounds, understanding distorted speech, and/or being able to lipread the speaker. Additionally, most hearing aid fitting procedures rely on the same measures as the audiogram, which capture very precisely the properties of the hearing system but are not able to measure real-life performance.

[1] World Health Organization (WHO). Deafness and hearing loss. https://www.who.int/health-topics/hearing-loss. Last accessed: 11th January 2023.

The process of fitting a patient with a hearing aid usually requires several visits to the audiologist (Dillon, 2012). During this process, the patient adapts gradually to the new listening experience, reports the issues he/she is experiencing with the hearing aid, and the audiologist adjusts the device accordingly. One of the main concerns during this procedure is that the user will stop using hearing aids. According to Hartley et al. (2010), about one out of four patients do not use their prescribed hearing aid. Furthermore, the vast majority (80%) of elderly people that could benefit from wearing a hearing aid, do not use one, as reported by McCormack & Fortnum (2013). Therefore, the first weeks of a new hearing aid recipient are crucial, and finding the right hearing aid configuration is most essential.

The scope study by McCormack & Fortnum (2013) identified that the most significant reasons for non-use of hearing aids are related to “hearing aid value/speech clarity” and “fit and comfort of the hearing aid”. The most common issues related to “hearing aid value” were that the hearing aid does not help in noisy situations, it provides poor benefit, and that the sound quality is poor. Current hearing tests and fitting procedures are not able to identify these issues, as they are focused on the physical properties of the hearing rehabilitation rather than the real-life “hearing aid value”. For this reason, the assessment in the clinic should be as complete as possible, it should capture these problems as early as possible, and it should provide a full picture of the hearing capabilities of an individual in a real-life scenario.

The term "ecological validity" is often used to describe how a test in a laboratory or clinical experiment can reflect behaviors and functions in real-life. In the sixth Eriksholm Workshop (Keidser et al., 2020), the consensual definition for this term in hearing research was: "In hearing science, ecological

validity refers to the degree to which research findings reflect real-life hearing-related function, activity, or participation." Increasing the ecological validity of clinical evaluations might be the key for a better assessment of hearing impairment and a more satisfactory hearing rehabilitation. In Keidser et al. (2020), integrating new and emerging technologies, such as virtual reality, is one of the key areas where research should focus to improve ecological validity. One reasonable assumption is that when the realism of a laboratory experiment increases, the ecological validity of the result increases too. These emerging technologies give the opportunity to create realistic immersive simulations to evaluate problems that appear in complex scenes, e.g., a virtual reality simulation of a conversation inside a restaurant. They can target issues that appear in real-life, such as hearing aids not performing well in noisy environments. Using virtual reality technologies in audiological clinics is possible, as they have become available for the general public in recent years.

Before introducing such technologies in clinical experiments, it is necessary to validate them in the context of hearing research and audiology. Some of the research questions (RQ) that need to be answered are:

- RQ1. What are the effects of visual cues and virtual reality in psychoacoustic experiments?
- RQ2. Are these virtual reality simulations ecologically valid, i.e., the results of audiovisual experiments reflect better real-life hearing-related activities?
- RQ3. Are these technologies and test procedures ready for clinical procedures? Are the target populations willing to accept them?

1.1. Aim and Scope of the Thesis

The work presented here aims at increasing the realism of hearing laboratory experiments, and at validating and introducing immersive technologies in hearing research. By validating such technologies and new procedures, clinics might adopt them in their hearing centers to provide a more holistic diagnosis of hearing impairment.

Each chapter of this thesis focuses on a different hearing research topic. These topics were chosen because of their relevance for the hearing impaired. Chapter 2 is about speech intelligibility, as a common reason for going to the audiologist is the difficulty of hearing and understanding other people. Chapter 3 focuses on loudness perception, as a major complaint of hearing aid users is that the hearing aid reaches uncomfortably high loudness levels on some occasions (Anderson et al., 2018; Dillon, 2012). Chapter 4 is about preference and acceptance of audiovisual technologies. The acceptance and preference of different audiovisual setups is measured to understand its applicability in clinical environments.

In relation to the proposed research questions, this work tries to understand the effects of visual cues (RQ1). The experiments are designed with the goal to increase ecological validity (RQ2) instead of focusing on basic science and audiovisual perception/integration. This work used several audiovisual technologies and compared them to understand their effects and the acceptance of target populations (RQ3). The following paragraphs explain the relationship between the chapters and the research questions proposed. These research questions are used as a guideline for the design of the experiments in each chapter.

Regarding the RQ1 (effect of visual cues), all experiments in this thesis compare audio-only stimuli to audiovisual stimuli. In Chapter 2, speech intelligibility is measured in an experiment with audio-only and audiovisual speech. Acoustic speech is presented through headphones and visual speech, when available, is shown on a computer monitor. In Chapter 3, the loudness judgments of vehicles are evaluated in different laboratory conditions. The audio-only condition uses a single loudspeaker, the first audiovisual condition uses stereo loudspeakers and a computer monitor, and the second audiovisual condition uses stereo loudspeakers and 360º videos with a head-mounted display.

RQ2 (ecological validity) serves as a guideline for the design of the experiments in this thesis, as currently there are no formal ways to evaluate and quantify the ecological validity of a study (Keidser et al., 2020). In this thesis, the ecological validity scale depicted in Figure 1.1 is used as a guideline. The scale shows that an increase in ecological validity is related to an increase in the realism of a laboratory simulation. On the left and in yellow, the laboratory with audio-only stimuli is represented. In the middle and in blue, the laboratory with audiovisual and immersive visual cues appears. Different audiovisual technologies are represented on the scale, with the most immersive technologies being on the right side closer to real-life experiences. On the right and in green, real-life situations and controlled field experiments are represented.

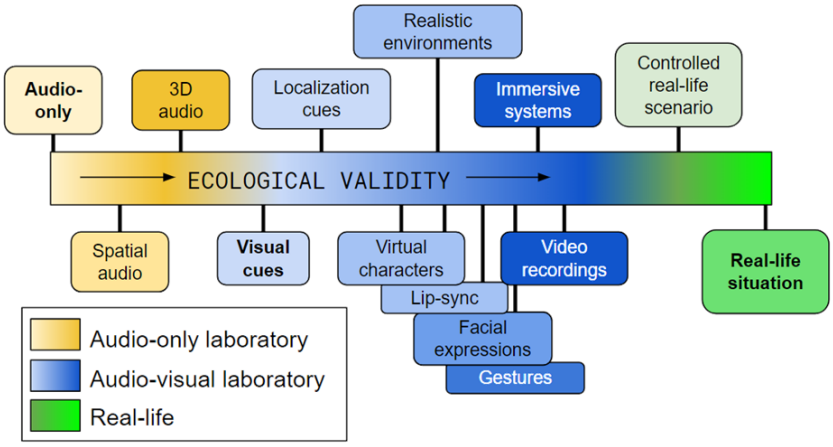


Figure 1.1. This graph represents the increasing complexity of a laboratory setup until reaching a real-life scenario. The goal of these methodologies is to increase the realism, i.e., the ecological validity, of the hearing tests.

In Chapter 2, testing speech perception with audiovisual stimuli is considered as more ecologically valid than using audio-only speech, at least regarding face-to-face communication. Although body gestures and facial expressions can convey as much information as acoustic speech, in Chapter 2 only lip movements are considered to preserve the structure of the audiological test which the experimental design is based on. Chapter 3 includes a comparison between field and laboratory loudness perception. The field measurement, with a hybrid format according to Keidser et al. (2020), is considered the ecologically valid reference for the laboratory measurements in that chapter. The aim of the experiment is to discover how immersive the laboratory setup needs to be to obtain the same loudness ratings as in the field measurement.

RQ3 is concerned about the applicability of the research done in the thesis. As much as discovering the effects of visual cues is relevant, the applicability of

the research into the clinic and audiological practices is crucial. The work presented in Chapter 2 is specially targeted at creating a speech intelligibility test that is applicable in the clinic. Visual cues are added to a standardized audio-only speech intelligibility test, i.e., the test procedure and audio-only stimuli are preserved. Because the audio-only test is currently used in clinical environments, the integration in the clinic of its audiovisual version is straight forward. The experiment includes young normal-hearing participants. The loudness perception experiment in Chapter 3 aims at raising awareness about the disparities between loudness perception in the laboratory and in the field. By identifying these differences, if any found, the research remarks on the importance of measuring loudness perception with more realistic stimuli. The audiovisual setups in Chapter 3 are specifically chosen because they are easy to implement in the clinic. The experiment included normal-hearing and hearing-impaired participants. Chapter 4 compares the preferences for different audiovisual technologies. The experiment specifically asks the participants which of the setups would they prefer, and which ones would they reject in a clinical setup. Young normal-hearing, older normal-hearing, and older hearing-impaired participants were recruited for the experiment.

1.2. Summary of Results

In Chapter 2 it was found that when including visual cues, speech is easier to understand, as expected. The novelty of the work in Chapter 2 is that the visual cues are recorded and added to existing acoustic speech recordings. The visual speech was recorded while listening to the original speech, thus creating a complementary visual signal to the speech material. Of course, small asynchronies exist between the visual and acoustic speech with such procedure, but those are below the perceivable range. Such method to create visual speech

on top of audio-only speech can be extend to other speech material, therefore permitting to create audiovisual speech material while keeping the validity and reproducibility of the audio-only speech material.

The experiment in Chapter 2 validates the dubbed speech material by finding an increase in speech intelligibility in respect to audio-only speech. In Figure 1.2 the differences between audio-only and audiovisual speech reception can be seen. In the experiment, the acoustic levels changed adaptively to make the participant understand 80% of the content. The vertical axis represents the acoustic level, either the signal-to-noise ratio (SNR) in noisy conditions or the sound pressure level (SPL) in quiet conditions. The different conditions are distributed in the horizontal axis. A difference of 5 dB SNR was found for the speech-in-noise conditions and a difference of 7 dB SPL was found for the speech-in-quiet conditions.

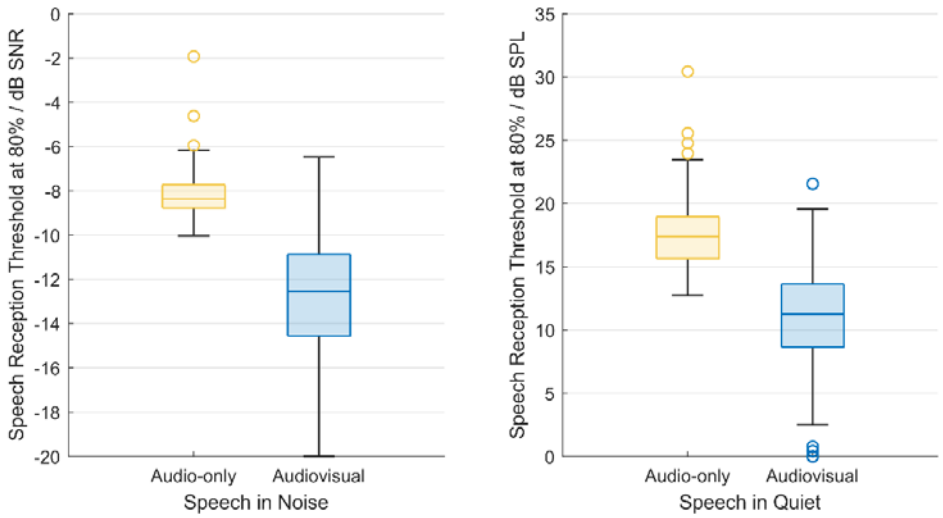


Figure 1.2. Speech reception thresholds (SRTs) of Chapter 2 in audio-only (yellow) and audiovisual (blue) conditions. Low SRTs indicate understanding 80% of the words at challenging listening conditions. In all conditions participants were able to understand

80% of the speech. When the speaker was visible, participants were able to understand speech in challenging acoustic conditions.

Chapter 3 investigates loudness perception differences between the field and different laboratory setups. The stimuli to rate are the driving actions of four urban vehicles. The participants tended to rate the driving actions louder when the realism of the laboratory stimuli decreased. The ratings of the most simplistic laboratory condition, an audio-only setup with one loudspeaker, yielded the most differences in comparison to the field loudness ratings. Following the tendency, the laboratory condition that used virtual reality and stereo audio was the one that achieved loudness ratings closer to the field ratings. In Figure 1.3 the differences between the loudness ratings in different conditions are shown. In the vertical axis the loudness scale is shown. In the horizontal axis the conditions are sorted from less realistic to more realistic following the ecological validity scale shown in Figure 1.3. No differences were found between hearing types (normal-hearing and hearing-impaired).

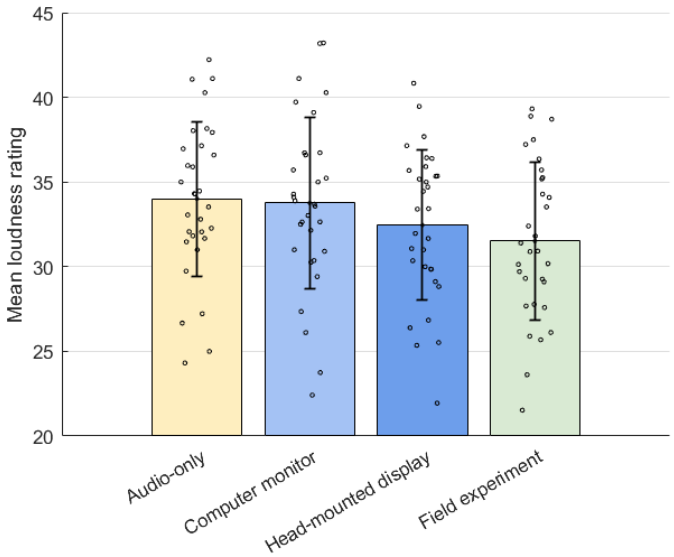


Figure 1.3. Average loudness ratings of the experiment in Chapter 3. The vertical axis does not show the whole loudness response alternative range.

Chapter 4 compares audiovisual technologies and setups. One relevant finding is that all participants (young, older and older with hearing impairment) were willing to use immersive audiovisual setups (head-mounted displays) for hearing experiments. This finding indicates that such technologies have a promising future in clinics. The most realistic visual cues (video recordings) were preferred, and the curved screen was preferred over the head-mounted display only by the older normal-hearing participants. The work in this chapter validated the use of audiovisual technologies in these kind of hearing experiments with older normal-hearing and older hearing-impaired participants. Figure 1.4 shows the laboratory setup presented in chapter 4.

Figure 1.4. Picture of the laboratory setup in Chapter 4. A conversation between four talkers was projected into a curved screen using video recordings, virtual characters, and no visual cues. A replica of the laboratory was recreated in a virtual reality simulation

for the head-mounted display, which showed the conversation in the same conditions as with the curved screen.

References

Anderson, S., Gordon-Salant, S., & Dubno, J. R. (2018). Hearing and Aging Effects on Speech Understanding: Challenges and Solutions. *Acoustics Today*, *14*(4). https://doi.org/10.1121/at.2018.14.4.12

Dillon, H. (2012). *Hearing Aids* (2nd Edition). Boomerang Press.

Hartley, D., Rochtchina, E., Newall, P., Golding, M., & Mitchell, P. (2010). Use of hearing aids and assistive listening devices in an older australian population. *Journal of the American Academy of Audiology*, *21*(10). https://doi.org/10.3766/jaaa.21.10.4

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It. *Ear and Hearing*, *41*. https://doi.org/10.1097/AUD.0000000000000944

McCormack, A., & Fortnum, H. (2013). Why do people fitted with hearing aids not wear them? *International Journal of Audiology*, *52*(5). https://doi.org/10.3109/14992027.2013.769066

Rutherford, B. R., Brewster, K., Golub, J. S., Kim, A. H., & Roose, S. P. (2018). Sensation and psychiatry: Linking age-related hearing loss to late-life depression and cognitive decline. *American Journal of Psychiatry*, *175*(3). https://doi.org/10.1176/appi.ajp.2017.17040423

Tareque, M. I., Chan, A., Saito, Y., Ma, S., & Malhotra, R. (2019). The Impact of Self-Reported Vision and Hearing Impairment on Health Expectancy. *Journal of the American Geriatrics Society*, *67*(12). https://doi.org/10.1111/jgs.16086

Chapter 2

Development and evaluation of video recordings for the OLSA matrix sentence test

Published in:

Llorach, G., Kirschner, F., Grimm, G., Zokoll, M. A., Wagener, K. C., & Hohmann, V. (2022). Development and evaluation of video recordings for the OLSA matrix sentence test. *International Journal of Audiology*, *61*(4). https://doi.org/10.1080/14992027.2021.1930205

Abstract

Objective: The aim was to create and validate an audiovisual version of the German Matrix Sentence Test, which uses the existing audio-only speech material.

Design: Video recordings were recorded and dubbed with the audio of the existing German matrix sentence test (MST). The current study evaluates the MST in conditions including audio and visual modalities, speech in quiet and noise, and open and closed-set response formats.

Sample: 1 female talker recorded repetitions of the German MST sentences. 28 young normal-hearing participants completed the evaluation study.

Results: The audiovisual benefit in quiet was 7.0 dB in sound pressure level (SPL). In noise, the audiovisual benefit was 4.9 dB in signal-to-noise ratio (SNR). Speechreading scores ranged from 0% to 84% speech reception in visual-only sentences (mean = 50%). Audiovisual speech reception thresholds (SRTs) had a larger standard deviation than audio-only SRTs. Audiovisual SRTs improved successively with increasing number of lists performed. The final video recordings are openly available.

Conclusions: The video material achieved similar results as the literature in terms of gross speech intelligibility, despite the inherent asynchronies of dubbing. Due to ceiling effects, adaptive procedures targeting 80% intelligibility should be used. At least one or two training lists should be performed.

2.1. INTRODUCTION

Speech audiometry is an essential element in audiology (Sanchez Lopez et al., 2018; Talbott & Larson, 1983). It assesses the ability to understand speech acoustically, which is crucial for human communication. The matrix sentence test (MST) (K. C. Wagener & Brand, 2005) is a well-established method in speech audiometry, and it exists in several languages (Kollmeier et al., 2015). MSTs use sentences of 5 words with a "noun - verb - number - adjective - object" structure. There are 10 possible words for each word category (e.g., 10 nouns, 10 verbs, etc.); these are combined to create semantically unpredictable, syntactically correct sentences. Lists of 20 sentences are commonly used to test speech intelligibility.

Although speech can be understood through sounds only, it is a multimodal process. Being able to see the speaker provides additional cues such as lip movements, which make speech much easier to understand (Sumby & Pollack, 1954). Audiovisual speech perception has been mentioned as a predictor of real-world hearing disability (Corthals et al., 1997) but it is usually not considered in audiometry (Woodhouse et al., 2009). Visual information supports speech intelligibility, particularly severely impaired listeners are relying on visual information in adverse listening conditions (Schreitmüller et al., 2018). The MST is also intended as a speech test for severely impaired listeners, therefore an audiovisual version is an important extension for its applicability. Nevertheless, audiovisual (or auditory-visual) MSTs with video recordings have only been developed in Malay, New Zealander English, and Dutch (Jamaluddin, 2016; Trounson, 2012; van de Rijt et al., 2019).

The ability to speechread (most commonly known as lipreading) plays a key role in audiovisual speech tests. In particular, audiovisual MSTs are highly affected by speechreading ability. In the Malay MST (Jamaluddin, 2016) young, normal-hearing participants scored from 25% to 85% speech reception just by speechreading, i.e., in the visual-only condition. Such visual-only scores indicate that participants are able to understand speech without any acoustic cues. This means that there is a ceiling effect in the audiovisual MSTs: even if speech is completely masked by noise and not heard, participants achieve their visual-only score.

Recording and validating an MST is quite an extensive undertaking: selection of the phonetically balanced speech material, recording of the speech, cutting and processing of the sound files, making each word equally intelligible to the others, evaluation, and validation (Kollmeier et al., 2015).

In order to reduce cost and effort in the creation of an audiovisual MST from scratch, existing audio-only MST can be reused. Because audio-only MSTs already exist and have been used extensively, it is reasonable to reuse the audio material in audiovisual tests. New audio recordings cannot be compared directly to other recordings of the same language, as the speaker influences the intelligibility of the MST (up to 6 dB differences between talkers) (Hochmuth et al., 2015). Reusing the audio material ensures validity across studies, and saves time and effort. If the audio recordings are newly created, they need to be optimized to allow for a steep intelligibility function (a prerequisite for an accurate test), which includes measuring the intelligibility functions for each word of the test in a large number of participants. This would multiply the effort in comparison to producing dubbed videos.

One approach that has been proposed uses virtual characters with lip-synchronization together with existing audio-only speech tests (Devesse et al., 2018; Grimm et al., 2019; Schreitmüller et al., 2018). The advantage of virtual characters is that they can be set in different configurations with relatively little effort (Llorach et al., 2018). The proposed approach in this paper is to create video recordings dubbed with existing audio for speech tests. A video recording usually provides better quality and realism than a virtual character. Nevertheless, asynchronies between the audio and the video have to be kept below 45ms (audio ahead) and 200ms (audio delayed) in order to pass unnoticed (Başkent & Bazo, 2011) and not affect speech intelligibility (Grant et al., 2003). Additionally, further considerations must be taken into account, such as the head movements and facial expressions of the speaker (Jamaluddin, 2016).

One of the advantages of MSTs is that the sentences are unpredictable and there are too many word combinations to be memorized, so consecutive tests in different conditions can be carried out. Nevertheless, the simple sentence structure and the limited number of words enable participants to learn and improve their results. This training effect has already been shown in audio-only MSTs (Ahrlich, 2013; K. Wagener et al., 1999) and is particularly noticeable in the first list of 20 sentences, where differences in SRTs of about 1 dB are expected. After 2-4 lists, there is usually an absolute improvement of 2 dB, and the training effects in the following lists are quite small. In audiovisual MSTs, it is expected that participants further improve their SRTs by becoming familiar with the speaker and the visual material (Lander & Davies, 2008) and because training effects have been found to be stronger in audiovisual speech (Lidestam et al., 2014).

Another factor to take into account is the response format of the MST. After hearing a sentence, participants either repeat what they heard (open-set response format) or select the answers from all possible words (closed-set response format). In the open-set format, a researcher must be present in order to assess whether the answer is correct, while in the closed-set format, participants can do the test by themselves. The closed-set format may give participants an advantage, since they are provided with a list of all possible words; in fact, SRTs have been found to be lower with closed-set type in some MSTs (Hochmuth et al., 2012; Puglisi et al., 2014), although not for German and other languages (Kollmeier et al., 2015). Whether such effects appear in audiovisual MSTs has not yet been investigated.

In this work we created an audiovisual version of the female German MST (AV-OLSAf). We recorded videos with a female speaker, dubbed them with the original sentences of the female speaker (Ahrlich, 2013; K. C. Wagener et al., 2014) and evaluated the material. Our first contribution is the methodology for producing the dubbed videos and getting the best synchronized video recordings. The final video recordings for the AV-OLSAf can be found in Llorach et al., 2020. Our second contribution is the evaluation of the AV-OLSAf with normal-hearing listeners in different conditions: we show the audiovisual training effects in the open-set and closed-set responses; we discuss the speechreading scores and the effects of speechreading in the audiovisual SRTs; and we compare the audio-only and audiovisual SRTs in noise and in quiet conditions. To conclude, we discuss the implications and recommendations for using the AV-OLSAf.

2.2. Method

2.2.1. Recording the Video Material

Although in theory there are 100,000 possible sentences (5 word categories with 10 words per category, Table 2.1), the female OLSA uses only 150 predetermined sentences. This relatively small number of sentences permitted us to record videos of the spoken sentences in a single afternoon. We were able to recruit the same speaker that recorded the audio-only version of the German female MST (OLSA) (Ahrlich, 2013; K. C. Wagener et al., 2014). She was a speech therapist and a singer. During the recording session, the speaker had to speak the sentences simultaneously while hearing them through an earphone on the right ear. Each sentence was played five times consecutively. Three short "beep" signals were given before each repetition started. The first repetition was used as a reference: the speaker was to listen only in order to know what sentence was coming. In the remaining 4 repetitions, she was to speak simultaneously while hearing the sentence.

The videos of the female speaker were recorded in the studio of the Media Technology and Production of the CvO University of Oldenburg. The available lights of the studio were set up to achieve a homogeneous illumination of the face and of the background green chroma key. The videos were recorded with a Sony α7S II camera at 50pfs / full HD, and a condenser microphone in front of the speaker at the height of the knees. The speech was recorded in one channel with a 48 kHz sampling rate and a 16 bit linear pulse-code modulation (LPCM) sample format. An image sample of the final video recordings is shown in Figure 2.1.

Table 2.1. Set of words used in the German Matrix Sentence Test. The sentences are combinations of 5 words from different categories, e.g., "*Doris malt neun nasse Sessel*" or "Nina bekommt vier rote Schuhe". The order shown here is the same as it was shown to the participants in the closed-set response format.

Noun	Verb	Number	Adjective	Object
Britta	bekommt	zwei	alte	Autos
Doris	gewann	drei	große	Bilder
Kerstin	gibt	vier	grüne	Blumen
Nina	hat	fünf	kleine	Dosen
Peter	kauft	sieben	*nasse*	Messer
Stefan	*malt*	acht	rote	Ringe
Tanja	nahm	*neun*	schöne	Schuhe
Thomas	schenkt	elf	schwere	*Sessel*
Ulrich	sieht	zwölf	teure	Steine
Wolfgang	verleiht	achtzehn	weiße	Tassen

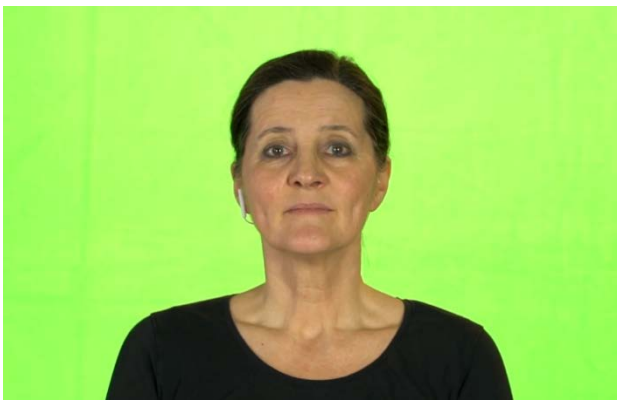


Figure 2.1. Example of a frame of the video material.

A computer was used to reproduce the original OLSA sentences, which at the same time was sending a linear time code (LTC) signal to the second audio channel of the camera. This way, the recorded speech of the session and the

original sentences could be synchronized. The recording session lasted around 2 hours in total.

2.2.2. Selection of the Videos

We manually discarded videos in which the speaker smiled or showed other non-neutral facial expressions. The recorded speech signals were synchronized to the reproduced original sentences using the LTC signal. When dubbing speech, there are inevitable asynchronies: time offsets (words spoken too early or too late) and/or words spoken slower or faster than the original words. As all these asynchronies could happen in one single sentence, we used dynamic time warping (DTW) (Sakoe & Chiba, 1978) to find the best match between the recordings and the original sentences. The DTW quantified the temporal misalignment between the original and recorded sentences. The algorithm compares two temporal signals and provides a warping path. We computed the mel spectrograms of the signals and used them for the DTW function. The mel spectrograms were done using frame windows of 46 ms with a frame shift of 23 ms. An example of the mel spectrograms and the corresponding warping path can be seen in Figure 2.2 and Figure 2.3. Once the warping path was calculated, we used equations 2.1 and 2.2 to compute the asynchrony score:

$$wp_{ij}(n) = DTW(melSpec_i(n), melSpec_j(m)) \tag{2.1}$$

$$async_{ij} = RMS(wp_{ij}(n) - n) \tag{2.2}$$

for $i = 1,2,3,\dots,150$ (original sentence number)

and $j = 1,2,3,4$ (recording nº per original sentence)

where the $melSpec_i(n)$ is the mel spectrogram of the original sentence i, $melSpec_j(n)$ is the mel spectrogram of its corresponding recording (4 recordings

j per sentence i), n is the frame number of the $melSpec_i$, m is the frame number of the $melSpec_j$, $wp_{ij}(n)$is the warping path between the mel spectrograms in frames, $(wp_{ij}(n) - n)$ is the difference in frames, RMS is the root mean square, and $async_{ij}$ is the asynchrony score between the i^{th} original sentence and the j^{th} recording of that sentence. The RMS was used because it represents the asynchrony score over a whole sentence. As our main interest is the speech intelligibility of the whole sentence, we did not consider momentary asynchronies, such as maximum asynchrony, as a measure to choose the best video recording. The asynchrony score can be further expressed in seconds instead of frames, as it represents a temporal difference.

We checked the sensitivity of this measure by comparing each recording to its corresponding original sentence and to the remaining unmatching original sentences (Figure 2.4). For each original sentence, we chose the video recording with the smallest asynchrony score. Of the final best selections, we found three outliers, with asynchrony scores greater than 80 ms, that had to be manually corrected with time offsets. Once corrected, these outliers were shown to 5 normal-hearing participants along with the best-matched sentences; the outliers could not be distinguished from the best-matched sentences and no asynchronies were noticed. We decided that the mean asynchrony score was small enough (˜40 ms) to minimize the perceptual asynchrony/dubbing effects when measuring speech intelligibility with lists of 20 sentences: in Grant et al., 2003), the authors evaluated the speech intelligibility of different timing misalignments with video and audio. According to them, visual asynchronies from -45ms to +200ms are not perceivable and speech recognition does not decline. Therefore, we proceeded with the evaluation of the material. The asynchrony score, maximum asynchronies and asynchrony over time of each

sentence can be found in Table S1 in the supplemental material[2]. The final video recordings can be found in (Llorach, Kirschner, et al., 2020).

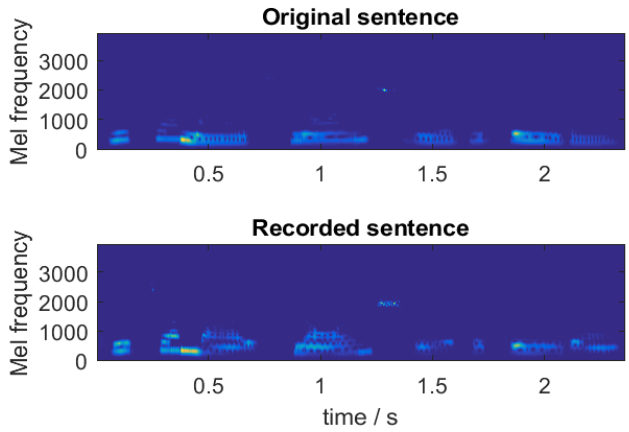


Figure 2.2. Mel spectrogram of original sentence and one of the four recordings of that sentence.

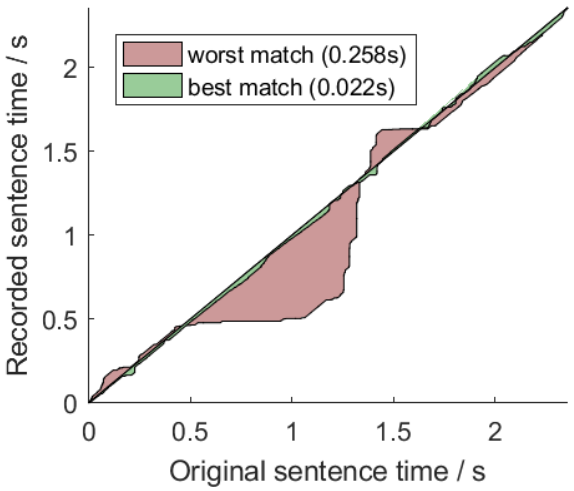


Figure 2.3. Warping path between the original and two recorded sentences. The best match and the worst match are shown. The size of the shaded surface corresponds to the asynchrony score.

[2] Interested readers can access the supplemental material at http://tandfonline.com/doi/suppl.

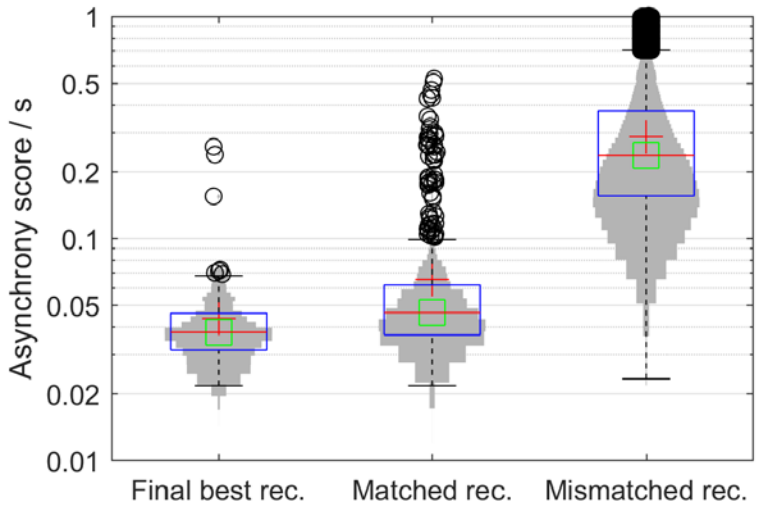


Figure 2.4. Asynchrony scores comparing the original sentences and their best-matched recordings before manual correction of the three outliers (left; 150 scores), the original sentences and all 4 of their recordings (middle; 600 scores) and the original sentences and the mismatched recordings (right; 150x149x4 scores). The vertical axis is on a logarithmic scale. The mean is represented as a red cross and the median as a green square. The outliers are depicted with black circles.

2.2.3. Evaluation of the Audiovisual Material

Participants

28 normal-hearing participants (14 female, 14 male) took part in the evaluation measurements. Their ages ranged from 20 to 29 years (mean age 24.9 years). They had normal or corrected-to-normal vision and their pure tone averages (PTAs) in the better ear were between -5 and 7.5 dB HL (mean -0.31 dB HL). The PTAs were computed using the frequencies 0.5, 1, 2 and 4 kHz. Participants were recruited through the database of the Hörzentrum Oldenburg GmbH and were paid an expense allowance. Permission was granted by the ethics committee of the CvO University of Oldenburg.

Setup

Participants were seated in a chair inside a soundproof booth. The evaluation measurements were done using binaural headphones (Sennheiser HDA 200). A 22'' touchscreen display with full HD (ViewSonic TD2220, ViewSonic Corp. Walnut, CA, USA) was placed in front of the participant within arm's reach at a height of 0.8 meters. The experiment was programmed in Matlab2016b. The videos and original sentences were reproduced with VLC 3.03. The acoustic signal was routed with RME Total Mix with an RME Fireface 400 sound card.

The acoustic levels were calibrated using a sound level meter placed at the approximate head position where participants would be seated. The sound and video reproduction were calibrated for synchronization using an external camera. For this purpose, we reproduced a video with frame numbering together with a LTC signal using the experiment setup. The external camera recorded the display screen of the experiment. The LTC signal was connected directly to the external camera instead of the headphones. Using the recording of the external camera we found a consistent asynchrony of 80 ms, which we corrected by delaying the audio signal in the experiment setup.

Stimuli

The acoustic stimulus was the female version of the German matrix sentence test (OLSA) (Ahrlich, 2013; K. C. Wagener et al., 2014) and the visual stimuli was the best-matched video recording (see Section 2.2). For the conditions with noise, we used continuous test-specific noise (TSN) based on the female speech material. The presentation level of the noise was kept constant at 65 dB SPL. The speech level of the first sentence was 60 dB SPL for conditions with and

without noise. The adaptive procedure used varied the speech presentation level depending on the responses of the participant.

Conditions

There were nine conditions in the experiment (see Table 2.2). Each condition used a list of 20 sentences. The sentences in each list were predefined by the MST. In total, we used 45 different predefined lists. The speech presentation levels were adapted after each sentence in order to reach an individual SRT of 80%, i.e. 4 out of 5 words correctly recognized per sentence. During the open-set response format, participants were asked to repeat orally what they understood after each sentence. In the closed-set response format, participants chose the words they understood from an interface displayed on the touch screen after stimulus presentation. The closed-set interface showed all 50 possible words plus one no-answer option per word category. In the visual-only condition (VONoiseClosed), there was no acoustic speech but only test-specific noise at 65 dB SPL. In this condition, the speech could only be understood through speechreading. For this condition, the percentage of correct words per sentence was averaged over 20 sentences (a list). In all conditions, no feedback was given about correctness of responses.

Table 2.2. Conditions tested for the evaluation and validation of the AV-OLSA.

		Audio-only (AO)	Audiovisual (AV)	Visual-only (VO)
Noise	Closed-set response	**AONoiseClosed**	**AVNoiseClosed**	**VONoiseClosed**
	Open-set response	**AONoiseOpen**	**AVNoiseOpen**	
Quiet	Closed-set response	**AOQuietClosed**	**AVQuietClosed**	-
	Open-set response	**AOQuietOpen**	**AVQuietOpen**	

Adaptive procedure

We chose a SRT of 80% to avoid ceiling effects in audiovisual conditions due to the visual-only contribution, i.e., some participants might be able to understand more than 50% of the content just by speechreading (Jamaluddin, 2016; Trounson, 2012; van de Rijt et al., 2019). The adaptive procedure used in this experiment is described in Brand & Kollmeier, 2002 and in Brand et al., 2011. It is an extended staircase method that changes its step size depending on the responses. The change in the presentation level is done in two stages. The first stage follows the equation presented in Brand & Kollmeier, 2002:

$$\Delta L = -\frac{f(i)\cdot(prev-tar)}{slope} \qquad (2.3)$$

where ΔL is the increment level, *prev* is the current result, *tar* is the target value, and *slope* is set to 0.1 dB^{-1} in this study. The function *f(i)* defines the convergence rate, where *i* is the number of reversals in the presentation level,

i.e. i increases every time the participant goes from being above/below threshold. In our study the current result is the discrimination value of the previous sentence and the target value is 0.8 (80% SRT). The value of *f(i)* is defined by $1.5 \;/\; 1.41^{i}$ and its set to 0.25 for $\boldsymbol{i \leq 6}$. The step size gets smaller when the participant crosses the target value. The second stage is described and examined in Brand et al., 2011). In this second stage, the step size is multiplied by 2 when two conditions are met: the step is a decrement (it lowers the presentation level) and *f(i)* is bigger than 0.5. This last condition is usually met in the first sentences of a list.

$$\Delta L = \begin{cases} 2 \cdot \Delta L & for\ f(i) \geq 0.5\ and\ \Delta L < 0 \\ \Delta L & otherwise \end{cases} \tag{2.4}$$

The final level estimate of a list is computed using a maximum-likelihood method and discrimination function described in (Brand & Kollmeier, 2002).

Training

We added training lists prior to evaluating the nine conditions in order to assess the training effects of the AV-OLSA. We also tested the participants in two different sessions (test, retest). In the first session, 4 audiovisual lists were presented in noise (80 sentences in total). Participants were randomly assigned to do the 4 training lists in open-set or closed-set formats (AVNoiseClosed or AVNoiseOpen); 13 participants completed the training in the closed-set format and 15 participants in the open-set format. In the second session, the training was a single list with the same format as the first session (20 sentences in AVNoiseClosed or AVNoiseOpen).

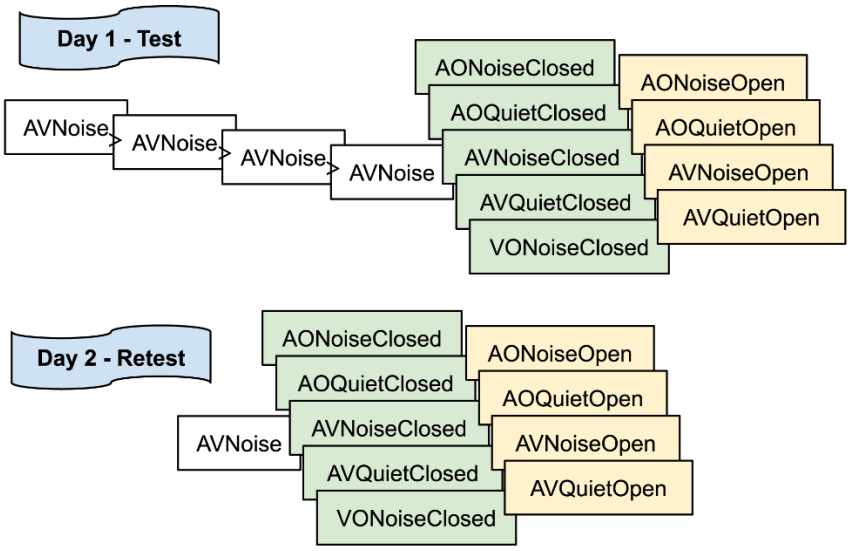


Figure 2.5. Ordering of the lists in the test and retest sessions. Conditions stacked in columns (green and yellow) were pseudo-randomized within the column. If the participants were trained in AVNoise with the open-set format, they performed the open-set format lists before the closed-set lists; if they were trained with the closed-set format, they proceeded with the closed-set format lists before doing the open-set lists.

Procedure

For the test lists, participants started with the same response format (open-set or closed-set) as in the training. Next, they did the conditions with the opposite response format. The conditions of one response format were presented in pseudo-randomized order (Figure 2.5). On the retest session, participants performed a training list with the same response format as on training lists of the first session; then they continued with the conditions with that same format before doing the ones with the other format, as on the test session. The test and retest sessions were temporally spaced from one day to two weeks.

2.3. Results

For each list, a final level estimate was computed as described in Section 2.3.5. This value, i.e. the SRT at 80%, was expressed in dB SNR for the conditions in noise and in dB SPL for the conditions in quiet. For the VONoiseClosed lists it was different: the percentage of words understood over all 20 sentences was computed (i.e., the speechreading score). For each participant there were 5 audiovisual training lists, 4 in the first session and 1 in the second session, and 18 test lists, 9 in each session (see Figure 2.5). For the analysis of the results, we removed an outlier of +9 dB SNR belonging to an AONoiseClosed list of the first session (test).

2.3.1. Training Effects

In general, audiovisual SRTs tended to improve across lists. Figure 2.6 shows the mean SRTs during the training lists, and the test and retest for the audiovisual in noise condition. In the aforementioned figure, the participants and its SRTs are separated in two groups depending on the training response format (open vs closed). On average, participants improved their SRTs by -1.6 dB SNR on their third training list. The total improvement between the first training list and the test list was -2.9 dB SNR. On the second session, participants retained the same SRT scores in the training as in their last list of the first session. SRTs improved further on the list of the retest session, by -3.8 dB SNR relative to the first training list of the first session. Figure 2.6 shows that there was a consistent difference of ~1.8 dB SNR between the mean SRTs of the open-set and closed-set lists.

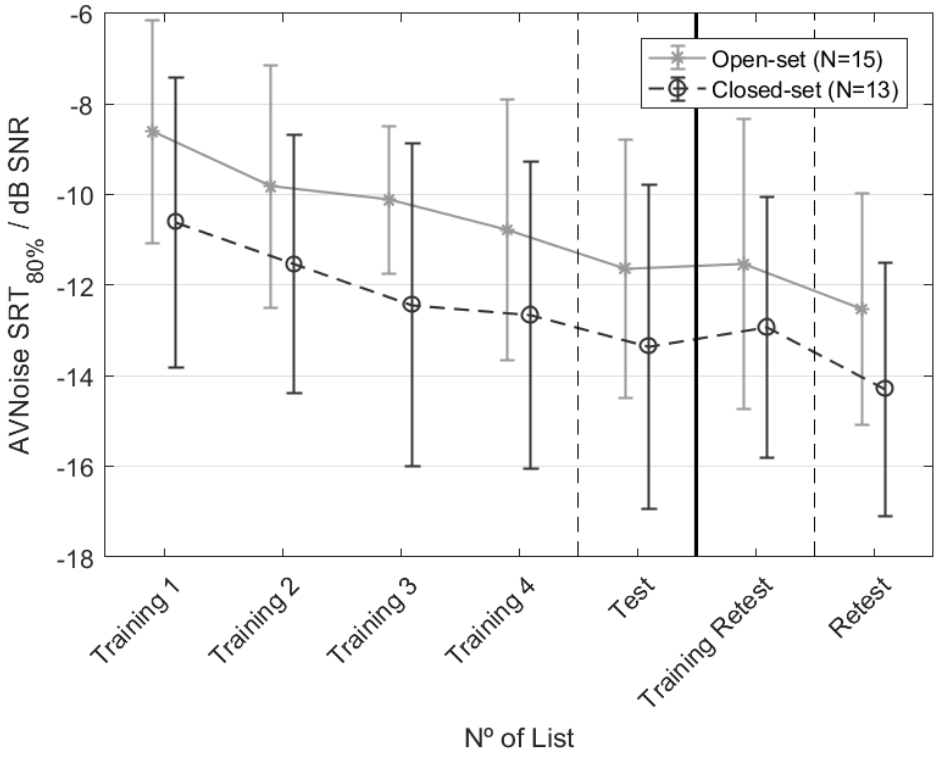


Figure 2.6. Audiovisual training effects. The average and the standard deviation of SRTs over groups are shown. The black dashed line with circles shows the SRTs of the 13 participants that did the training in closed response format. The continuous grey line with whiskers shows the SRTs of the 15 participants that did the training in open response format. It should be noted that, due to the other measurement conditions, there could be up to 4 lists in between the Training 4 and Test lists and between Training Retest and Retest lists.

A repeated-measures ANOVA was performed with response format as the between-subjects factor (open vs. closed) and position within the initial training lists (1st, 2nd, 3rd, and 4th training list) as the within-subjects factor. The dependent variable was the SRT. The sphericity assumption had not been violated according to Mauchly's test ($\chi^2(5) = 2.33$; $p = 0.80$). A significant

main effect was found for the within-subjects factor (training list order) ($F(3, 78) = 10.96$; $p < 0.001$). No significant effect was found for the between-subjects factor (response format in the training lists) ($F(1, 26) = 4.17$; $p = 0.052$), although it was close to being significant. No significant interaction was found between the training list's position and the response format ($F(3, 78) = 0.21$; $p = 0.82$). Multiple comparisons with Bonferroni corrections showed that the SRT of the first list was significantly different from the SRTs of the other three. The SRTs of the second, third and fourth list did not differ significantly.

2.3.2. Audio-only and Audiovisual SRTs

Mean SRTs and standard deviations of the lists for the different conditions are shown in Table 2.3. In the table, test and retest SRTs are grouped together per condition. The average SRT differences between audio-only and audiovisual lists were 5.0 dB SNR for speech in noise and 7.0 dB SPL in quiet. The listeners' PTAs were not significantly correlated with the audio-only in quiet scores (Pearson's $r = 0.15$, $p = 0.11$).

Table 2.3. Mean audio-only and audiovisual SRTs and between-subjects standard deviations in the test and retest sessions (56 scores per cell).

	Mean SRT / dB SNR		**Mean SRT / dB SPL**
AONoiseClosed	-8.2 (0.9)	**AOQuietClosed**	17.6 (3.2)
AONoiseOpen	-8.2 (1.1)	**AOQuietOpen**	17.8 (2.4)
AVNoiseClosed	-13.4 (3.2)	**AVQuietClosed**	10.9 (4.4)
AVNoiseOpen	-12.9 (3.4)	**AVQuietOpen**	10.5 (4.6)

2.3.3. Ceiling Effects

Participants reached SNRs below -20 dB and speech presentation levels below 0 dB SPL (no sound pressure) in the audiovisual conditions. At these levels, there is no contribution of acoustic information to speech reception: the speech detection threshold for the female OLSA is around -16.9 dB SNR in audio-only tests with TSN (Schubotz et al., 2016), a threshold that can be theoretically lowered by around -3 dB when adding visual speech (Bernstein et al., 2004). Therefore, below these thresholds (-20 dB SNR and 0 dB SPL), participants used only visual speech in this experiment, i.e., they were speechreading. In consequence, the scores below these thresholds do not represent audiovisual speech perception, but rather visual-only. Figure 2.7 shows that during the adaptive procedure, participants could reach levels where there was no acoustic contribution.

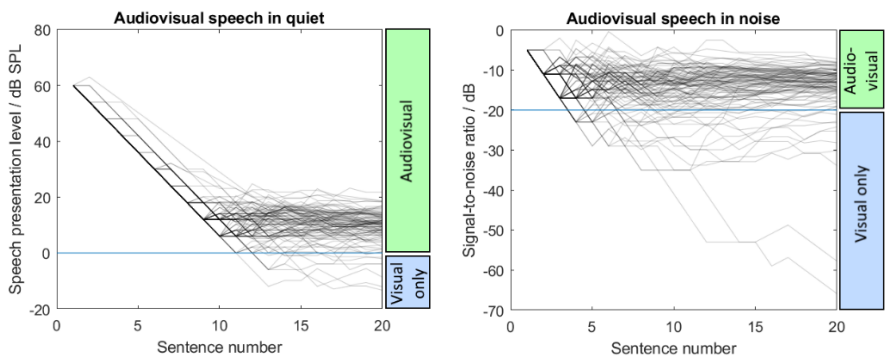


Figure 2.7. Adaptive SNRs and speech presentation levels for AVQuiet (left) and AVNoise (right) conditions. The adaptive procedure changed the speech levels to reach 80% intelligibility. Below the blue horizontal line, participants understood speech using only visual cues. Each line shows a single list, adding up to 4 lines per participant in each subfigure.

For the analysis of the data, we decided to limit the values that were below the acoustic speech detection thresholds, as they were not representative of audiovisual speech reception. In total, 18 out of 364 SRTs of audiovisual lists (5%) were modified by limiting them to -20 dB SNR for speech in noise and 0 dB SPL for speech in quiet. We decided to include these scores as they were representing the best speechreading scores. The lists affected had varied conditions: of the 18 lists, 3 were training lists, 5 AudiovisualNoiseOpen, 5 AudiovisualNoiseClosed, 3 AudovisualQuietOpen, and 2 AudiovisualQuietClosed. Of the 28 participants, 6 were able to go below the speech detection thresholds.

2.3.4. Speechreading and Audiovisual Benefit

Participants had a wide range of speechreading abilities. The individual VONoiseClosed scores ranged from 0 to 84% intelligibility, had an average of 50% and a standard deviation of 21.4%. Figure 2.8 shows the distribution of the visual-only scores. There was an average intelligibility improvement of 6.1% in the retest over the test session, although not all participants improved their scores.

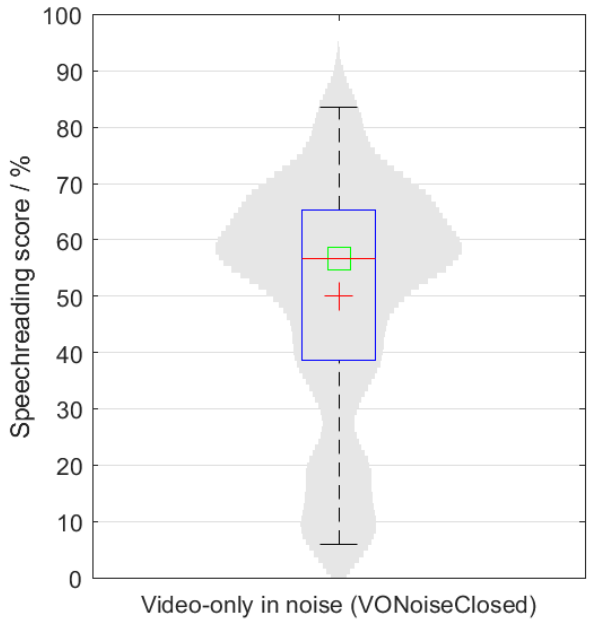


Figure 2.8. Boxplot and distribution of the speechreading scores. In this figure, each participant has a data point: the average word scoring percentage over 40 sentences. The mean and the median are represented as a red cross and a green square, respectively.

Speechreading scores were correlated with the audiovisual benefit (i.e., the SRT difference between audiovisual and audio-only condition). This correlation can be seen in Figure 2.9, where the visual-only scores are plotted against the individual SRT benefits in different conditions. The Pearson's r correlation scores between the speechreading scores (VONoiseClosed) and the audiovisual benefits were -0.76 ($p<0.001$) for AVNoiseClosed minus AONoiseClosed, -0.69 ($p<0.001$) for AVNoiseOpen minus AONoiseOpen, -0.65 ($p<0.001$) for AVQuietClosed minus AOQuietClosed, and -0.65 ($p<0.001$) for AVQuietOpen minus AOQuietOpen. Participants that were good speechreaders gained more from having visual information in the audiovisual lists. Whether participants were trained in open-set or closed-set formats did not make any difference for the audiovisual benefit.

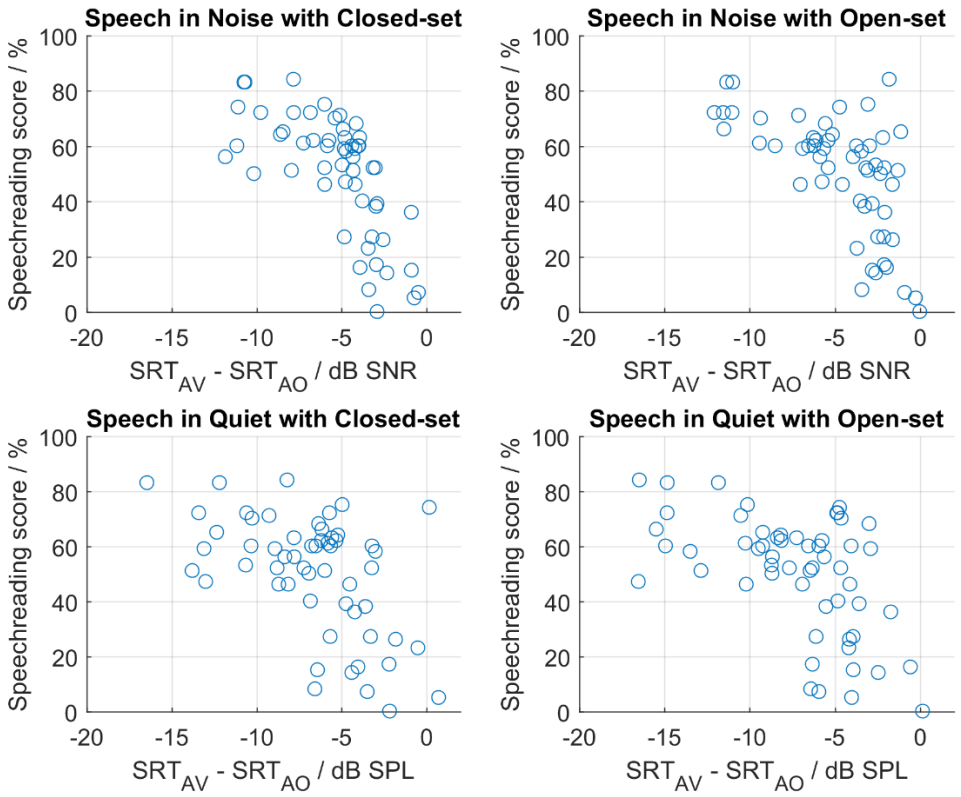


Figure 2.9. Speechreading scores (VONoiseClosed) shown against the audiovisual benefit of each participant; each participant has two circles per plot for test and retest lists. Top left: audiovisual benefit in noise with closed-set response. This condition was the most similar to the visual-only condition, as both had noise and a closed-set format. Top right: audiovisual benefit in noise with open-set format. Bottom: audiovisual benefit in quiet with closed-set (left) and open-set formats (right).

2.3.5. Test-retest differences

The within-subject and the between-subject standard deviations of the SRTs are shown in Figure 2.10. The standard deviations of the within-subject differences (test minus retest) are shown as gray bars. The between-subjects standard deviations are shown as white bars. The 2σ criterion is shown as a

thick black line. The 2σ criterion represents the threshold where it is possible to differentiate significantly between individuals: if the between-subject standard deviation is higher than the double of the within-subject standard deviation, i.e., the 2σ criterion, it means that it is possible to differentiate significantly between subjects (K. C. Wagener & Brand, 2005). None of the conditions but the VONoise exceeded the 2σ criterion. The audiovisual conditions had a higher within-subject and between-subject variability in comparison to their respective audio-only conditions.

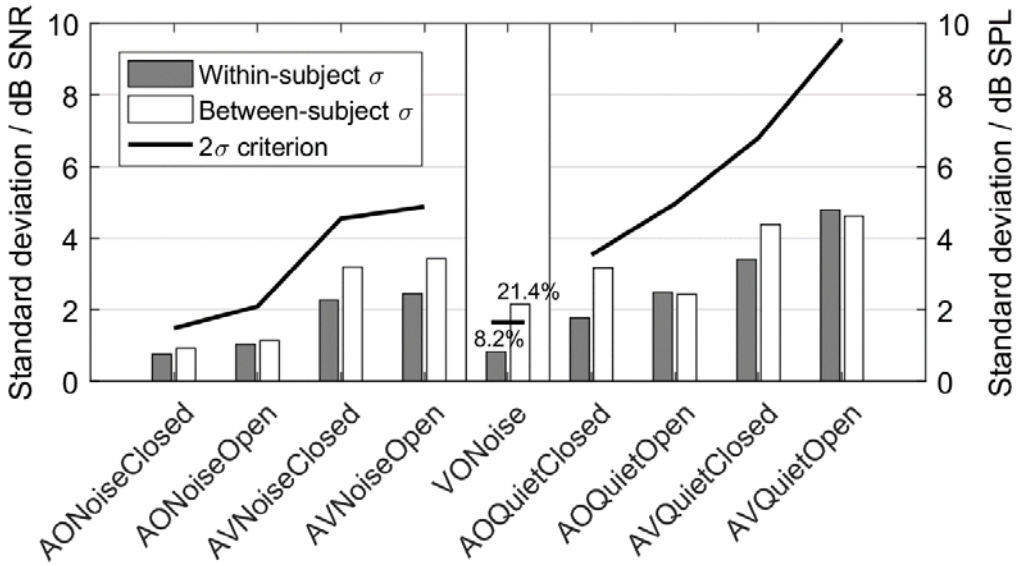


Figure 2.10. Within-subject (gray bars) and between-subject (white bars) standard deviations for all conditions. The 2σ criterion is indicated as a thick black line. On the left, STDs of speech in noise conditions expressed in dB SNR; on the middle, STDs of the speechreading scores expressed in percentage; and on the right, STDs of speech in quiet expressed in dB SPL.

2.4. Discussion

2.4.1. Validity of the Video Material

The audio-only and audiovisual scores found were similar to those expected based on the literature. A difference of 3 dB between audio-only and audiovisual scores was reported previously (van de Rijt et al., 2019), whereas we found a difference of more than 5 dB in the equivalent conditions. This difference could arise from the specific speaker, as some speakers are easier to speechread than others (Bench et al., 1995), or from language differences (Kollmeier et al., 2015).

The results of the audiovisual MST were in concordance with the literature, thus validating the video material for measuring speech reception thresholds using lists of 20 sentences. Nevertheless, due to the inherent dubbing asynchronies, the material presented here might not be suitable for investigating fine-grained effects of audiovisual interactions. Audiovisual asynchronies can detriment speech perception (Grant et al., 2003), but we believe that these asynchronies did not affect severely the results. In another publication (Llorach & Hohmann, 2019), the data of this study was analyzed on a word level. Speech intelligibility detriments due to dubbing asynchronies were not looked at, but it was shown that if a word was harder to understand in the audiovisual version it was because this word was hard to speechread. In other words, audiovisual benefits and detriments were explained in S. L. Taylor et al., 2012 by how easy to speechread a word was.

In our study, participants were not specifically asked about audiovisual asynchronies in the audiovisual material during the evaluation, and none reported any temporal artifacts.

2.4.2. Advantages of Optimized and Validated Audio Material

As mentioned in the introduction, one of the advantages of using existing audio material is that it maintains the validity of acoustic speech. For example, van de Rijt et al., 2019 reported a large variability in intelligibility across words, which probably arose because the word acoustic levels were not balanced and optimized, as is usually done in MSTs. Nevertheless, this does not mean that a non-optimized MST is not usable: MSTs without level adjustments (Nuesse et al., 2019) are used in research and can be used to evaluate speech recognition thresholds with almost the same precision.

Another advantage of using existing audio material is that it makes the recording procedure simpler. Limiting the final number of sentences (150) simplifies and speeds up the recording process. Jamaluddin, 2016 did not have a final selection of sentences, and so they created all 100,000 possible sentences by re-mixing 100 recorded sentences. During the recording session, they had to ensure that the speaker's head was in the same physical position so that the videos could be cut and blended without artifacts. For this purpose, they had to fabricate a head-resting apparatus to keep the head in the same position. The material required an additional evaluation step to validate the re-mixed recordings, resulting in 600 final sentences.

Another possible solution for creating visual speech, and one that offers more flexibility and control, is animated virtual characters. Ideally, the virtual character's lip-syncing should achieve the same intelligibility scores as the videos of real speakers. Some of the current virtual characters used in audiological research improve speech intelligibility. Schreitmüller et al., 2018

used the German MST with virtual characters: CI and NH participants achieved 37.7% and 12.4% average word scoring in the visual-only condition, respectively. These values are below the scores we found in this study (50%), but they cannot be compared directly because we only considered young normal-hearing listeners. Similarly, Grimm et al., 2019 used the German MST with virtual characters and compared it to the material presented here (AV-OLSAf), but no SRT improvements were found. Devesse et al., 2018 reported an SRT improvement of 1.5 to 2 dB SNR with virtual characters, while we found a 5 dB SNR improvement; the speech material in that study was different from ours and thus cannot be compared directly.

For each research application one has to find the best compromise when creating audiovisual MSTs. For some it might be enough to use synthetic speech and virtual characters with lip-sync, whereas others might need audiovisual synchronous recordings with balanced word acoustic levels. We found that dubbed videos were the most cost-effective solution for the research applications in our laboratory and that it might be a useful technique for others to measure gross audiovisual speech intelligibility.

On a side note, we would like to encourage audiovisual MSTs as a tool for evaluating the lip-syncing animations of virtual characters. Most current research in lip animation and visual speech does not consider human-computer communication and speech understanding in their evaluation procedures (Jamaludin et al., 2019; S. Taylor et al., 2017).

2.4.3. Speechreading

The ceiling effects found in the audiovisual MST resulted from the visual speech contribution. These ceiling effects change how the audiovisual MST can

be tested. Some participants achieved scores up to 84% just by speechreading. If the audiovisual MST is tested with an adaptive procedure targeting 50% SRT, there will be quite some participants that will be able to speechread half the material without using acoustic information.

Even at 80% SRT, we found few participants that could achieve SNRs were the sentences are not audible anymore. Excluding these data points would have been equivalent to removing the best audiovisual scores. But keeping them as they were would have led to unrealistic audiovisual SNR benefits (some participants reached scores below -60 dB SNR in audiovisual lists). We decided that limiting these values to the level were acoustic information disappears was the best trade-off. Another sensible approach would be to use the median SNR instead of the mean.

These effects could be because the limited set of words in the MST is easy to learn, to differentiate visually, and to speechread. Additionally, because there are only 150 possible sentences, some participants might memorize some of them after several repetitions. However it is rather difficult to memorize the sentences because of their syntactical structure with low context (Bronkhorst et al., 2002). In sentences for which participants have no previous knowledge of content, one would expect lower speechreading scores, of around 30% (Duchnowski et al., 2000; Fernandez-Lopez & Sukno, 2017). Nevertheless, it can be argued that having some expectations about sentence content is probably closer to a real-life conversation.

Another possible factor is that the female speaker was easy to speechread. Additionally, Bench et al., 1995 reported that young female speakers were judged to be easier to speechread than males and older females. We did not

make a selection of speakers, as we wanted to have the same person that recorded the audio-only MST. Furthermore, female speakers have been recommended as a compromise between the voice of an adult male and a child (Akeroyd et al., 2015), so this was a reasonable starting point. Selecting speakers that are more difficult to speechread would probably reduce the ceiling effects.

An interesting alternative to audiovisual MSTs would be to develop a viseme-balanced MST. The audio-only MST is designed to be phonetically balanced, but this does not mean that the visual speech is balanced, as each phoneme does not necessarily correspond to a viseme (S. L. Taylor et al., 2012). Visual cues were previously reported to affect word intelligibility and word error for the AV-OLSAf (Llorach & Hohmann, 2019), demonstrating that acoustic speech and visual speech provide different information. Therefore it is possible that the visual speech found in the current MST sentences is not representative of the language tested. Language-specific viseme vocabularies (Fernandez-Lopez & Sukno, 2017) should be developed for this purpose.

That the audiovisual lists were correlated with the speechreading scores was expected (Macleod & Summerfield, 1987; Summerfield, 1992; van de Rijt et al., 2019). The better a participant was at speechreading, the less acoustic information he or she needed to understand speech. This correlation was present in noise and in quiet conditions; the audiovisual benefit was therefore resilient to the acoustic condition.

2.4.4. Training Effects

An improvement of 2.2 dB SNR between the 1st and the 8th list at 50% speech reception is expected in audio-only MSTs (Ahrlich, 2013). We found a

~3 dB SNR improvement at 80% speech reception between the first training list and the test list; this additional dB probably arose from the participants learning to speechread the material and becoming familiarized with the speaker (Lander & Davies, 2008). According to the statistical report, the training effect disappeared after one training list. Nevertheless, an average constant improvement was observed. This training effect was not reported in the audiovisual Dutch MST (van de Rijt et al., 2019) after a familiarization phase with the complete set of words and a training list of 10 audiovisual sentences.

2.4.5. Within- and between- subject variability

In the audio-only speech in noise SRTs, we found little within- and between-subject variability, which was expected, as all participants were young and did not have any hearing disability (Souza et al., 2007). Both within- and between-subject variability increased in the audio-only speech in quiet lists, which is expected in quiet conditions (Smoorenburg, 1992). Hearing thresholds and noise-induced hearing loss are usually correlated with speech in quiet scores: the worse the hearing levels, the worse the speech intelligibility in quiet (Smoorenburg, 1992). Nevertheless, we did not find this correlation in our study, probably because the PTAs were all very similar and we did not include hearing-impaired participants.

The speechreading scores were highly individual and diverse in a homogeneous group of participants, which was expected from the literature (Jamaluddin, 2016; van de Rijt et al., 2019). The test-retest analysis showed that the visual-only lists could differentiate significantly between individuals, meaning that the visual-only MST can assess the speechreading ability of an individual.

The larger between-subject variability found in audiovisual lists can be explained by individual speechreading abilities. If a participant had a high speechreading score, it would be reflected in its audiovisual score. Nevertheless, when looking at the test-retest differences, the within-subject variability in the audiovisual scores did not permit to differentiate between participants significantly. Why could the audiovisual MST not differentiate between participants in the audiovisual modality, given that they all had the same hearing abilities but very different speechreading scores? One possible explanation for the within-subject variability in the audiovisual condition is that the asynchronies of the audiovisual material reduced the test-retest reliability. Another plausible explanation is that the integration between two types of modalities (acoustic and visual) led to a variance that could not be accounted for, assuming that audiovisual integration is an independent modality (Grant, 2002). Further research should look into the within-subject variability in audiovisual speech perception, as it cannot be derived from this study.

Audiovisual MSTs are particularly relevant for testing severe-to-profound hearing-impaired listeners in the clinic. These listeners cannot perform audio-only intelligibility tests and therefore the audiovisual MST would be useful for investigating whether hearing aid or cochlear implant provision improves their audiovisual speech comprehension. Additionally, the test provides information about the speechreading abilities of an individual. If the individual can speechread well, further recommendations could be provided to the patient for everyday-live situations, such as placing yourself in a position where you can see the mouth of the speakers.

We believe that our material can be used for clinical purposes, when taking into account aforementioned effects: in order to minimize ceiling effects, an 80% SRT is recommended; and at least one or two training list should be used to minimize training effects. Further research should evaluate the AV-OLSAf with hearing-impaired and elderly participants, as some effects are expected: hearing-impaired listeners tend to be better speechreaders (Auer & Bernstein, 2007), and the ability to speechread decreases with age (Tye-Murray et al., 2007). Furthermore, the influence of the type of noise could change in the audiovisual version and should be investigated (K. C. Wagener & Brand, 2005). Audiovisual integration needs to be further investigated with specific tests of audiovisual integration and different subject groups, as it has been suggested as an indicator of audiovisual speech intelligibility in noise, especially for those individuals with a hearing loss (Gieseler et al., 2020).

2.5. Conclusions

- The method presented here keeps the validity of the original audio material while introducing concordant visual speech. Dubbed video recordings gave similar benefit in terms of gross speech intelligibility measures as naturally synchronous audiovisual recordings, according to literature data, and thus are applicable for our purposes of assessing audiovisual speech intelligibility scores. Other fine-grain effects of audiovisual interaction may not be accessible through the dubbed recordings.
- The audiovisual MST suffers from ceiling effects, which are closely related to the speechreading abilities of the participant. These effects should be considered when designing experiments for audiovisual

perception. High target SRTs such as 80% SRT are recommended instead of 50% SRT in adaptive procedures.

- Audiovisual stimuli gave an SRT benefit of 5 dB SNR in test-specific noise and 7 dB SPL in quiet in comparison to audio-only stimuli for young, normal-hearing participants. Reference values for 80% SRT found in this study were -13.2 dB SNR for audiovisual speech in noise and 10.7 dB SPL for audiovisual speech in quiet.
- At least one training list should be completed in order to avoid statistically significant training effects. These effects may continue after a certain number of training lists. It is therefore recommended that two training lists are used to evaluate an audiovisual condition.
- Audiovisual SRTs correlated with speechreading abilities. The better participants could speechread, the more they benefited in the audiovisual conditions.
- The visual-only MST can be used to differentiate between the speechreading abilities of young normal-hearing individuals. Due to the variability in the audiovisual SRTs, we recommend including a visual-only condition when assessing audiovisual speech perception with the AV-OLSAf.

ACKNOWLEDGMENTS

This work received funding from the EU's H2020 research and innovation program under the MSCA GA 675324 (ENRICH), from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 352015383 (SFB 1330 B1 and C4) and from the European Regional Development Fund – Project "Innovation network for integrated, binaural hearing system technology (VIBHear)". We would like to thank the Media

Technology and Production of the CvO University of Oldenburg for helping out with the recordings. Special thanks to Anja Gieseler for giving feedback on the evaluation procedures and the manuscript and to Bernd T. Meyer for counseling on the video selection metric.

REFERENCES

Ahrlich, M. (2013). *Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit. Optimization and evaluation of the female OLSA and investigation of the speaker's effects on speech intelligibility* [Bachelor Thesis]. Carl von Ossietzky Universität Oldenburg.

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., Gagné, J. P., Lutman, M., Wouters, J., Wong, L., & Kollmeier, B. (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. *International Journal of Audiology*, *54*. https://doi.org/10.3109/14992027.2015.1030513

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research*, *50*(5). https://doi.org/10.1044/1092-4388(2007/080)

Başkent, D., & Bazo, D. (2011). Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and Hearing*, *32*(5). https://doi.org/10.1097/AUD.0b013e31820fca23

Bench, J., Daly, N., Doyle, J., & Lind, C. (1995). Choosing talkers for the BKB/A Speechreading Test: A procedure with observations on talker age and gender. *British Journal of Audiology*, *29*(3). https://doi.org/10.3109/03005369509086594

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1-4 SPEC. ISS.). https://doi.org/10.1016/j.specom.2004.10.011

Brand, T., Kissner, S., Jürgens, T., Berg, D., & Kollmeier, B. (2011). Adaptive Algorithmen zur Bestimmung der 80%-Sprachverständlichkeitsschwelle. Adaptive algorithms for determining the 80% speech intelligibility threshold. *14th Annual Meeting of the Deutsche Gesellschaft Für Audiologie (DGA)*.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, *111*(6). https://doi.org/10.1121/1.1479152

Bronkhorst, A. W., Brand, T., & Wagener, K. (2002). Evaluation of context effects in sentence recognition. *The Journal of the Acoustical Society of America*, *111*(6). https://doi.org/10.1121/1.1458025

Corthals, P., Vinck, B., de Vel, E., & van Cauwenberge, P. (1997). Audiovisual speech reception in noise and self-perceived hearing disability in sensorineural hearing loss. *International Journal of Audiology*, *36*(1). https://doi.org/10.3109/00206099709071960

Devesse, A., Dudek, A., van Wieringen, A., & Wouters, J. (2018). Speech intelligibility of virtual humans. *International Journal of Audiology*, *57*(12). https://doi.org/10.1080/14992027.2018.1511922

Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., & Braida, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE Transactions on Biomedical Engineering*, *47*(4). https://doi.org/10.1109/10.828148

Fernandez-Lopez, A., & Sukno, F. M. (2017). Automatic viseme vocabulary construction to enhance continuous lip-reading. *VISIGRAPP 2017 - Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, *5*. https://doi.org/10.5220/0006102100520063

Gieseler, A., Rosemann, S., Tahden, M., Wagener, K. C., Thiel, C., & Colonius, H. (2020). Linking audiovisual integration to audiovisual speech recognition in noise. *OSF Preprints*, *September*.

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America*, *112*(1). https://doi.org/10.1121/1.1482076

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2003). Discrimination of Auditory-Visual Synchrony. *AVSP 2003 - International Conference on Audio-Visual Speech Processing*.

Grimm, G., Llorach, G., Hendrikse, M. M. E., & Hohmann, V. (2019). Audio-visual stimuli for the evaluation of speech-enhancing algorithms.

Proceedings of the International Congress on Acoustics, *2019-September*. https://doi.org/10.18154/RWTH-CONV-238907

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., & Kollmeier, B. (2012). A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology*, *51*(7). https://doi.org/10.3109/14992027.2012.670731

Hochmuth, S., Jürgens, T., Brand, T., & Kollmeier, B. (2015). Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation? *International Journal of Audiology*, *54*. https://doi.org/10.3109/14992027.2015.1088174

Jamaluddin, S. A. (2016). Development and Evaluation of the Digit Triplet and Auditory-Visual Matrix Sentence Tests in Malay. In *University of Canterbury*.

Jamaludin, A., Chung, J. S., & Zisserman, A. (2019). You Said That?: Synthesising Talking Faces from Audio. *International Journal of Computer Vision*, *127*(11–12). https://doi.org/10.1007/s11263-019-01150-y

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. In *International Journal of Audiology* (Vol. 54). https://doi.org/10.3109/14992027.2015.1020971

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7). https://doi.org/10.1080/17470210801908476

Lidestam, B., Moradi, S., Pettersson, R., & Ricklefs, T. (2014). Audiovisual training is better than auditory-only training for auditory-only speech-in-noise identification. *The Journal of the Acoustical Society of America*, *136*(2). https://doi.org/10.1121/1.4890200

Llorach, G., Grimm, G., Hendrikse, M. M. E., & Hohmann, V. (2018). Towards realistic immersive audiovisual simulations for hearing research capture, virtual scenes and reproduction. *AVSU 2018 - Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Co-Located with MM 2018*. https://doi.org/10.1145/3264869.3264874

Llorach, G., & Hohmann, V. (2019). Word error and confusion patterns in an audiovisual German matrix sentence test (OLSA). *Proceedings of the International Congress on Acoustics*, *2019-September*. https://doi.org/10.18154/RWTH-CONV-239621

Llorach, G., Kirschner, F., Grimm, G., & Hohmann, V. (2020). *Video recordings for the female German Matrix Sentence Test (OLSA)*. Zenodo. https://zenodo.org/record/3673062

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2). https://doi.org/10.3109/03005368709077786

Nuesse, T., Wiercinski, B., Brand, T., & Holube, I. (2019). Measuring Speech Recognition With a Matrix Test Using Synthetic Speech. *Trends in Hearing*, *23*. https://doi.org/10.1177/2331216519862982

Puglisi, G. E., Astolfi, A., Prodi, N., Visentin, C., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2014). Construction and first evaluation of the Italian Matrix Sentence Test for the assessment of speech intelligibility in noise. *Proceedings of Forum Acusticum*, *2014-January*.

Sakoe, H., & Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1). https://doi.org/10.1109/TASSP.1978.1163055

Sanchez Lopez, R., Bianchi, F., Fereczkowski, M., Santurette, S., & Dau, T. (2018). Data-Driven Approach for Auditory Profiling and Characterization of Individual Hearing Loss. *Trends in Hearing*, *22*. https://doi.org/10.1177/2331216518807400

Schreitmüller, S., Frenken, M., Bentz, L., Ortmann, M., Walger, M., & Meister, H. (2018). Validating a Method to Assess Lipreading, Audiovisual Gain, and Integration during Speech Reception with Cochlear-Implanted and Normal-Hearing Subjects Using a Talking Head. *Ear and Hearing*, *39*(3). https://doi.org/10.1097/AUD.0000000000000502

Schubotz, W., Brand, T., Kollmeier, B., & Ewert, S. D. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *The Journal of the Acoustical Society of America*, *140*(1). https://doi.org/10.1121/1.4955079

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America*, *91*(1). https://doi.org/10.1121/1.402729

Souza, P. E., Boike, K. T., Witherell, K., & Tremblay, K. (2007). Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise. *Journal of the American Academy of Audiology*, *18*(1). https://doi.org/10.3766/jaaa.18.1.5

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*(2). https://doi.org/10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* (Vol. 335, Issue 1273). https://doi.org/10.1098/rstb.1992.0009

Talbott, R., & Larson, V. (1983). Research Needs in Speech Audiometry. *Seminars in Hearing*, *4*(03). https://doi.org/10.1055/s-0028-1091432

Taylor, S., Kim, T., Yue, Y., Krahe, M. M., Rodriguez, A. G., Hodgins, J., & Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics*, *36*(4). https://doi.org/10.1145/3072959.3073699

Taylor, S. L., Mahler, M., Theobald, B. J., & Matthews, I. (2012). Dynamic units of visual speech. *Computer Animation 2012 - ACM SIGGRAPH / Eurographics Symposium Proceedings, SCA 2012*.

Trounson, R. H. (2012). *Development of the UC Auditory-visual Matrix Sentence Test*. University of Canterbury.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, *18*(10). https://doi.org/10.3766/jaaa.18.10.7

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00335

Wagener, K., Brand, T., Kollmeier, B., & Kühnel, V. (1999). Entwicklung und Evaluation eines Satztests in deutscher Sprache. Teile I, II und III. *Zeitschrift Für Audiologie*, *38*.

Wagener, K. C., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, *44*(3). https://doi.org/10.1080/14992020500057517

Wagener, K. C., Hochmuth, S., Ahrlich, M., Zokoll, M. A., & Kollmeier, B. (2014). Der weibliche Oldenburger Satztest. *17. Jahrestagung Der Deutschen Gesellschaft Für Audiologie*.

Woodhouse, L., Hickson, L., & Dodd, B. (2009). Review of visual speech perception by hearing and hearingimpaired people clinical implications. In *International Journal of Language and Communication Disorders* (Vol. 44, Issue 3). https://doi.org/10.1080/13682820802090281

Chapter 3

Vehicle noise: comparison of loudness ratings in the field and the laboratory

Published in:

Llorach, G., Oetting, D., Vormann, M., Meis, M., & Hohmann, V. (2022). Vehicle noise: comparison of loudness ratings in the field and the laboratory. *International Journal of Audiology*, 1–10. https://doi.org/10.1080/14992027.2022.2147867

Abstract

Objective: Distorted loudness perception is one of the main complaints of hearing aid users. Measuring loudness perception in the clinic as experienced in everyday listening situations is important for loudness-based hearing aid fitting. Little research has been done comparing loudness perception in the field and in the laboratory.

Design: Participants rated the loudness in the field and in the laboratory of 36 driving actions. The field measurements were recorded with a 360° camera and a tetrahedral microphone. The recorded stimuli, which are openly accessible, were presented in three conditions in the laboratory: 360° video recordings with a head-mounted display, video recordings with a desktop monitor and audio-only.

Study samples: Thirteen normal-hearing participants and 18 hearing-impaired participants with hearing aids.

Results: The driving actions were rated as louder in the laboratory than in the field for the condition with a desktop monitor and for the audio-only condition. The less realistic a laboratory condition was, the more likely it was for a participant to rate a driving action as louder. The field–laboratory loudness differences were bigger for louder sounds.

Conclusion: The results of this experiment indicate the importance of increasing realism and immersion when measuring loudness in the clinic.

3.1. INTRODUCTION

One of the common complaints of hearing-impaired (HI) participants with hearing aids is about loudness: some sounds are too loud, and others are not heard (Anderson et al., 2018). When participants are provided with hearing aids, the hearing aids are fitted and adjusted in the clinic with controlled acoustic situations and audiometric tests, which are far from reflecting real-life scenarios. These disparities between the clinic and the field may lead to inaccurate estimates of loudness perception and, in consequence, to inappropriate settings in the hearing aids (Keidser et al., 2008).

To overcome these problems, loudness-related measurements in the laboratory should become more ecologically valid (Keidser et al., 2020) than established methods, i.e., they should better reflect real-life loudness perception. Loudness perception differences between the field and the laboratory have rarely been studied, as the complexity of a field situation is rather difficult to reproduce in the laboratory. Among the few existing studies, the experiment of Smeds et al. (2006) showed some interesting disparities between the field and the laboratory. Normal-hearing (NH) participants and participants with hearing loss were instructed to use research hearing aids in the field for a week. They could adjust the loudness through the volume control, and, when they did, the research hearing aid recorded the gain of the device and the sound pressure level of the field situation. Then, the participants were invited to the laboratory, where they had to adjust the volume of their research hearing aids, this time in a controlled audiovisual laboratory experiment. The stimuli in the laboratory, which consisted of recordings of a bushwalk, an office situation, a small gathering, a motorway and sawing wood with a power tool, were presented through a television screen and two frontal loudspeakers. The NH

participants chose lower gains in the laboratory than in the field, whereas the participants with hearing loss did the opposite: they chose higher gains in the laboratory than in the field. Several explanations were given in the article, such as the difficulty of imagining being in a particular situation in the laboratory, the possibility of the participants with hearing loss using lower gains in the field because of undesired soft background noises and the possibility of the NH participants using higher gains in the field to compensate for the reduced frequency range of the hearing aids.

A key factor when measuring loudness perception in the laboratory is visual information: visual cues have been found to influence loudness perception. When sounds were presented together with congruent visual cues, they were usually perceived as less loud (Fastl, 2004). In further experiments, the differences between immersive audiovisual simulations (i.e., a car simulator and videos via a head-mounted display) and audio-only reproduction were investigated. The loudness judgments, which were measured with a free-modulus magnitude estimation task, were decreased by about 15% in the immersive audiovisual simulations, in some individual cases by more than 50%. In free-modulus magnitude estimation, the participant is asked to assign a numerical value to the first stimulus. The following stimuli are rated consecutively relative to that number, e.g., if the first stimulus had a rating of 10 and the next one a rating of 5, that means a reduction of 50% for the second stimulus. These findings were confirmed in similar experiments, reviewed in Fastl & Florentine (2011).

The aim of our work was to compare loudness perception for field and different laboratory setups and to further explore the factors influencing loudness perception in laboratory experiments. We measured loudness

perception in the field and in the laboratory with the same participants. NH and HI participants were included, as the study of Smeds et al. (2006) showed differences between these groups. We recorded the stimuli in the field and replicated them in the laboratory with different setups. The laboratory setups ranged from immersive experiences (head-mounted display and stereo audio) to more simple clinical setups (only audio with a single frontal loudspeaker), as we wanted to know which requirements a clinical setup should have to measure loudness perception as in the field.

The methods and results of the field experiment can be found in Llorach et al. (2019) for the NH participants and in Oetting et al. (2020) for the participants with hearing loss. Our work provides an addition to the findings of Smeds et al. (2006), where a direct comparison between the stimuli in the laboratory and the field could not be done, due to the uncontrolled nature of the field situations and to the work of Patsouras (2003), where there were no field measurements to compare to the audiovisual simulations. To the best of our knowledge, this is the first work that compares field and laboratory loudness perception using the same kind of stimuli and the same participants. Implications for fitting procedures for the participants with hearing loss are discussed in Oetting et al. (2020).

3.2. Materials and Methods

Figure 3.1. Vehicles used in the experiment. From left to right: car (Opel Corsa 2016), motorbike (Suzuki VX 800 800 cc 1994), van (Ford Transit FT100 1999) and street sweeper (Kärcher MC 50). The figure is taken from Llorach et al. (2019)**.**

The participants were asked to rate the perceived loudness of different driving actions, using the response scale of the categorical loudness scaling (CLS) procedure (ISO 16832:2006, 2006) for loudness. The CLS uses an ordinal scale with name tags from "not heard" and "very soft", to "loud" and "extremely loud". The field experiment was conducted in a private street on a former military facility. The participants were distributed in four different sessions/dates. The listening positions were on a side of the street, and the participants rated the driving actions of four vehicles (see Figures 3.1 and 3.2). These driving actions were recorded with a 360° camera (Xiaomi Mi Sphere Camera, Xiaomi, Hong Kong), a tetrahedral microphone (Core Sound TetraMic, Core Sound, LLC, Teaneck, USA), and a sound level meter.

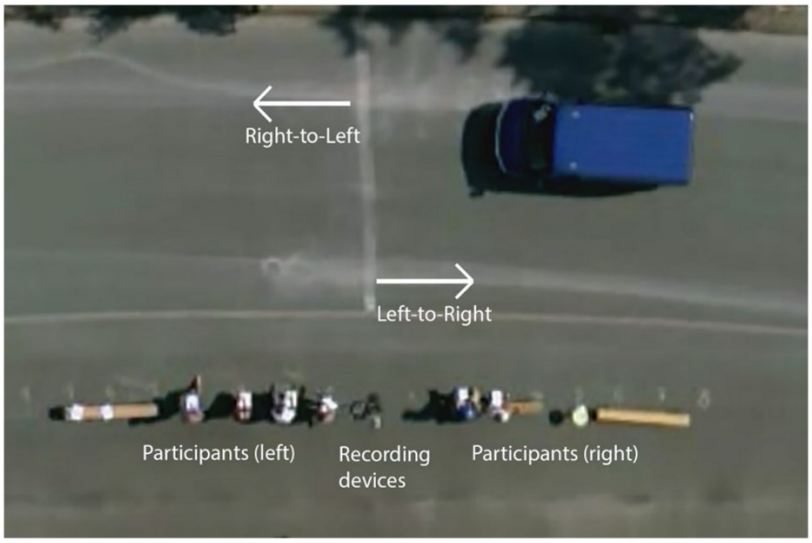


Figure 3.2. Setup of the field experiment. The figure is taken from Llorach et al. (2019).

In the laboratory experiments, the recorded driving actions were played back in three conditions: (1) 360° video playback with a head-mounted display (HMD) and stereo audio with loudspeakers at ±60° (360VID); (2) video playback with a computer monitor and stereo audio with loudspeakers at ±60° (2DVID) and (3) audio-only with a frontal loudspeaker (AO).

With such a design it is not possible to discern the effect of visual cues independently, as the audio setup was different in the audio-only condition. Rather than measuring the effect of visual cues, this experiment compares two audiovisual setups and a setup (AO) that represents the simplest clinical setup for loudness measurements. Because the audiovisual setups had the same audio setup, a comparison between the two visual displays (HMD and computer monitor) was possible.

3.2.1. Participants

Thirteen NH participants (six female and seven male) and 18 participants with hearing loss (11 female and seven male) participated in the field and in the laboratory experiments. The NH participants had a pure-tone average across the frequencies 500, 1000, 2000 and 4000 Hz between −2 and 13 dB HL. The age of the NH participants ranged from 27 to 72 years with an average of 53.5 years. The pure-tone average of the HI participants was between 34 and 52 dB HL with an average of 42.4 dB HL. The difference between the pure-tone average of the left and right ears was below 15 dB, so all participants had symmetric hearing loss. The age of the HI participants ranged from 69 to 80 years with an average of 74.9 years. Ten HI participants were experienced with hearing aids and eight were new users. Phonak Audéo B90-312 hearing aids were fitted with trueLOUDNESS (program 1) and with NAL-NL2 (program 2) (Oetting et al., 2018). The two fitting methods were used as part of the experiment described in Oetting et al. (2020). In this work, only the ratings with the trueLOUDNESS fitting were considered, which accounts for binaural loudness summation and aims at avoiding under- and over-amplification. In particular, to derive trueLOUDNESS gains, binaural broadband loudness summation was measured in each participant with hearing loss according to the procedure described in Oetting et al. (2016), which employs loudness scalings of narrowband noise signals and the IFnoise, a wideband signal with the long-term speech spectrum. The approach of Oetting et al. (2018) was then used to modify frequency-specific gains derived from narrowband loudness scaling by a binaural broadband gain correction taken from a 3D-gaintable (Oetting et al., 2018, Figure 4). The binaural broadband gain correction was fixed for an interaural level difference parameter of $\Delta L = 0$ and a bandwidth parameter of $B = 9.3$, which corresponds to the bandwidth

estimation for the speech shaped noise signal (IFnoise, (Holube, 2011)), that was used to measure the binaural broadband loudness summation.

The gains for program 1 were adjusted according to the trueLOUDNESS gain calculations for levels of 50, 65 and 80 dB SPL of the IFnoise signal. An acoustician manually adjusted the gains to match the target trueLOUDNESS functions and the gain functions of the hearing aid. Program 2 used the fitting method NAL-NL2 and its corresponding software to calculate the gains.

Individual ear moulds (cShells, when possible) or domes (open, closed or power dome according to the recommendations of the Phonak fitting software) were used for acoustic coupling. In the laboratory experiments, the trueLOUDNESS fitting with the same hearing aids and earmolds as in the field experiment was used. More details of the hearing aid fitting and a description of the HI participants can be found in Oetting et al. (2020). Ethical permission was granted by the ethics committee of the CvO Universität Oldenburg (Drs. 1r63/2016). The participants were recruited, contacted and reimbursed through Hörzentrum Oldenburg GmbH.

3.2.2. Stimuli

Four vehicles were used, which are shown in Figure 3.1: a white car (Opel Corsa 2016), a red motorbike (Suzuki VX 800 800 cc 1994), a dark blue van (Ford Transit FT100 1999) and a street sweeper (Kärcher MC 50). Loudness for the first three vehicles was rated in 10 conditions (five driving actions, once on each side of the street). These actions were “stand by with the engine on”, “stand by to drive forward”, “pass by at 30 km/h”, “pass by at 50 km/h” and “brake until stopping”. The vehicles drove towards the end of the street and turned back, once out of the sight of the participants, to do the next driving

action, this time on the other side of the street. For example, a vehicle would "stand by to drive forward" on the participant's street side, reach the end of the street, turn back and "pass by at 30 km/h" on the other side of the street. Loudness ratings for the street sweeper were assessed for six driving situations (three actions, once on each side of the street): "stand by with the engine on", "stand by with the brushes on" and "stand by to move and brush forward".

Each driving action was repeated eight times (four sessions, test and retest for the NH participants and program 1 and program 2 for the participants with hearing loss). The drivers aimed to repeat the driving actions identically. The sound level for each driving action had an average standard deviation (SD) of 1.7 dB and a reliability coefficient of 0.96 ($p < 0.001$). The sound pressure levels of the driving actions were measured with a sound level metre (Nor140, Norsonic Tippkemper GmbH, Oelde-Stromberg, Germany) and were calculated as the maximum level in dB SPL in windows of 125 ms. The average level for each driving action is shown in Table 3.1.

Table 3.1. Vehicle driving actions with average maximum level in dB SPL (125 ms windows). The actions are numbered with the order of presentation during the experiment. LR and RL stand for the direction of the driving: Left-to-Right (LR) and Right-to-Left (RL). The table is taken from Llorach et al. (2019).

Maximum level (dB SPL) of the driving actions in the field

	1A. Stand by (close)	2A. Accelerate LR (close)	3A. 30 km/h RL (far)	4A. 50 km/h LR (close)	5A. Break and stop RL (far)	6A. Stand by (far)	7A. Accelerate RL (far)	8A. 30 km/h LR (close)	9A. 50 km/h RL (far)	10A. Break and stop LR (close)
Car	71.2	84.3	73.3	81.5	75.2	67.9	80.1	75.2	76.9	77.1
Motor bike	83.5	91.5	82.5	89.7	81.1	78.4	86.6	89.0	88.1	84.0
Van	82.7	88.4	81.1	90.1	80.5	80.3	87.8	84.5	85.9	82.8

	1B. Stand by (close)	2B. Brushes on (close)	3B. Forward LR (close)	4B. Stand by (far)	5B. Brushes on (far)	6B. Forward RL (far)
Street sweeper	83.6	91.1	92.6	76.9	83.7	83.5

The recorded signals in the field were edited for the laboratory experiment. Out of the eight recordings for each driving action, the one that contained less noise and distractions (birds chirping, wind, coughing) was selected for each driving action, leading to 36 final recordings for the laboratory. Each driving action recording was edited and cut to last 12 seconds. The acoustic recordings of the Tetrahedral microphone were synthesised to a stereo format (XY microphone setup) using the VVMic software from VVAudio. The faces of the participants were blurred for anonymity in the video recordings of the 360° camera. The sound levels of the selected driving actions ranged from 67.8 to 94.6 dB SPL (maximum level in windows of 125 ms). The acoustic levels in the laboratory were adjusted using a sound level meter (Nor140, Norsonic Tippkemper GmbH, Oelde-Stromberg, Germany) to match the sound pressure levels recorded in the field. The sound level meter was placed at the approximate position of the participant's ears in the laboratory. A global gain was set for all driving actions to adjust the sound levels. Due to the room acoustics of the laboratory and the signal differences between driving actions, variability of ±2 dB between the levels of the field and the laboratory was present. This sound level variability was not controlled for each driving action, as it was similar to the variability of the repetition of the driving actions (SD of 1.7 dB SPL). The audiovisual recordings of the driving actions for the laboratory experiment can be found in Llorach et al. (2020).

3.2.3. Setup

In the field experiments, the participants sat on the side of the road where t he vehicles were driving (see Figure 3.2). The participants sat on benches and c hairs and they kept their sitting position for the whole experiment.

In the laboratory experiment, the participants sat on a chair in an acoustically treated room. They sat in the middle of a circle of 12 spectrally flat loudspeakers GENELEC 8030 BPM (Genelec Oy, Olvitie, Finland). The loudspeakers were at a distance of 1.2 m from the center, at a height of 1.2 m and were located every 30°. Only the loudspeakers placed at ± 60° (stereo) and the frontal direction (mono) were used. For the 360VID and 2DVID conditions, the stereo loudspeakers were used. The frontal loudspeaker was used for the AO condition. In the 2DVID condition, the participants had a computer monitor in front of them, where the videos were displayed. The computer monitor was at an approximate height of 70 cm and within arm's reach of the participant. This computer monitor was moved away from the participant in the other two conditions because they used the head-mounted display (HMD) for the 360VID condition and they did not have any visual stimuli in the AO condition. The head-mounted display was the HTC Vive (HTC Corporation, New Taipei City, Taiwan). The videos were reproduced with the "Media Player Classic - Home Cinema" software in condition 2DVID and with the "Steam 360 Video Player" in condition 360VID. The computer used Windows 10 with an NVIDIA Quadro M5000 graphics card. The participants had a button on their lap that would mute the playback, in case of emergency or extreme discomfort.

3.2.4. Procedure

Field experiment

The participants were distributed across four sessions, as there was a limited number of seats. In each session, all 36 driving actions were done, then there was a pause of 30 min, and the 36 driving actions were repeated. For the NH participants, this was a test and retest of the ratings. The participants with

hearing loss were tested for the first 36 driving actions using the trueLOUDNESS fitting, and after the pause, the NAL-NL2 fitting.

The participants were instructed to rate the loudness of the driving actions. A researcher indicated the number of the driving action to rate when the driving action was being executed (see video recordings in Llorach et al. (2020)). The indication was given to instruct the participants to rate the current action. This was especially important for the static driving actions, e.g., stand by, as they had to know that that was actually an action to be rated (the vehicles had to move to that position beforehand and that could be mistaken for a driving action). Once all participants had rated the current driving action, the next driving action was executed. The driving actions followed the order shown in Table 3.1 and each vehicle did all its driving actions consecutively. The car started first, followed by the motorbike, the van and the street sweeper.

Laboratory experiment

The laboratory experiments used the same participants. The field and laboratory experiments were separated by approximately eight months. For the participants with hearing loss, the same hearing aids with the trueLOUDNESS fitting were used. An audiologist measured the audiometric threshold to detect changes relative to their previous audiograms (none were found) and assisted with the hearing aids during the experiment.

The HMD was shown and given to the participants to familiarise them with the technology. The interpupillary distance of the participants was measured and the HMD was adjusted correspondingly. The straps of the HMD were adjusted to the head of the participants while the driving actions of the car

were shown through the device without sound. During this adaptation phase, the participants were asked to explore the 360° environment by head movements and to make themselves comfortable with the HMD. This phase lasted less than 2 min.

The order of the driving actions was the same as in the field experiment. The researcher who indicated the number of the driving action in its loudest instant in the field was visible in the videos. After each driving action, the video was paused until the participant indicated the perceived loudness. During this pause, the driving action number and the response scale were shown in the video, and no sounds were played back. In the 360VID condition, an additional letter was added for each loudness category in the questionnaire appearing in the video. In this way, the participants could answer verbally without taking off the head-mounted display. The order of the laboratory conditions was balanced (Latin square design): each condition was in first, second or third place the same number of times as the other conditions across participants.

Data processing

Not all participants experienced the same sound levels during the field experiment, as they were seated in different positions along the road (see Figure 3.2). The sound pressure levels that they experienced in the laboratory were different from the ones they were exposed to in the field for most driving actions, as the levels in the laboratory were not adjusted individually. We approximated the sound pressure level differences by assuming that the sound sources were omnidirectional and that there were no spectral differences. We used the following equation to compute the sound level differences:

$$dB_{diff} = sgn(d_2 - d_1) \cdot |20 \cdot \log\left(\frac{d_1}{d_2}\right)| \quad (3.1)$$

where dB_{diff} is the calculated sound level difference between the recording device and the participant, d_1 is the approximate distance between the position of the sound level meter and the position of the vehicle at its loudest instant of a driving action, d_2 is the approximate distance between the sitting position of the participant and the position of the vehicle in its loudest instant of a driving action, and sgn is the sign function, which determines if the dB difference is positive or negative. The driving actions that had equal levels for all participants (Table 3.1. 3A, 4A, 8A, 9A) had a 0 dB difference. The level differences between the laboratory and the field stimuli had an average value of 1.9 dB with a SD of 2.3 dB, with a range from -0.8 dB to 8.1 dB across all participants and driving actions.

We removed the ratings of the participants where the sound level difference was bigger than 1.5 dB. If a participant experienced a level difference above the set threshold according to our estimate, his/her loudness ratings of that driving action were removed for all conditions (field, 360VID, 2VID, AO). The value of the threshold was chosen to have a non-skewed distribution of level differences while preserving ratings for all participants and driving actions. Overall, 36% of the ratings were removed (19% NH, 17.0% HI), with a maximum of 61% for one participant. None of the 36 driving actions were completely removed. Figure 3.3 shows the distribution of the sound level differences.

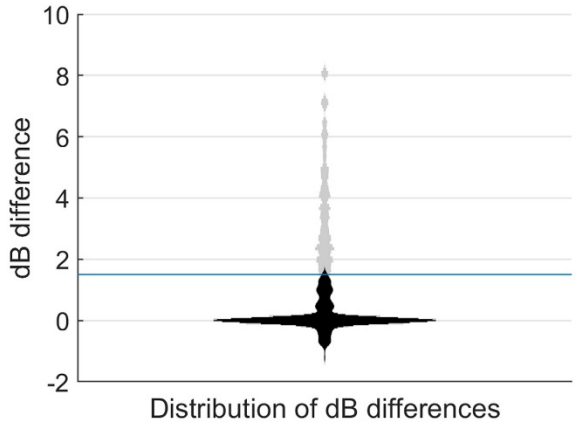


Figure 3.3. Distribution of the sound level differences between the field and the laboratory for all ratings due to differences in sitting position and the driving actions. Grey indicates the differences for which ratings were removed. Black shows the differences for the remaining ratings. The criterion for removing the ratings for a given driving action is marked with a blue horizontal line.

Statistical analysis

The differences in loudness perception ratings were analysed with two different approaches: metric-model analysis (repeated-measures ANOVA and Bonferroni-corrected pairwise comparisons) and ordinal analysis (non-overlap of all pairs (NAP) (Parker & Vannest, 2009) and group comparisons with Mann–Whitney U tests). The repeated-measure ANOVA analysed the effects on a group level and assumed that the rating data were metric, whereas the NAP measure analysed the effect size on an individual level and used the ordinal ratings. These two analyses were complementary: NAP scores only provided information about what happened with each participant, whereas the repeated-measures ANOVA analysed effects on a general level. The design of these two complementary approaches is described in this section.

The repeated-measures ANOVA indicated if the loudness ratings were affected by the condition, if there were differences between groups (NH vs HI), and if there were interactions between condition and group. For this analysis, each participant had four numerical values as the dependent variable (one for each condition, being the average of the loudness ratings for that condition) and a group factor (NH or HI). In other words, the within-subject factor was condition (Field, 360VID, 2DVID, AO) and the between-subject factor was hearing type (normal hearing or hearing impaired). To obtain a numerical value for each condition as the dependent variable, the loudness ratings of a given condition were averaged, of which there were 36 in the best case and 14 in the worst case due to data removal (see Figure 3.3). To average them, the loudness categories were transformed to a monotonically increasing numerical scale between 0 and 50 in steps of 5 for each loudness category/response alternative, as recommended by the ISO 16832:2006 standard. We assumed that the loudness categories were equidistant (see Discussion section). Because the NH participants gave ratings for two field measurements (test and retest), the mean of the test and retest rating was used to calculate the average rating for the field condition. For the participants with hearing loss, we used the field ratings that were done with the trueLOUDNESS fitting for the averaging, as the same fitting was used in the laboratory conditions. Bonferroni-corrected pairwise comparisons, if a main effect was found, indicated which conditions/groups were different from each other and the direction of the effect.

Metric-model analyses, such as ANOVA, are often used for analysis of behavioural ordinal data. Nevertheless, Liddell & Kruschke (2018) showed that this can lead to errors. Therefore, we included NAP to measure the nonparametric effect size and to complement the metric-model analysis. NAP

provided a score for a comparison between conditions for each participant, i.e., it compared the ratings in condition A to the ratings in condition B of a participant. Using confidence intervals, the NAP scores indicated how many participants rated a condition significantly louder, quieter, or similarly loud than another condition. The NAP scores of the NH and HI participants were compared with Mann–Whitney U tests to check if there were differences between groups.

The result of NAP is an intuitive number from 0 to 1: if all ratings in condition A are bigger than in condition B, the NAP score is 1; if all ratings are equal in the two conditions, the score is 0.5; if all the ratings for A are below the ratings for B, the score is 0. Six comparisons were done in the analysis: Field-360VID, Field-2DVID, Field-AO, 360VID-2DVID, 360VID-AO, 2DVID-AO. We modified the NAP formula to compare a driving action rating of condition A to its corresponding one in condition B, instead of comparing a rating of condition A to all the ratings in condition B. This was done because the loudness ratings of our experiment were paired: a rating in condition A had its equivalent in condition B. As a result, ratings of unrelated driving actions were not compared to each other. The modified formula is the following:

$$NAP_k = \frac{1}{n*TRT}\sum_{i=1}^{n}\sum_{j=1}^{TRT}\left[I\left(r_{i,j}^{B} > r_i^{A}\right) + 0.5\, I\left(r_{i,j}^{B} = r_i^{A}\right)\right]; \qquad (3.2)$$

where k is the participant number, n is the number of driving action ratings (between 14 to 36 for each participant), TRT is the test/retest rating (2 for NH participants and 1 for HI participants for the field condition, 1 for all other conditions), r is the rating of the driving action i, and A and B are the conditions being compared. The ratings were a numerical scale between 0 and 50 (the loudness categories were transformed to a numerical scale as

recommended by ISO 16832:2006). For the HI participants, we selected the ratings of the field condition when the trueLOUDNESS fitting was used.

To determine if a NAP value was significantly above or below the chance level (0.5), the confidence intervals were computed. If the confidence intervals contained the 0.5 value, the two conditions being compared were not different from each other for that participant. Otherwise, the conditions compared were significantly different for that participant. The confidence intervals were computed using the standard error formulas proposed in Newcombe (2006) in Method 6, namely:

$$SE_k = \sqrt{\frac{n-1}{n^2} NAP_k(1 - NAP_k)\left(\frac{1}{n-1} + \frac{1-NAP_k}{2-NAP_k} + \frac{NAP_k}{1+NAP_k}\right)}; \qquad (3.3)$$

where k is the participant number, and n is the number of driving action ratings for each participant. The confidence intervals were computed as $NAP_k \pm z \cdot SE_k$; where z was defined as 1.645 for a confidence interval of 90%.

Group differences were analyzed with Mann-Whitney U test using the NAP scores. Six Mann-Whitney U tests were done, one for each comparison of conditions. The test indicated if the differences of loudness perception between conditions were different for the NH and HI groups.

To understand the NAP results and the variability of the ratings better, an additional comparison between the field test and field retest of the NH ratings was added to the analysis.

3.3. Results

The results of the repeated-measures ANOVA are described as the following: Levene's test showed that the variances for the dependent variable were equal.

Mauchly's test indicated that the assumption of sphericity was violated, $\chi 2(5) = 14.504$, $p = 0.013$, and therefore, the Greenhouse-Geisser correction was used. There was no interaction between condition and hearing type, $F(2.214, 6.586) = 0.280$, $p = 0.781$. Hearing type did not have a significant effect on the mean loudness ratings, $F(1, 1.491) = 0.022$, $p = 0.882$. The mean loudness rating differed significantly between conditions, $F(2.243, 131.618) = 5.591$, $p = 0.004$. Pairwise comparisons with Bonferroni correction showed that the mean loudness ratings for the field condition were significantly different from those for the AO condition ($p = 0.018$), but not from those for the 2DVID condition ($p = 0.060$) or the 360VID condition ($p = 1.0$). The laboratory conditions did not differ significantly from one another, according to pairwise comparisons. Figure 3.4 shows the distributions of the mean loudness ratings for the four conditions. Overall, the loudness ratings were slightly higher in the laboratory than in the field. The two laboratory conditions that were not significantly different from the field were the 360VID and the 2DVID, which included visual cues and stereo audio. The 2DVID condition, which was less immersive than the 360VID, was borderline non-significant ($p = 0.06$).

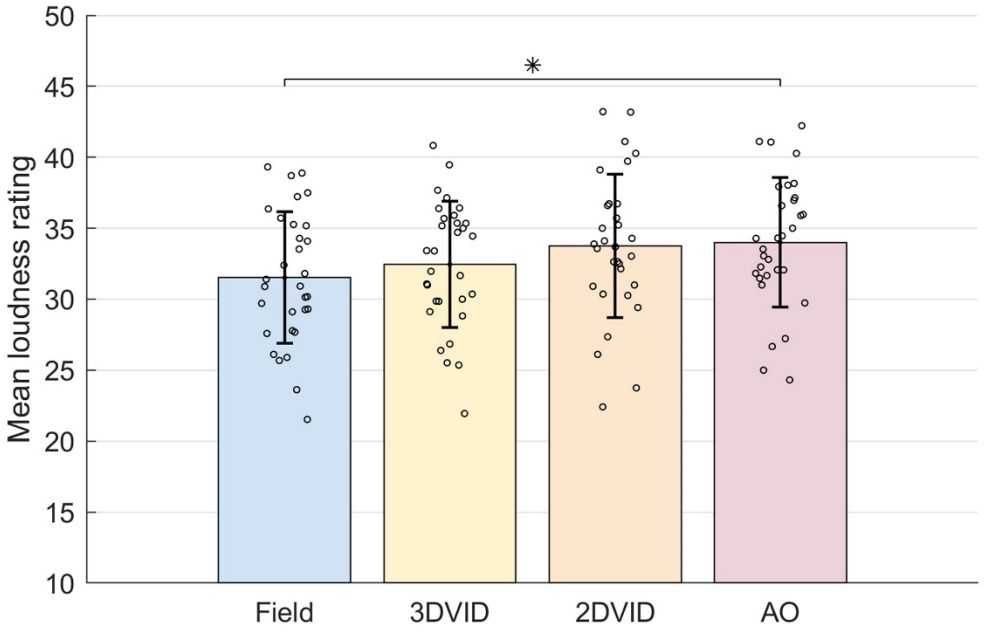


Figure 3.4. Distribution of the mean loudness ratings. Each bar has 31 black dots on top, one for each participant. Each dot is the mean of the loudness ratings of that participant for that condition (average of 14–36 ratings for each participant). Each bar represents the mean for each condition: Field, 360VID, 2DVID and AO. The vertical line in the middle of each bar indicates the standard deviation of the distribution. The one significant difference is indicated with an asterisk ($p < 0.05$).

For the ordinal analysis, NAP scores were computed for each individual and pairwise comparison (31×6). Additionally, NAP scores were computed for the test-retest field ratings of the 13 NH participants. To summarise each comparison, we report the number of participants that rated loudness significantly higher in condition A, the number of participants that did not rate loudness significantly different, and the number of participants that rated loudness significantly higher in condition B. Confidence intervals described in the previous section were used to determine if there was a significant difference. Table 3.2 summarises the scores.

According to NAP scores, loudness perception in the 360VID condition and the Field condition did not follow a specific tendency: 12 participants rated loudness higher in the 360VID condition and 10 participants rated loudness higher in the Field condition, as shown in Table 3.2. The differences between the field and the other two laboratory conditions indicated that loudness was usually rated higher in those laboratory conditions: 18 participants rated loudness higher in the 2D condition and 16 did in the AO condition, whereas only 5 participants rated loudness higher in the Field condition in comparison to the 2D and AO conditions (see Table 3.2). The results of the repeated-measures ANOVA were somewhat in concordance with the NAP scores: the AO-Field difference was significant, while the 2D-Field difference was borderline significant ($p = 0.06$).

Table 3.2. Number of participants with a certain loudness perception difference or similarity between conditions. The number of participants is determined by the NAP scores and their confidence intervals.

Laboratory vs Field	**Num. of participants**	**Laboratory conditions**	**Num. of participants**
360VID vs Field	360VID > Field: 12 360VID = Field: 9 Field > 360VID: 10	2DVID vs 360VID	2DVID > 360VID: 12 2DVID = 360VID: 14 360VID > 2DVID: 5
2DVID vs Field	2DVID > Field: 18 2DVID = Field: 8 Field > 2DVID: 5	AO vs 360VID	AO > 360VID: 14 AO = 360VID: 12 360VID > AO: 5
AO vs Field	AO > Field: 16 AO = Field: 10 Field > AO: 5	AO vs 2DVID	AO > 2DVID: 11 AO = 2DVID: 14 2DVID > AO: 6
Field Retest vs Field Test (NH participants only)		**Num. of participants**	
Test vs Retest Field		Retest > Test: 2 Retest = Test: 7 Test > Retest: 4	

If a laboratory condition was less realistic, loudness was usually rated higher in that condition. Twelve participants rated loudness higher in the 2D condition than in the 360VID condition, and five participants did the opposite. Similarly, 14 participants rated loudness higher in the AO condition than in the 360VID condition where 5 participants did the opposite. The difference between the 2D and AO conditions was less pronounced but in the same direction: eleven participants rated loudness higher in the AO condition than in the 2D condition, whereas six participants did the opposite. The pairwise comparisons of the repeated-measures ANOVA did not show significant differences between laboratory conditions.

Loudness in the laboratory conditions was similar for more participants than in the field versus laboratory comparisons (see Table 3.2). In the comparisons between laboratory conditions, the number of participants with similar loudness ratings ranged between 12 (39%) and 14 (45%), whereas in the comparisons between laboratory conditions and the field, the number of participants ranged from 8 (26%) to 10 (32%) participants. When looking at the test-retest comparison of the NH participants, the relative number of participants with similar ratings was higher (7 out of 13 – 54%).

Six Mann–Whitney U tests (one for each comparison) were conducted using the NAP scores and the hearing type to determine if there were differences between groups. None of the tests showed significant differences. The U values ranged between 92 and 107, and the p values were between 0.326 and 0.704. The variances (Levene's test) and normality (Shapiro–Wilk test) of the NAP scores were equal between groups.

To assess whether these differences differed for loud and soft noises, we computed the correlation between the sound pressure level of the driving actions and the laboratory-field loudness rating differences. The loudness rating differences were computed between the field and the laboratory conditions for each driving action and participant. For each driving action we computed the average laboratory-field difference across participants, resulting into 36 data points. Figure 3.5 shows the loudness laboratory-field difference for each driving action. Each circle represents the difference for a driving action. The Spearman correlation coefficient between the 360VID-Field loudness rating differences and the sound pressure levels was 0.05 ($p = 0.79$), the 2DVID-Field loudness rating differences and the sound pressure levels was 0.43 ($p < 0.01$) and between the AO-Field loudness rating differences and the sound pressure levels was 0.36 ($p = 0.03$). If there were differences between the laboratory and the field ratings, these were higher when the sounds had a higher level. This correlation was only significant for the 2DVID-Field and the AO-Field differences. The loudness ratings of this experiment can be found in Llorach et al. (2022).

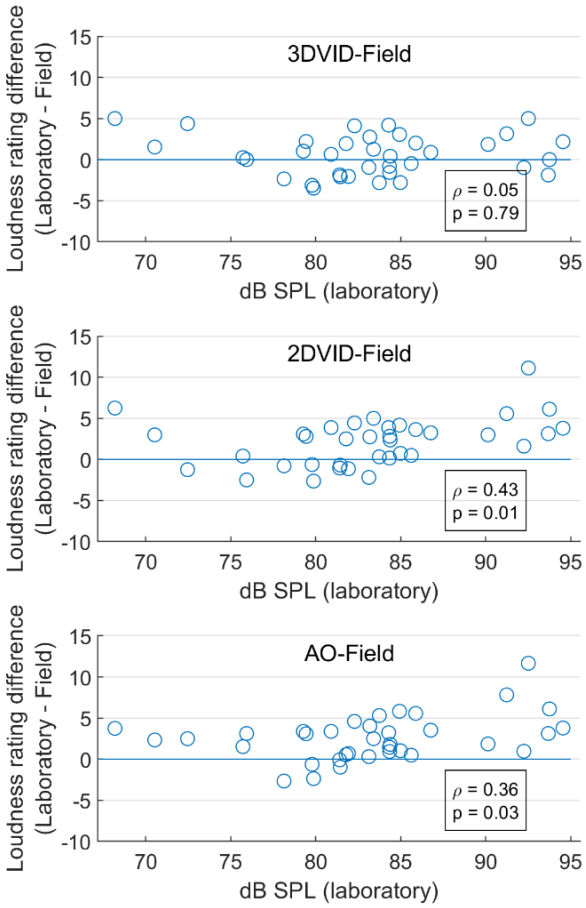


Figure 3.5. Relationship between the sound levels and the laboratory–field differences in loudness ratings. Each circle (36 for each panel) represents the mean loudness difference for a driving action. The average is done between participants: 31 ratings or less due to data removal. The relationship with each laboratory condition is represented in a different panel: 360VID (top), 2DVID (centre) and AO (bottom). The Spearman correlation coefficient (ρ) and its p value are shown on the bottom-right of each panel. If the driving action circles are above zero, these driving actions were rated louder in the laboratory.

3.4. Discussion

The vehicle driving actions were perceived as louder in the laboratory than in the field for the 2DVID condition (computer monitor and stereo loudspeakers), and for the AO condition (no visuals and a single loudspeaker): the repeated-measures ANOVA showed a significant difference between the AO condition and the field, and the NAP scores showed a higher percentage of participants rating the loudness in the 2DVID and the AO conditions higher than in the field.

When using immersive visual cues and stereo audio, loudness perception was similar in the field and in the laboratory: the 360VID condition showed similar loudness ratings to the field condition on average (no significant difference found in the metric-model analysis) and the NAP scores for the Field-360VID comparison were balanced (similar number of participants who rated one or the other condition as higher). The 360VID condition (HMD with 360° videos and stereo audio) was realistic enough to elicit the same loudness perception as in the field.

The results suggest that as the realism of the laboratory increased, the loudness ratings were lower and resembled more the ones from the field: the NAP scores for the comparisons between laboratory conditions showed that the least realistic condition had always a higher percentage of participants with higher loudness ratings (see Table 3.2). Therefore, immersive and realistic simulations should be considered for clinical evaluations of loudness perception that target ecological validity.

When comparing the 360VID and the 2DVID conditions, only the visual cues changed (from a head-mounted display to a computer monitor). Using

immersive visual cues instead of a computer monitor made participants rate loudness lower according to the NAP scores, in line with the literature (Fastl & Florentine, 2011). Regarding the loudness differences between the AO condition and the two other laboratory conditions (360VID and 2DVID), which factor (visual cues or stereo audio) had more influence could not be determined: the 2DVID and 360VID conditions had visual cues and stereo audio and the AO condition used mono audio and no visual cues.

The loudness perception differences between the field and the laboratory became more apparent for higher sound levels in the AO and the 2DVID conditions, meaning that the field–laboratory differences might be more apparent when using intense stimuli and undetectable for low-level sounds. Clinical evaluations should pay special attention to these differences, as intense sounds are the ones that usually cause loudness discomfort.

Although the field–laboratory differences were small on average in terms of categorical units, these differences should be considered in the methods for measuring loudness perception and in hearing aid fitting procedures. According to Heeren et al. (2013), the functions relating CUs and levels in dB SPL can have slopes of more than 0.1 CU per dB SPL. Although the field–laboratory rating differences found here were below one CU, these could be equivalent to 10 dB SPL in some situations. Gain adjustments in the hearing aid of that magnitude could influence listening comfort with hearing aids. As stated by van Beurden et al. (2018): "[...] there is need to adjust fitting rules for bilaterally fitted hearing aids to take the large individual differences in loudness summation into account". Therefore, research institutes and clinical facilities should be aware that increasing the ecological validity of their methodologies

may provide a better assessment of real-life hearing experiences and consequently better hearing aid fitting.

In the following paragraphs the limitations and challenges of comparing field and laboratory loudness perception are described. These should be considered when interpreting the results of this experiment.

3.4.1. Limitations

Making an exact replica of a field situation in the laboratory is very challenging, if not impossible (Keidser et al., 2020), and requires expensive equipment and expertise (Llorach et al., 2018). In this experiment, we tried to reproduce the field stimuli in the laboratory as accurately as possible using a setup that could be used in other labs or clinics. This means that marked differences between the laboratory and field setups were present and could have influenced the results.

The participants sat in different positions in the field experiment. They did not see and hear the same stimuli as the recording devices. By being in a different sitting position, the sound pressure levels, and the spectral shape of the driving actions changed. We tried to minimise this factor in the experimental design by doing the measurements in four sessions, in order to have fewer participants for each session and to have them sitting closer to the middle position and the recording devices. Nevertheless, we still had to remove about one third of the collected loudness ratings.

The driving actions were repeated eight times in the field and only one of those repetitions was used in the laboratory. Therefore, most participants did not experience the driving actions in the same way, as they were only present

for two of those eight repetitions in the field. Nevertheless, the repetition of the driving actions was quite accurate in terms of sound pressure levels (Pearson correlation coefficient = 0.96, $p < 0.001$) (Llorach et al., 2019) and the test-retest reliability of the ratings of the NH participants was high (Spearman correlation coefficient = 0.85, $p < 0.001$) (Llorach et al., 2019). Therefore, the effect on the ratings may be minimal.

The driving repetitions with less background noise and distractions were selected for the laboratory stimuli and for the open data publication (Llorach et al., 2020). This selection was done to create stimuli that can be used in future experiments where the main content is the driving actions. Nevertheless, this curation of the material could have added a bias to the differences between the laboratory and the field, as the laboratory stimuli were the ones with less noise. Not enough data were collected to find out if a bias existed. But as mentioned before, the test-retest reliability of the ratings of the NH participants was high enough to consider that this bias, if present, was minimal.

The acoustic experience in the laboratory was not the same as in the field. In the laboratory, the sound came from one or two visible loudspeakers, and although the room was acoustically treated, it was not fully anechoic. Acoustic reflections, room modes and distance to the loudspeakers (Mershon et al., 1981) could have affected the loudness ratings and added variability to the field-equivalent sound pressure levels. We wanted the design of our laboratory experiment to be closer to a clinical test than an exact reconstruction of the field experiment. Therefore, we did not provide any acoustic context in the laboratory: in the field experiment, the participants heard the vehicles when they were getting ready for each driving action and there was background noise

between driving actions. They could expect a certain loudness, which did not happen in the laboratory.

The field and laboratory experiments were separated by eight months due to technical preparations and time availability of the researchers. Separating two phases of this kind of experiment for such a long extent of time is not recommended. Hearing abilities may worsen, and participants may become unavailable for the second session after such a long time.

These differences and limitations between the laboratory and field experiment could explain the variabilities of the NAP scores in the ordinal analysis. There were no comparisons between conditions where all participants had the same tendency, i.e., all participants rated one or the other condition higher. The test-retest field comparison of the NH participants showed that 54% (7 out of 13) of the participants had similar ratings, as indicated by the NAP scores. The 360VID-Field comparison, where loudness perception was not significantly different, had only nine participants (29%) with similar ratings. It would be expected that the percentage of participants with similar ratings increases when loudness perception is similar. Nevertheless, the 360VID-Field comparison had a small percentage of participants with similar ratings.

The variability in the NAP scores can be explained by the differences and limitations between the field and the laboratory, but individual differences in loudness perception are a factor to consider. Previous literature has shown that there are individual differences in loudness perception within a homogeneous group. In fact, the trueLOUDNESS fitting is based on such individual differences: Oetting et al. (2018) and found large individual differences in binaural loudness summation, a measure that is usually not considered when

fitting hearing aids. Unfortunately, individual binaural loudness summation was not recorded for all participants and were not considered in this experiment. We considered hearing type, as Smeds et al. (2006) found differences between NH and HI participants when measuring field–laboratory gain preferences. We did not find differences in loudness ratings between hearing groups, even though we had a bigger sample size. The repeated-measures ANOVA did not show a significant difference between groups nor interactions, and the Mann–Whitney U tests on the NAP scores of the condition comparisons did not show significant differences between groups. The general tendency in our experiment was that the loudness ratings were higher in the 2DVID and AO laboratory conditions than in the field for both groups. Smeds et al. (2006) found a similar effect for the NH participants in a condition comparable to the 2DVID condition, i.e., NH participants chose lower hearing aid gains in the laboratory. In our study, the HI participants rated the stimuli as louder in the 2DVID conditions than in the field in opposition to what was found by Smeds et al. (2006): HI participants chose higher gains in the laboratory than in the field. In Smeds et al. (2006) participants were asked to set the preferred loudness, whereas in our study we asked them to rate perceived loudness. These two measures are different (preference vs perception) and could explain the differences found between the studies, e.g., NH and HI could have the same loudness perception in the laboratory, but the HI impaired chose to set the gains higher in Smeds et al. (2006).

3.4.2. Categorical loudness scaling

In our experiment we did not follow some of the standard procedures of categorical loudness scaling described by ISO 16832:2006. For example, the whole audible range should be presented (from not heard to too loud) and each

signal should be presented at five sound levels. In our experiment, the lowest sound level was well above the hearing level (>65 dB SPL) and each driving action was presented at the same level for each laboratory condition. These limitations should be taken into consideration when comparing the CU ratings to other studies using the same rating scale.

The standard procedure calculates the average of the sound levels that belong to a loudness category. In our case, we calculated the average of the loudness categories for a condition once these were transformed to a numerical scale, to be able to compare between conditions in the metric-model analysis (repeated-measures ANOVA). We assumed that the categorical units have a linear relationship with dB SPL and the loudness categories are equidistant, as suggested by ISO 16832:2006. The loudness function, ie the relationship between loudness categories and sound pressure levels, of narrowband noise signals has been fitted in previous work using two straight lines (Brand & Hohmann, 2002). For binaural broadband noise signals, the loudness function tends to be a single straight line (Oetting et al., 2016). Therefore, the linear relationship between loudness categories and sound pressure levels can be justified.

3.4.3. Future work

Future work should test laboratory audiovisual conditions with participants who were not in the field, as the participants experienced the same actions in the field and in the laboratory. The hypothesis is that the rating differences between the audio-only and the audiovisual conditions will become significant and bigger, as in previous work (Fastl, 2004). Another possible experiment would be to let the participants adjust the volume/gain of the stimuli, as in Smeds et al. (2006). The hypothesis is that the chosen levels would be lower for

the audio-only condition than for the audiovisual conditions and to the levels recorded in the field.

A further improvement to the study design would be to add other urban vehicles, such as electrical scooters. Such quieter vehicles would give references for the quieter categories of the loudness scale and thus increase its validity.

The 360VID condition of this experiment was the most realistic and the one that achieved similar loudness perception as in the field. Future research should test immersive audio reproduction techniques, e.g., Ambisonics or Vector Base Amplitude Panning, together with immersive visual cues. The hypothesis of such research would be that increasing the realism of an immersive simulation does not affect loudness perception, once the simulation is realistic enough to elicit ecologically valid loudness perception. Defining the "realistic enough" simulation could provide insightful indicators for clinical setups targeting ecological validity.

ACKNOWLEDGEMENTS

This work received funding from the EU's H2020 research and innovation program under the MSCA GA 675324 (ENRICH), from the Deutsche Forschungsgemeinschaft (DFG, Cluster of Excellence EXC 1077/1 "Hearing4all", and SFB1330 Projects B1 and C4). Thanks to Julia Schütze, Anja Kreuteler for helping conduct the experiments, Petra von Gablenz for helping with the ordinal analysis, and Melanie Krüger for contacting the participants. Special thanks to the personnel of the old military facility.

REFERENCES

Anderson, S., Gordon-Salant, S., & Dubno, J. R. (2018). Hearing and Aging Effects on Speech Understanding: Challenges and Solutions. *Acoustics Today*, *14*(4). https://doi.org/10.1121/at.2018.14.4.12

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, *112*(4). https://doi.org/10.1121/1.1502902

Fastl, H. (2004). Audio-visual interactions in loudness evaluation. *ICA, The 18th International Congress on Acoustics*.

Fastl, H., & Florentine, M. (2011). *Loudness in Daily Environments*. https://doi.org/10.1007/978-1-4419-6712-1_8

Heeren, W., Hohmann, V., Appell, J. E., & Verhey, J. L. (2013). Relation between loudness in categorical units and loudness in phons and sones. *The Journal of the Acoustical Society of America*, *133*(4). https://doi.org/10.1121/1.4795217

Holube, I. (2011). Speech intelligibility in fluctuating maskers. *Proceedings of the International Symposium on Auditory and Audiological Research, Vol. 3*, 57–64.

ISO 16832:2006. (2006). *Acoustics—Loudness scaling by means of categories*. International Organization for Standardization. https://doi.org/https://www.iso.org/standard/32442.html

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner,

T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It. *Ear and Hearing*, *41*. https://doi.org/10.1097/AUD.0000000000000944

Keidser, G., O'Brien, A., Carter, L., McLelland, M., & Yeend, I. (2008). Variation in preferred gain with experience for hearing-aid users. *International Journal of Audiology*, *47*(10). https://doi.org/10.1080/14992020802178722

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*. https://doi.org/10.1016/j.jesp.2018.08.009

Llorach, G., Grimm, G., Hendrikse, M. M. E., & Hohmann, V. (2018). Towards realistic immersive audiovisual simulations for hearing research capture, virtual scenes and reproduction. *AVSU 2018 - Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Co-Located with MM 2018*. https://doi.org/10.1145/3264869.3264874

Llorach, G., Grimm, G., Vormann, M., Hohmann, V., & Meis, M. (2020). *Vehicle driving actions for loudness and annoyance perception*. Zenodo. https://zenodo.org/record/3822311

Llorach, G., Oetting, D., Krüger, M., Vormann, M., Fitschen, C., Schulte, M., Hohmann, V., & Meis, M. (2019). Vehicle noise: Loudness ratings, loudness models and future experiments with audiovisual immersive simulations. *INTER-NOISE 2019 MADRID - 48th International Congress and Exhibition on Noise Control Engineering*.

Llorach, G., Oetting, D., Vormann, M., Fitschen, C., Krüger, M., Schulte, M., Meis, M., & Hohmann, V. (2022). *Loudness and annoyance ratings of vehicle noise*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.6519277

Mershon, D. H., Desaulniers, D. H., Kiefer, S. A., Amerson, T. L., & Mills, J. T. (1981). Perceived loudness and visually-determined auditory distance. *Perception*, *10*(5). https://doi.org/10.1068/p100531

Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine*, *25*(4). https://doi.org/10.1002/sim.2324

Oetting, D., Bach, J.-H., Krueger, M., Vormann, M., Schulte, M., & Meis, M. (2020). Subjective loudness ratings of vehicle noise with the hearing aid fitting methods NAL-NL2 and trueLOUDNESS. *Proceedings of the International Symposium on Auditory and Audiological Research, Vol. 7*, 289–296.

Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research*, *335*. https://doi.org/10.1016/j.heares.2016.03.010

Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Ewert, S. D. (2018). Restoring perceived loudness for listeners with hearing loss. *Ear and Hearing*, *39*(4). https://doi.org/10.1097/AUD.0000000000000521

Parker, R. I., & Vannest, K. (2009). An Improved Effect Size for Single-Case Research: Nonoverlap of All Pairs. *Behavior Therapy*, *40*(4). https://doi.org/10.1016/j.beth.2008.10.006

Patsouras, C. (2003). *Geräuschqualität von Fahrzeugen-: Beurteilung, Gestaltung und multimodale Einflüsse*. Shaker.

Smeds, K., Keidser, G., Zakis, J., Dillon, H., Leijon, A., Grant, F., Convery, E., & Brew, C. (2006). Preferred overall loudness. II: Listening through hearing aids in field and laboratory tests. *International Journal of Audiology*, *45*(1). https://doi.org/10.1080/14992020500190177

Chapter 4

Comparison between a Head-Mounted Display and a Curved Screen

Submitted to Acta Acustica as:

Llorach, G., Hendrikse, M. M. E., Grimm, G., & Hohmann, V. (2023). Comparison between a Head-Mounted Display and a Curved Screen in a Multi-Talker Audiovisual Listening Task.

Abstract

Introduction: Virtual audiovisual technology and its methodology has yet to be established for psychoacoustic research. This study examined the effects of different audiovisual conditions on preference when listening to multi-talker conversations. The study's goal is to explore and assess audiovisual technologies in the context of hearing research.

Methods: The participants listened to audiovisual conversations between four talkers. Two displays were tested and compared: a curved screen (CS) and a head-mounted display (HMD). Using three visual conditions (audio-only, virtual characters and video recordings), three groups of participants were tested: seventeen young normal-hearing, ten older normal-hearing, and ten older hearing-impaired listeners.

Results: Open interviews showed that the CS was preferred over the HMD for older normal-hearing participants and that video recordings were the preferred visual condition. Young and older hearing-impaired participants did not show a preference between the CS and the HMD.

Conclusions: CSs and video recordings should be the preferred audiovisual setup of laboratories and clinics, although HMDs and virtual characters can be used for hearing research when necessary and suitable.

Key words: virtual reality, head-mounted display, hearing impaired, technology acceptance.

4.1. INTRODUCTION

In recent years, audiovisual technologies have become more prevalent in hearing research (Ahrens et al., 2019; Assenmacher et al., 2005; Devesse et al., 2018; Hendrikse et al., 2019; Kohnen et al., 2016; Llorach, Oetting, Vormann, Meis, et al., 2022; Rummukainen, 2016; Schutte et al., 2019; Seol et al., 2021; Stecker, 2019; van de Par et al., 2022). One of the motivations of using such technologies is to increase the ecological validity of the experiments in the laboratory and the clinic (Keidser et al., 2020), i.e., that the results in the laboratory reflect real-life hearing-related function (Bentler, 2005; Cord et al., 2004). This is particularly important when fitting hearing aids for the first time. New users tend to give up using hearing aids if these don't improve their hearing situation in their daily environments, thus leading to a poorer quality of life in the long term (McCormack & Fortnum, 2013; Tareque et al., 2019).

Audiovisual technologies such as head-mounted displays (HMDs) and surrounding screens are already established and are available in the market. For hearing and hearing aid research, however, the applicability and acceptance of different audiovisual technologies has hardly been investigated. Seol et al. (2021) tested speech perception with and without HMDs and asked the participants about technology preference and its applicability in the clinic. Almost all participants were willing to complete the test in a clinical setup with the HMD, but the weight of the device and the participant's unfamiliarity with it were concerning issues. Seol et al. (2021) mentioned it is crucial to test and validate the audiovisual technology used in audiological experiments, "as it could be one of many factors that professionals and patients would consider before employing and performing the test in clinics". More data are therefore

needed to characterize and establish audiovisual technology options in hearing research (Llorach et al., 2018).

Individual characteristics, in particular age and hearing status, could elicit different technology acceptance. Most research in technology acceptance has been done with young normal hearing (YNH) participants, thus data is lacking for older and hearing-impaired participants. Philpot et al. (2017) found no difference in preference for young adults between the CS system and the HMD when watching a 360-degree angle documentary. Hendrikse et al. (2018) found that YNH participants preferred video recordings over animated virtual characters in a listening task with a curved screen. Older and hearing-impaired participants might be more reluctant than YNH participants to use intrusive and immersive audiovisual technologies, although the opposite is possible as well. Surrounding screens and HMDs differ in several characteristics that could affect acceptance in laboratory measurements. A HMD is worn on the head and it occludes all visual references to the real space. The field of view is reduced, as current consumer-grade HMDs cannot cover the whole human field of view (210^{o} horizontal). The HMD requires head straps that could interfere with the positioning and performance of the hearing aids, and therefore could be uncomfortable for hearing aid users. Additionally, in Seol et al. (2021), HMDs were reported as heavy and difficult to use if one is not familiar with the device. With surrounding screens, the real space is always visible. The user's head movement and vision are less constrained, as no device is worn on the head. To understand such differences between displays and visual conditions, the current study provides comparative data for different readily available audiovisual technology options, in particular for curved screens (CSs) vs. head-mounted displays (HMDs), and for video recordings (VID) vs. virtual

characters (VC) vs. audio-only (AO). Technology acceptance is particularly important for the applicability of such technologies in clinical setups. If we want to use such technologies in the clinic, we must make sure that older and hearing-impaired participants are willing to use such technologies and that these technologies do not deter the participants from getting their hearing abilities checked. In consideration of that, this study included YNH participants, older normal hearing participants (ONH), and older hearing-impaired participants (OHI).

In this experiment, participants listened to conversations in different audiovisual conditions and answered questions about the content afterwards; subjective ratings and open comments of preference and technology acceptance were collected and analyzed. The results of this study are meant to provide useful insight to guide future research and implementation in hearing clinics and research laboratories using audiovisual technologies.

This study replicates parts of the experimental setup of our previous study Hendrikse et al. (2018b). The current work extends it by adding the HMD as a display type, by testing older participants, with and without hearing impairment, and by measuring technology acceptance with open interviews. Head orientation and gaze were measured in the experiment, but the analysis and results are to be presented in a future article. A direct comparison between several audiovisual setups is presented here, as two display types were combined with three visual conditions. Additionally, older and older hearing-impaired participants were included to compare the applicability of the audiovisual setups in clinical environments.

4.2. Methods

The participants were asked to listen to conversations under six different conditions: two display types combined with three visual conditions. The display types were a CS and a HMD, and the visual conditions were audio-only, virtual characters, and video recordings. In Figure 4.1, the conditions can be seen as they looked in the experiment. The top row shows the conditions with the HMD and the lower row the ones with the CS. The task of the participants was to answer three questions about the content of the conversation they just heard. After completing all six conditions, they had to do an interview and fill out questionnaires. The CS-AO condition was meant to represent an experiment in a hearing laboratory without visual cues, which is the standard case in hearing research. To counterbalance the number of conditions done with the CS, the HMD-AO condition was included.

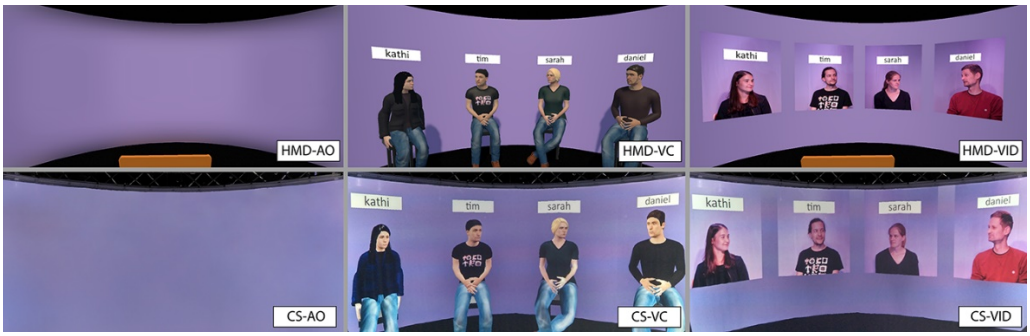


Figure 4.1. Images of the six conditions presented in this experiment. The images on the top row are screen captures of the virtual space used for the head-mounted display (HMD). The virtual space contained elements of the real space such as the chair where the participants were seated. The images on the bottom row are pictures of the curved screen (CS) taken inside the laboratory. From top to bottom: conditions with the HMD and the curved screen. From left to right: audio-only (AO), virtual characters (CS), and video recordings (VID).

4.2.1. Participants

Seventeen young normal-hearing subjects (YNH), ten older normal-hearing subjects (ONH), and ten older moderately hearing-impaired subjects with hearing aids (OHI) participated in the study. All but one of the YNH subjects were students of the Carl von Ossietzky Universität Oldenburg with a mean age of 24 years (STD 2.43, range 18-27). YNH subjects were specifically asked about their hearing: none of them reported hearing loss. The mean age of the older participants was 61.9 years (STD 5.3, range 50-69). ONH and OHI participants were recruited through Hörzentrum Oldenburg GmbH, where their audiograms were measured regularly: ONH participants had a mean pure tone average (PTA) between 125 Hz and 8 kHz of 10 dB HL; the mean PTA between 125 Hz and 8 kHz for OHI participants was 49.4 dB HL. The OHI participants had been using their hearing aids for more than six months and had a moderate symmetric hearing loss. They wore their hearing aids during the experiment. Participants were also specifically asked about visual impairments, which none of them reported (e.g., reduced vision not corrected by glasses or contact lenses). The ethics permission was granted by the ethics committee of the CvO Universität Oldenburg (Drs. 1r63/2016). The participants signed an informed consent.

4.2.2. Setup

The experiment was conducted inside a circular 'tent' within an acoustically semi-treated room (reverberation time (T60) = 0.13s). The inside of the tent and a top view of the room can be seen in Figure 4.2. The figure shows where the participant was sitting and how the projection looked for the CS-VID condition. The position of the elements of the tent is also shown in Figure 4.2. The tent was covered with a black blanket and it had a radius of 1.98 meters.

It consisted of a metal structure that supported a circular array of 16 loudspeakers (Genelec 8020B, Genelec Oy, Olvitie, Finland) and an acoustically transparent curved screen. The loudspeakers were spaced every 22.5-degree angle at a radius of 1.96 meters and a height of 1.60 meters. The curved screen was in front of this array of loudspeakers and was 2 meters tall with a 1.76-meter radius. Images were projected onto the screen from a close-field projector (NEC U321H, Sharp NEC Display Solutions, Munich, Germany) placed on top of the tent. The projector achieved a projection of 120-degree angle (horizontal) and had a refresh rate of 60 Hz with a resolution of 1920x1080 pixels. The HTC Vive system (HTC Corporation, New Taipei City, Taiwan) was used as HMD. The HTC Vive Base Stations and a camera for live feedback were placed above the curved screen. The HTC Vive display had a refresh rate of 90 Hz, a resolution of 1080x1200 pixels per eye, a 100-degree angle field of view (horizontal) and orientation and translation tracking. The background noise level inside the tent with all the devices working was 31.1 dB A.

A chair was placed in the center of the tent, facing towards the front, i.e., the 0-degree angle azimuth of the simulation. The chair was on an elevated platform with dimensions 120 cm by 120 cm. The platform was elevated 30 cm from the floor. When the participants were seated, the ears were at approximately the same level as the loudspeakers (1.60 meters). To the side of the participant, around 120-degree angle azimuth from the front, there was an emergency button at arm's reach: pressing this button stopped the simulation.

Three computers were used in the experiment: an Ubuntu 14.04 for the acoustic rendering, data logging and master control; an Ubuntu 14.04 for the screen projection with NVIDIA Quadro K6000; and a Windows 10 for the HMD rendering with NVIDIA Quadro M5000 and head tracking.

The 3D virtual acoustic environment was rendered with TASCAR (Grimm et al. 2019b) versions 0.175.2-0.177.5. The virtual 3D scene for the curved screen was created and rendered with the Blender Game Engine version 2.79 (Roosendaal 1995). The image warping for the projection was done with the graphics card and was manually configured and calibrated. The 3D scene for the HMD was rendered with the Unity game engine version 2017.1.0f3. All the sensor data was transmitted for central data logging in TASCAR via the Open Sound Control (OSC) (Wright and Freed 1997) and the LabStreamingLayer protocol (Kothe et al. 2018). The experiment was controlled and executed with Matlab 2016b and with the acoustic engine using the OSC protocol. Temporal alignment between visual and acoustic cues was adjusted manually.

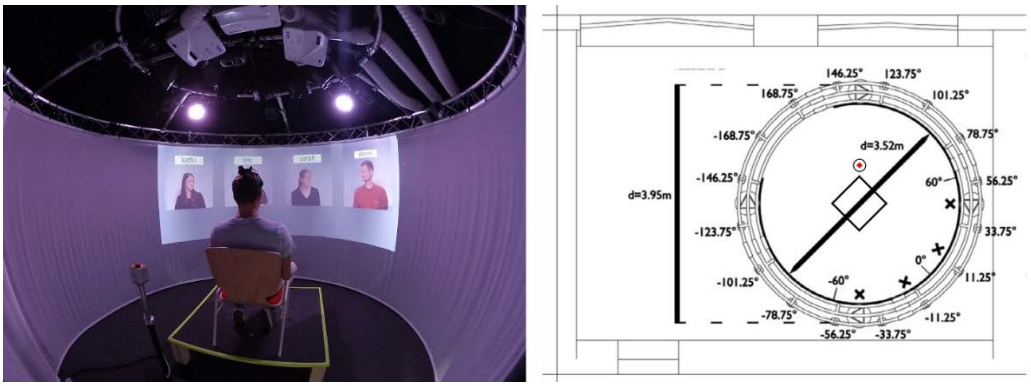


Figure 4.2. A, On the left: fish-eye picture of the inside of the tent in the condition with the curved screen and the video recordings. B, On the right: top view of the tent and the room. The angles on the outside of the metal ring (circular structure) indicate the position of the loudspeakers. The crosses indicate the position of the target speakers in this experiment. The square in the middle represents the platform where the participant was seated. The circle with a red dot, close to the platform, depicts the emergency button.

Head orientation was measured with two different devices for the CS and the HMD. For the CS, participants wore a head crown with a Vive Tracker (HTC Corporation, New Taipei City, Taiwan) attached. For the HMD, the device itself, i.e., the HTC Vive, was used for head tracking. The horizontal movement of the eyes was measured with two electrodes placed next to the eyes (electrooculography, EOG).

4.2.3. Stimuli

We used the same audiovisual material, casual acted conversations, as in our previous study (Hendrikse et al. 2018b). There are seven conversations available, one of which we used for the training and the six remaining for the experiment conditions. The material can be found in the database by Hendrikse et al. (2018a). The conversations lasted between 1 min 24 s and 1 min 39 s and the topics were food, holidays/travelling, weather, work, future plans, movies and anecdotes. Of the four talkers, two were females and fluent non-native talkers (German CEFR C1), and the other two were males and native talkers. In the 3D virtual scene, the actors were positioned at 45, 15, -15 and -45-degree angle in a radius of 1.7 meters away from the listener's position. After each conversation, one of the actors asked three multiple-choice questions about the content. In this experiment, we did not record the answers to the questions but this was unknown to the participants in order to keep them engaged. These questions can be found in the aforementioned database (Hendrikse et al. 2018a).

Acoustic Stimuli

The acoustic conditions were the same across all trials. The multi-talker conversations were played together with diffuse background noise. In our laboratory, the loudspeaker layout did not match the position of the target

talkers (see Figure 4.2). We used TASCAR to generate a virtual acoustic environment and to reproduce sound sources at the prescribed place. This virtual acoustic environment simulated a virtual source for each target talker at the predefined position, as if the clean speech was played back through a loudspeaker in the room at the respective position. The audio reproduction technique used for the target talkers was horizontal 7th-order Ambisonics with max-rE decoding (Daniel et al. 1998), rendered to the 16 loudspeakers on a ring at ear level. The diffuse background noise was a 1st-order Ambisonics recording of the cafeteria of the University of Oldenburg (Hendrikse et al. 2018a). To achieve a diffuse reproduction of the background sound field and to avoid spectral artifacts due to self-motion, the first-order signal was extended to 7^{th} order, and a frequency-dependent rotation similar to the method of Zotter et al. (2014) was applied. The average sound levels for each conversation were measured with a sound level meter at the position of the listener. The sound levels for the YNH were 45.2 $\pm$ 0.3 dB A for the conversations and 49.7 dB A for the cafeteria background noise. For the older participants, the speech levels had to be increased and the noise levels reduced, as the first two older participants complained that they could not hear the spoken instructions clearly inside the simulation (speech in quiet). The levels for the ONH were adjusted with an increase of 3.1 dB for speech (48.3 dB A) and a decrease of 3.7 dB for noise (46.0 dB A). The levels for the OHI were adjusted with an increase of 9 dB for speech (54.2 dB A) and a decrease of 6 dB for noise (43.7 dB A). These level adjustments were defined by the first participants (first ONH and first OHI). The speech level was raised so that they could understand the speech in quiet and the background noise level was reduced so that the conversation could be followed when background noise was present. We were

aiming for a realistic speech level around 65 dB A, but due to a calibration error, the actual speech levels were not in this range.

Visual Stimuli

Three different visual conditions were presented in this experiment (see Figure 4.1): audio-only (AO), virtual characters (VC) and video recordings (VID). In the CS-AO condition, the projection was turned off and a diffuse light was turned on. In the HMD-AO condition, a virtual laboratory was shown, so the participant would feel he/she was in the same real space and would have some reference points: the participant could see the chair underneath, the platform where the chair was, the cylindrical screen and the emergency button. This virtual scene was used for the other visual conditions. For the VC condition, the 3D virtual characters were created with Makehuman version 1.02 in resemblance to the real actors. The virtual characters were blinking and moving their lips with a speech-based lip-syncing (Llorach et al. 2016). The virtual characters also moved their head and eyes: they followed the conversation by looking towards virtual character who was speaking. These three animations were automated and generated in real-time. The effects of these animations can be found in the studies by Grimm et al. (2019b) and Hendrikse et al. (2018b). In the VID condition, the video recordings were shown through flat screens in the virtual scene (see Fig 1).

Experiment Procedure

The participants filled in an anonymization form and an informed consent. They were informed about the experiment through written forms, a video clip and orally. The interpupillary distance was measured with a ruler and the lenses of the HMD were adjusted accordingly. The head crown and the HMD were adjusted to the participant's comfort. If the participants used corrective

glasses, we let them try the HMD with and without them; they decided whether they wanted to do the HMD trials with or without glasses. The EOG electrodes were attached to the participant together with a Bluetooth transmitter and participants were instructed not to touch them during the experiment. They were instructed that they would have to answer verbally 'A', 'B' or 'C', to the multiple-choice questions presented after each conversation. After this introduction, they filled out the pre-exposure Simulator Sickness Questionnaire (SSQ; Kennedy et al. 1993) and were seated on the chair inside the tent. We included the SSQ in the experiment to assess whether participants suffered from cybersickness.

The participants started with the HMD or the curved screen randomly. They did the three randomized visual conditions with one display followed by three more with the other display. The order of the visual conditions was the same with the curved screen and the HMD for each participant. The conversations were randomized and each conversation was played equally often for each condition across all participants. The EOG required a calibration protocol, which was done once for the CS and once for the HMD before starting the trials.

Instructions about the task were repeated through a virtual character in the simulation. When using the HMD, an initial adaptation phase was added: a virtual character made suggestions for getting used to the room, to look at the chair they were sitting on and to find the emergency button behind them. If they did not find the emergency button, the researcher came inside the tent and made sure the participant could turn and see the button. The virtual button was in the same location as the physical one. This procedure was done to adapt the participants to the experience, e.g., some participants may be

unaware that they can move or turn their heads with the HMD. This adaptation phase lasted around 1 minute.

After the instructions, there was a training trial. The training trial used a conversation that was not used in the test trials. After each conversation, the participants answered verbally to the multiple-choice related questions. The participants came out of the tent to fill out the SSQ after all trials were completed. After this, we proceeded with the open interview recorded with a sound recorder.

4.2.4. Measures

The preference and acceptance of the audiovisual conditions were measured via a recorded interview. The participants were asked to give comments and impressions about the experiment once they completed all listening tasks. They were given a paper with six pictures (one for each condition) and a picture of each display device. We allowed a minimum of three minutes time and a maximum of 15 minutes for comments. Afterwards, the participants were asked to select one of the six conditions (see Figure 1) as the one they would like to experience in a future experiment. Then, they were asked to name the second-best condition. Finally, they were asked to choose if there was any condition they would not like to experience again. The participants that did not have a preference between displays or visual conditions could also answer combinations, i.e., first preference as the video regardless of the display. The increase in SSQ symptoms between pre- and post-exposure questionnaire was computed and the mean values for the total simulator sickness severity were below 13 for both groups. According to Kennedy et al. (1993), the cybersickness reported in this experiment is considered insignificant (10-15 Total Severity).

4.3. Results

4.3.1. Open Comments

We analyzed the recorded interviews and annotated the issues that were mentioned: these are summarized in Table 4.1. The interviews revealed that the speech was difficult to understand (Table 4.1 item 2); some subjects found the males talkers more difficult to understand (Table 4.1 items 3-5); some found the accent of the non-native female talkers hard to understand (Table 4.1 item 6). Three participants mentioned that moving their head changed their audio perception (Table 4.1 item 7). Five participants mentioned that the HMD was heavy and three older participants commented that they felt isolated when wearing the HMD (Table 4.1 items 8-9). Three YNH participants noticed that the screen of the HMD was brighter than the CS (Table 4.1 item 10). Seven participants mentioned that in the AO trials it was easier to concentrate than in the other trials, but for three participants it was the opposite (Table 4.1 items 12-13). Additionally, eight participants mentioned that it was easier to understand the conversation in the VID condition (Table 1 item 14). Four OHI and two ONH participants complained about the insufficient resolution of the lips of the virtual characters (Table 4.1 item 16), four participants mentioned that the virtual characters were too stiff (Table 4.1 item 17) and seven participants indicated that the characters were not realistic (Table 4.1 item 15).

Table 4.1. Comments by the participants during the open interviews. Only comments mentioned by three or more participants were noted in this table.

	Nº YNH out of 17	Nº ONH out of 10	Nº OHI out of 10	Total nº out of 37
Comments about the conversations and the acoustics				
1. It was hard to concentrate	2	1	3	6
2. It was difficult to understand	7	4	4	15
3. It was easier to listen to the female talkers	3	2	0	5
4. Daniel (+45-degree angle) was really hard to understand	0	2	2	4
5. Tim (-15-degree angle) was really hard to understand	2	2	0	4
6. It was hard to understand the accent	1	2	1	4
7. The head position changed the audio perception	1	2	0	3
Comments about the display				
8. I felt isolated with the head-mounted display (HMD)	0	1	2	3
9. The HMD was heavy	2	1	2	5
10. The image was brighter with the HMD	3	0	0	3
11. Wearing the HMD was distracting	1	2	0	3
Comments about the visual condition				
12. It was easier to concentrate in the audio-only (AO) condition	4	1	2	7
13. It was harder to concentrate in the AO condition	3	0	0	3
14. It was easier to listen to the video recordings	5	3	0	8
15. The virtual characters (VCs) were not realistic	4	0	3	7
16. The lips were not readable with the VCs	0	2	4	6
17. The VCs were too stiff	3	0	1	4

Table 4.2. Preferences for the visual conditions and displays.

		Chosen as 1st or 2nd condition			
		Nº of YNH out of 17	Nº of ONH out of 10	Nº of OHI out of 10	Total nº of participants out of 37
Visual condition	Video recordings	15	9	9	33
	Virtual characters	6	4	3	13
	Audio-only	5	3	3	11
Display	Head-mounted display	15	6	8	29
	Curved screen	15	10	9	34
		Never again condition			
		Nº of YNH out of 13	Nº of ONH out of 10	Nº of OHI out of 10	Total nº of participants out of 33
Visual condition	Video recordings	0	0	0	0
	Virtual characters	3	1	3	7
	Audio-only	6	0	5	11
Display	Head-mounted display	5	1	4	9
	Curved screen	1	0	1	2

4.3.2. Chosen conditions

The answers of the participants are shown in Table 4.2. We divided the preference results by visual conditions and display. The first and second

preferences were grouped together, e.g., the VID condition was chosen by thirty-four participants out of thirty-eight as the first and/or second preference. The first four subjects were not asked whether there was any condition they would not like to do again. Eighteen participants out of thirty-four were willing to do all the conditions again.

The VID condition was clearly chosen as the preferred visual condition and was never rejected. The other two visual conditions, VC and AO, were chosen with nearly equal preference. The YNH and the OHI participants showed no preference between the HMD and the CS displays. The ONH preferred the CS more often (all ONH participants chose the CS and six chose the HMD as first/second condition out of ten). In general, the HMD was more frequently rejected by the YNH and the OHI and the CS was rejected only by two participants out of thirty-three. The rejected conditions were always a combination of a display (HMD or CS) with the AO or VC condition. The AO condition was rejected by four participants more than the VC condition (eleven vs. seven participants out of thirty-three). No ONH participants rejected the AO condition, whereas almost half of the YNH and the OHI rejected it.

4.4. Discussion

As expected based on the study by Philpot et al. (2017), the YNH participants showed equal preference for the two displays. The ONH participants preferred the CS over the HMD, but almost equal preference was shown for the OHI participants. The HMD was rejected more often as a display and received more negative comments, such as that it was heavy, isolating and distracting. Therefore, the CS would be a better choice for the comfort of the participants. Nevertheless, the HMD was chosen quite often as a first or second

option and it was never rejected as a display alone, i.e., regardless of the visual condition. Most OHI were willing to use the device for future experiments. Therefore, it should be considered for clinical implementations, as a cheaper and simpler implementation.

The video recording condition (VID) was clearly the most preferred visual condition. This finding agreed with the previous study by Hendrikse et al. (2018b). The AO condition was the most rejected (by eleven participants out of thirty-three), showing a general preference for conditions with visual cues. The comments regarding the VC condition indicated that their quality, non-verbal behaviors and lip-readability should be improved. It is worth noticing that only the older participants (two ONH and four OHI participants) mentioned the lip-readability, indicating that older participants might look for this kind of visual cues specifically.

The accent was brought up as a difficulty for understanding, but the female talkers, which were the ones with the accent, were also specified to be easier to understand by other participants. We consider that the talkers were all equally intelligible, according to the open comments.

4.4.1. Outlook and Limitations

Depending on the research question, current hearing clinics and laboratories might want to use immersive visual cues (Keidser et al. 2020). Changing their methodologies from audio-only to audiovisual stimuli might be expensive and effortful. HMDs are more affordable and easier to setup than custom-built CSs. Nevertheless, specific procedures need to be done in the clinic for HMDs, such as measuring the interpupillary distance and adjusting the head-straps. Additionally, head-mounted displays introduce acoustic distortions (Genovese

et al., 2018; Lladó et al., 2022) that might affect the results collected in the clinic. Which audiovisual system to use will depend on each specific experiment and clinical setup.

Next generations of HMDs might improve some of the issues mentioned by the participants, such as the weight of the device. Mixed reality and augmented reality solutions should also be considered, as they might be less isolating than HMDs.

Further improvements need to be done to our virtual characters if they are to be used (see Llorach et al. 2018). This is supported by the comments of the participants and the significant differences found between the video recordings and the virtual characters. In this study we used open-source characters and animations available at the time.

ACKNOWLEDGMENTS

Thanks to: K. Schwarte for her help during the measurements and the analysis of the recorded conversations; M. Krüger and M. Zokoll for the support on recruiting older participants; A. Wagner for counseling; and J. Luberadzka for the helpful comments. This study was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675324 (ENRICH) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 352015383 - SFB 1330 B1.

Data Availability Statement

The audiovisual stimuli associated with this article is available in Zenodo ("Audiovisual recordings of acted casual conversations between four speakers in German"), under the reference https://doi.org/10.5281/zenodo.1257333.

References

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLoS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0214603

Assenmacher, I., Kuhlen, T., & Lentz, T. (2005). Binaural acoustics for CAVE-like environments without headphones. *9th International Workshop on Immersive Projection Technology - 11th Eurographics Symposium on Virtual Environments, IPT/EGVE 2005*.

Bentler, R. A. (2005). Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. In *Journal of the American Academy of Audiology* (Vol. 16, Issue 7). https://doi.org/10.3766/jaaa.16.7.7

Cord, M. T., Surr, R. K., Walden, B. E., & Dyrlund, O. (2004). Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology*, *15*(5). https://doi.org/10.3766/jaaa.15.5.3

Devesse, A., Dudek, A., van Wieringen, A., & Wouters, J. (2018). Speech intelligibility of virtual humans. *International Journal of Audiology*, *57*(12). https://doi.org/10.1080/14992027.2018.1511922

Genovese, A., Zalles, G., Reardon, G., & Roginska, A. (2018). Acoustic perturbations in HRTFs measured on mixed reality headsets. *Proceedings of the AES International Conference*, *2018-August*.

Grimm, G., Llorach, G., Hendrikse, M. M. E., & Hohmann, V. (2019). Audio-visual stimuli for the evaluation of speech-enhancing algorithms. *Proceedings of the International Congress on Acoustics*, *2019-September*. https://doi.org/10.18154/RWTH-CONV-238907

Hendrikse, M. M. E., Grimm, G., Llorach, G., & Hohmann, V. (2018). *Audiovisual recordings of acted casual conversations between four speakers in German*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.1257333

Hendrikse, M. M. E., Llorach, G., Grimm, G., & Hohmann, V. (2018). Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, *101*. https://doi.org/10.1016/j.specom.2018.05.008

Hendrikse, M. M. E., Llorach, G., Hohmann, V., & Grimm, G. (2019). Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life. *Trends in Hearing*, *23*. https://doi.org/10.1177/2331216519872362

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science: What It Is, Why It

Matters, and How to Advance It. *Ear and Hearing*, *41*. https://doi.org/10.1097/AUD.0000000000000944

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, *3*(3). https://doi.org/10.1207/s15327108ijap0303_3

Kohnen, M., Stienen, J., Aspöck, L., & Vorländer, M. (2016, July). Performance Evaluation of a Dynamic Crosstalk-Cancellation System with Compensation of Early Reflections. *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*.

Lladó, P., Mckenzie, T., Meyer-Kahlen, N., & Schlecht, S. J. (2022). Predicting Perceptual Transparency of Head-Worn Devices. *Journal of the Audio Engineering Society*, *70*(7/8), 585–600.

Llorach, G., Grimm, G., Hendrikse, M. M. E., & Hohmann, V. (2018). Towards realistic immersive audiovisual simulations for hearing research capture, virtual scenes and reproduction. *AVSU 2018 - Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Co-Located with MM 2018*. https://doi.org/10.1145/3264869.3264874

Llorach, G., Oetting, D., Vormann, M., Meis, M., & Hohmann, V. (2022). Vehicle noise: comparison of loudness ratings in the field and the laboratory. *International Journal of Audiology*, 1–10. https://doi.org/10.1080/14992027.2022.2147867

McCormack, A., & Fortnum, H. (2013). Why do people fitted with hearing aids not wear them? *International Journal of Audiology*, *52*(5). https://doi.org/10.3109/14992027.2013.769066

Philpot, A., Glancy, M., Passmore, P. J., Wood, A., & Fields, B. (2017). User experience of panoramic video in CAVE-like and head mounted display viewing conditions. *TVX 2017 - Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. https://doi.org/10.1145/3077548.3077550

Rummukainen, O. (2016). *Reproducing reality: Perception and quality in immersive audiovisual environments* [Doctoral thesis]. Aalto University.

Schutte, M., Ewert, S. D., & Wiegrebe, L. (2019). The percept of reverberation is not affected by visual room impression in virtual environments. *The Journal of the Acoustical Society of America*, *145*(3). https://doi.org/10.1121/1.5093642

Seol, H. Y., Kang, S., Lim, J., Hong, S. H., & Moon, I. J. (2021). Feasibility of Virtual Reality Audiological Testing: Prospective Study. *JMIR Serious Games*, *9*(3). https://doi.org/10.2196/26976

Stecker, G. C. (2019). Using Virtual Reality to Assess Auditory Performance. In *Hearing Journal* (Vol. 72, Issue 6). https://doi.org/10.1097/01.HJ.0000558464.75151.52

Tareque, M. I., Chan, A., Saito, Y., Ma, S., & Malhotra, R. (2019). The Impact of Self-Reported Vision and Hearing Impairment on Health Expectancy. *Journal of the American Geriatrics Society*, *67*(12). https://doi.org/10.1111/jgs.16086

van de Par, S., Ewert, S. D., Hladek, L., Kirsch, C., Schütze, J., Llorca-Bofí, J., Grimm, G., Hendrikse, M. M. E., Kollmeier, B., & Seeber, B. U. (2022). Auditory-visual scenes for hearing research. *Acta Acustica*, *6*, 55. https://doi.org/10.1051/aacus/2022032

Zotter, F., Frank, M., Kronlacher, M., & Choi, J.-W. (2014). Efficient phantom source widening and diffuseness in ambisonics. *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, *2*(April).

Chapter 5

General Discussion

In this thesis the effects of visual cues were investigated in speech reception, loudness perception and technology preference. Ecological validity was considered in the design of the experiments. Several audiovisual technologies were used to better understand their applicability in clinical setups.

In Chapter 2, video recordings of a talker were added to an audio-only speech intelligibility test. It was found that when adding visual cues, the speech reception increased, but that this increase was highly individual in a homogeneous group. This finding shows that when speech perception is only evaluated in audio-only experiments, the ability to communicate in face-to-face conversations is not fully captured. Some individuals might be able to lipread quite a lot of the content, whereas others not. Based on this information, audiologists could provide better recommendations and counseling to the patients. For example, participants with speechreading skills could be advised to situate themselves where they can see the faces of all participants in a conversation, as their speech reception would improve.

To the best of authors knowledge, this is the first audiovisual Matrix Sentence Test (MST) that uses audio dubbed with video recordings. Other audiovisual MST recorded new audiovisual speech instead of reusing previous balanced acoustic speech (Jamaluddin, 2016; van de Rijt et al., 2019). The main disadvantage of creating new material is that the results are not directly

comparable to previous studies. With the dubbed MST it was possible to compare the audio-only results to the literature and to evaluate the audiovisual benefit of the video recordings. Although small asynchronies were present in the audiovisual material, the audiovisual benefit is comparable to the literature (about -5 dB SNR and 7 dB SPL benefit) (van de Rijt et al., 2019), thus showing that these asynchronies did not detriment the expected audiovisual benefit. The main limitation of the German audiovisual MST is its ceiling effect: good speechreaders reached unexpected thresholds where acoustic speech was not audible. The test is precise enough to differentiate good from bad speechreaders, which is relevant for clinical recommendations.

The tools developed for creating synchronous visual dubbing have been published on a open-source repository (Llorach & Loïc Le Rhun, 2022) and are currently being used to develop the French version of the Matrix Sentence Test. They provide scripts and interfaces to facilitate the recording, selection, and cutting of the video recordings. With such tools, the development of audiovisual versions of the MST should be facilitated. Even more, these tools permit to extend any audio-only speech tests to its audiovisual version by using the published guidelines. To further validate this dubbing method, future work should evaluate the differences between a dubbed MST and a synchronously recorded MST.

The test developed here has been used in the literature already, thus proving the applicability and reproducibility of the test on different research topics: audiovisual integration in older populations (Gieseler et al., 2020), speech intelligibility when wearing a facial mask during the COVID-19 pandemic in normal and hearing-impaired populations (Sönnichsen, Llorach, Hochmuth, et al., 2022; Sönnichsen, Llorach, Hohmann, et al., 2022), the effects of

audiovisual speech on listening effort (Ibelings et al., 2019), and neural activity during audiovisual speech processing (Bálint et al., 2022).

Chapter 3 presents the first comparison, as far as the author knows, between laboratory and field loudness ratings of the same stimuli. The results showed that as the realism of the laboratory simulations increased, the ratings resembled more the ones obtained in the field. This finding remarks on the importance of evaluating loudness perception in realistic simulations. We did not find differences in loudness ratings between normal-hearing participants and hearing-impaired participants with hearing aids. Still, future research should include both groups, as Smeds et al., (2006) found group differences when measuring loudness gain preferences.

Further improvements could be made to the acoustics of the laboratory simulation, such as using surrounding audio in an anechoic room. Nevertheless, the laboratory setups used in Chapter 3 were chosen for its applicability in the clinic. Using a head-mounted display with stereo loudspeakers should be relatively feasible in a clinical environment, even more when the stimuli are in the format of 360º videos with stereo audio, i.e., a 3D virtual environment is not required (see Llorach et al. 2018). The research presented in Chapter 3 is still far from being standardized and transformed into an international loudness perception test. Future research should focus on standardizing loudness tests with realistic and everyday sounds that are causing loudness discomfort. This way the appropriate loudness settings of a hearing aid could be set up in the laboratory to improve the experience of the recipient in its everyday life. The data and stimuli have been published openly in hope that other researchers and audiologist can follow up the work (Llorach, Grimm, et al., 2020; Llorach, Oetting, Vormann, Fitschen, et al., 2022).

In Chapter 4 the applicability of different audiovisual setups for hearing research was studied in a multi-talker listening task. The most relevant contribution of this study is that head-mounted displays are accepted by older participants, with and without hearing impairment. This is particularly important for establishing clinical tests that use such technologies, as mentioned in (Seol et al., 2021).

Ecological Validity

Ecological validity is a goal or direction to follow when designing and performing experiments, but not something achievable. Researchers can aim at more ecological validity by, for example, increasing the resemblance of their tests to the real-life situation they want to investigate. Nevertheless, the intrinsic characteristics of laboratory experiments and the fact that we are measuring data mean that the results found cannot be 100% ecologically valid. As mentioned by Keidser et al. (2020): "no experiment is free of all threats to ecological validity".

In this thesis, it was assumed that an increase in the realism of laboratory simulations meant an increase in ecological validity. Of course, such assumption cannot be generalized, and it can only be considered for certain experimental setups and paradigms, e.g., being able to see the face of the speaker on a speech intelligibility task does not reflect the speech reception of a telephone call (audio-only). In the following paragraphs the ecological validity of each chapter is discussed.

In the case of speech perception (Chapter 2), being able to see the speaker is common in face-to-face communication or in video conferencing. We wanted to

evaluate speech perception in face-to-face communication, thus adding visual cues in an audio-only experiment was a clear step towards ecological validity.

In Chapter 3, loudness perception of vehicles was evaluated in the field and in the laboratory. It was assumed that the field measurements were more ecologically valid than the laboratory measurements because the participants were actually in the street where the vehicles were driving. Of course, it can be discussed that we do not experience loudness in real-life by sitting next to a road and by paying attention to the loudness of vehicles, as in the field experiment. Yet, the goal of the experiment was to compare the field to the laboratory ratings and to find out how realistic a laboratory setup needs to be to obtain the same results as in the field. The laboratory setup with the head-mounted display and stereo audio was found to elicit similar loudness perception as in the field, thus suggesting that in a laboratory experiment with a similar setup, the loudness perception should be more ecologically valid (or at least more similar to the field perception).

Technology preference and acceptance of two immersive visual systems were measured in Chapter 4 in a multi-talker conversation. Immersive simulations are a key towards ecological validity (Keidser et al., 2020) and validating their applicability in hearing research is a necessary step.

The path for creating standard tests that use head-mounted displays and virtual reality is in its beginning. In this thesis, only 68 participants were tested with head-mounted displays, and only in one study the participants were particularly asked about their experience and preference. More studies will be required to understand the implications of using virtual reality in hearing evaluation and to establish standard procedures. For example, motion sickness

is a common issue in virtual reality (Llorach et al., 2014), which depends on the experimental design. Avoiding discomfort and unpleasant virtual reality experiences is a must, as participants will be reluctant to try and use such devices in standard procedures. It is most recommended to read and follow developer guidelines when designing experiments, such as the ones published by Yao et al. (2014). Aside from user experiences, the shape of head-mounted displays influences and distorts the sound reaching the ears (Genovese et al., 2018), which could affect hearing aid processing and its ecological validity.

These experiments are a small step towards realistic and engaging simulations for hearing research. With the current available technologies, it should be possible to simulate situations in immersive environments where the user can have an active role. Nevertheless, there are still many challenges and steps to take scientifically, as recent research has shown: to evaluate the effects of different audiovisual displays (as this study did) and the effects of visual cues on listening behavior (Hendrikse et al., 2018), to standardize the virtual environments and the experimental designs with virtual reality across scientific communities (Hendrikse et al., 2019; van de Par et al., 2022), to establish realistic hearing tests to measure specific hearing disabilities (Seol et al., 2021), and to transfer these tests to the clinics successfully.

Looking into the future, the patient will be able to fit and test the hearing aids in the clinic with immersive simulations. These realistic simulations would be specifically tailored to the problems the patient is experiencing in real-life. For example, a hearing-aid user is experiencing discomfort during family dinners and contacts his/her clinic. The clinic could create a virtual reality simulation of a dinner with several talkers where several crucial hearing

properties are tested. The audiologist would adjust the hearing aid settings systematically to find the best configuration for the patient's comfort.

REFERENCES

Bálint, A., Wimmer, W., Caversaccio, M., & Weder, S. (2022). Neural Activity During Audiovisual Speech Processing: Protocol For a Functional Neuroimaging Study. *JMIR Research Protocols*, *11*(6), e38407. https://doi.org/10.2196/38407

Genovese, A., Zalles, G., Reardon, G., & Roginska, A. (2018). Acoustic perturbations in HRTFs measured on mixed reality headsets. *Proceedings of the AES International Conference*, *2018-August*.

Gieseler, A., Rosemann, S., Tahden, M., Wagener, K. C., Thiel, C., & Colonius, H. (2020). Linking audiovisual integration to audiovisual speech recognition in noise. *OSF Preprints*, *September*.

Hendrikse, M. M. E., Llorach, G., Grimm, G., & Hohmann, V. (2018). Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, *101*. https://doi.org/10.1016/j.specom.2018.05.008

Hendrikse, M. M. E., Llorach, G., Hohmann, V., & Grimm, G. (2019). Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life. *Trends in Hearing*, *23*. https://doi.org/10.1177/2331216519872362

Ibelings, S., Holube, I., Schulte, M., & Krüger, M. (2019, March 6). Audiovisuelle Erweiterung des subjektiven Höranstrengungsmessverfahrens ACALES. *22. Jahrestagung Der Deutschen Gesellschaft Für Audiologie*.

Jamaluddin, S. A. (2016). Development and Evaluation of the Digit Triplet and Auditory-Visual Matrix Sentence Tests in Malay. In *University of Canterbury*.

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It. *Ear and Hearing*, *41*. https://doi.org/10.1097/AUD.0000000000000944

Llorach, G., Evans, A., & Blat, J. (2014). Simulator Sickness and Presence using HMDs : comparing use of a game controller and a position estimation system. *VRST '14 Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, 137–140.

Llorach, G., Grimm, G., Vormann, M., Hohmann, V., & Meis, M. (2020). *Vehicle driving actions for loudness and annoyance perception*. Zenodo. https://zenodo.org/record/3822311

Llorach, G., & Loïc Le Rhun. (2022). *Tools for dubbing acoustic speech with video recordings*. Online repository. https://github.com/gerardllorach/audiovisualdubbedMST

Llorach, G., Oetting, D., Vormann, M., Fitschen, C., Krüger, M., Schulte, M., Meis, M., & Hohmann, V. (2022). *Loudness and annoyance ratings of*

vehicle noise. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.6519277

Seol, H. Y., Kang, S., Lim, J., Hong, S. H., & Moon, I. J. (2021). Feasibility of Virtual Reality Audiological Testing: Prospective Study. *JMIR Serious Games*, *9*(3). https://doi.org/10.2196/26976

Smeds, K., Keidser, G., Zakis, J., Dillon, H., Leijon, A., Grant, F., Convery, E., & Brew, C. (2006). Preferred overall loudness. II: Listening through hearing aids in field and laboratory tests. *International Journal of Audiology*, *45*(1). https://doi.org/10.1080/14992020500190177

Sönnichsen, R., Llorach, G., Hochmuth, S., Hohmann, V., & Radeloff, A. (2022). How Face Masks Interfere with Speech Understanding of Normal-Hearing Individuals: Vision Makes the Difference. *Otology and Neurotology*, *43*(3). https://doi.org/10.1097/MAO.0000000000003458

Sönnichsen, R., Llorach, G., Hohmann, V., Hochmuth, S., & Radeloff, A. (2022). Challenging Times for Cochlear Implant Users – Effect of Face Masks on Audiovisual Speech Understanding during the COVID-19 Pandemic. *Trends in Hearing*, *26*, 233121652211343. https://doi.org/10.1177/23312165221134378

van de Par, S., Ewert, S. D., Hladek, L., Kirsch, C., Schütze, J., Llorca-Bofí, J., Grimm, G., Hendrikse, M. M. E., Kollmeier, B., & Seeber, B. U. (2022). Auditory-visual scenes for hearing research. *Acta Acustica*, *6*, 55. https://doi.org/10.1051/aacus/2022032

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in

Audiovisual Speech Perception. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00335

Yao, R., Heath, T., Davies, A., Forsyth, T., Mitchell, N., & Hoberman, P. (2014). *Oculus VR Best Practices Guide*. http://brianschrank.com/vrgames/resources/OculusBestPractices.pdf

Statement of own contributions

Contributions that each author lead are underlined in the following statement.

Article: Llorach, G., Kirschner, F., Grimm, G., Zokoll, M. A., Wagener, K. C., & Hohmann, V. (2022). Development and evaluation of video recordings for the OLSA matrix sentence test. *International Journal of Audiology*, *61*(4). https://doi.org/10.1080/14992027.2021.1930205

Author contributions:

Gerard Llorach Tó: research question formulation, recording and organization of the video material, study design and setup, data collection, data analysis, paper writing, review and editing.

Frederike Kirschner: recording and cutting of the video material.

Giso Grimm: research question formulation, recording and cutting of the video material, study design and setup, data analysis, paper review and editing.

Melanie A. Zokoll: study design and setup, data collection, data analysis, paper review and editing.

Kirsten C. Wagener: supervision, study design, data analysis, paper review and editing.

Volker Hohmann: supervision, research question formulation, study design, paper review and editing.

Article: Llorach, G., Oetting, D., Vormann, M., Meis, M., & Hohmann, V. (2022). Vehicle noise: comparison of loudness ratings in the field and the laboratory. *International Journal of Audiology*, 1–10. https://doi.org/10.1080/14992027.2022.2147867

Author contributions:

Gerard Llorach Tó: research question formulation, field and laboratory measurements, field recordings, study design and setup, data analysis, paper writing, review and editing.

Dirk Oetting: research question formulation, field measurements, study design and setup, data analysis, paper review and editing.

Matthias Vormann: research question formulation, field and laboratory measurements, field recordings, study design and setup, data analysis, paper review and editing.

Markus Meis: research question formulation, field measurements, study design, paper review and editing.

Volker Hohmann: supervision, research question formulation, study design, data analysis, paper review and editing.

Article: Llorach, G., Hendrikse, M. M. E., Grimm, G., & Hohmann, V. (2022). Comparison of a Head-Mounted Display and a Curved Screen in a Multi-Talker Audiovisual Listening Task. *Submitted to Acta Acustica.*

Author contributions:

Gerard Llorach Tó: research question formulation, study design and setup, data collection, paper writing, review and editing.

Maartje M.E. Hendrikse: research question formulation, recording of the video material, study design and setup, paper review and editing.

Giso Grimm: research question formulation, recording of the video material, study design and setup, paper review and editing.

Volker Hohmann: supervision, research question formulation, study design, paper review and editing.

I hereby declare that I have written the remaining thesis chapters independently and I have only used the specified bibliography and resources.

I hereby confirm that Gerard Llorach Tó contributed to the aforementioned studies as stated above.

List of publications

Llorach, G., Kirschner, F., Grimm, G., Zokoll, M. A., Wagener, K. C., & Hohmann, V. (2022). Development and evaluation of video recordings for the OLSA matrix sentence test. *International Journal of Audiology*, *61*(4). https://doi.org/10.1080/14992027.2021.1930205

Llorach, G., Oetting, D., Vormann, M., Meis, M., & Hohmann, V. (2022). Vehicle noise: comparison of loudness ratings in the field and the laboratory. *International Journal of Audiology*, 1–10. https://doi.org/10.1080/14992027.2022.2147867

Llorach, G., Hendrikse, M. M. E., Grimm, G., & Hohmann, V. (2022). Comparison of a Head-Mounted Display and a Curved Screen in a Multi-Talker Audiovisual Listening Task. *Submitted to Acta Acustica.*

Corpora

Llorach, G., Kirschner, F., Grimm, G., & Hohmann, V. (2020). *Video recordings for the female German Matrix Sentence Test (OLSA)*. Zenodo. https://zenodo.org/record/3673062

Llorach, G., Grimm, G., Vormann, M., Hohmann, V., & Meis, M. (2020). *Vehicle driving actions for loudness and annoyance perception*. Zenodo. https://zenodo.org/record/3822311

Curriculum Vitae

Gerard Llorach Tó, born in Barcelona on the 22nd February of 1991, studied a Bachelor's degree in Audiovisual Systems Engineering at Universitat Pompeu Fabra (2009 – 2013). During his studies he did an exchange program at the University of Adelaide, Australia, where he did his Bachelor's thesis (Honours Music Technology). After finishing the Bacherlor's degree, he worked at the Interactive Technologies Group (GTI) at Universitat Pompeu Fabra (2013 – 2017). He developed several computer graphics' technologies, published in several conferences, and worked in two European projects (IMPART, KRISTINA). While working at GTI, he started collaborating with the Department Medizinische Physik und Akustik at the University of Oldenburg (2015-2017). He developed several virtual environments and transferred some of the technology and knowledge developed at GTI. In 2017 he was awarded a Marie Sklodowska-Curie Fellowship at Hörzentrum gGmbH Oldenburg as an early-stage researcher in a European Training Network (ETN – ENRICH). During this time, he collaborated with several researchers, led different research projects, and finished a Master in Hearing Technology and Audiology at the University of Oldenburg. In 2020 his PhD officially started at the Auditory Signal Processing and Hearing Devices group at the University of Oldenburg. He is an assistant lecturer of Computer Graphics at Universitat Pompeu Fabra and works at the "Institut de Ciències del Mar" of the Spanish National Research Council (CSIC)