Carl von Ossietzky
Universität
Oldenburg

Self-conducted speech audiometry using automatic speech recognition

Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines

— **Doktor der Naturwissenschaften (Dr. rer. nat.)** —

angenommene Dissertation

von Herrn
Jasper Ooster
geboren am 30.06.1990 in Jever

24. Juni, 2021

Gutachter: Prof. Dr. Bernd T. Meyer
Zweitgutachter: Prof. Dr. Dr. Birger Kollmeier
Externe Gutachterin: Prof. Dr. Odette Scharenborg

Tag der Disputation: 18.10.2021

Abstract

Speech communication is one of the core elements of social interaction. The ability to hear and understand speech, especially in noisy environments, is often limited for listeners with a hearing impairment. This is a major issue, since 1.5 billion people are affected by some kind of hearing loss globally. Speech audiometry is an important method for assessing the speech recognition ability of a listener. While several well-established clinical measurement procedures exist, most come with the drawback of a high measurement effort, since a specialist needs to conduct the speech audiometric test. This can be a limiting factor for the clinical conduction of such tests and their overall accessibility. A self-measurement application utilizing automatic speech recognition (ASR) can overcome this limitation and even extend it to screening applications. However, even the best ASR systems produce some errors in the transcription of subjects' responses, which potentially decrease the measurement accuracy of speech audiometric self-measurement.

This thesis utilizes the German matrix sentence test as a development prototype for ASR-based measurements that are automatically conducted. It investigates the influence of ASR errors on the measurement accuracy using Monte-Carlo simulation methods and proposes self-measurement procedures using ASR for two different application scenarios, namely: a well-controlled laboratory environment with a locally running ASR system developed specifically for this purpose and a screening application using the ASR component of a smart speaker - i.e. a commercially available high-quality speaker connected to a virtual assistant, which results in less control over experimental parameters. The two systems proposed above are evaluated with 139 subjects in total - covering a wide range of hearing abilities: normal-hearing listeners, mildly-, moderately- and severely hearing-impaired subjects, as well as listeners with cochlear im-

plants. For the well-controlled environment, two application methods are considered: unaided measurements using headphone presentation of the stimuli, which results in clean audio recordings, as well as aided measurements using loudspeaker presentation, which results in noisy recordings. For the screening application, three different reverberation scenarios are considered, since the acoustic conditions are rather unclear in self-measurement at home.

Monte Carlo simulation methods are used to reproduce the adaptive measurement procedure and they show that the adaptive procedure of the matrix sentence test is generally very robust against errors from the ASR system, which in turn shows that an automated measurement procedure is possible without compromising measurement accuracy. This is validated through measurements where no difference was found between measurement results obtained with the automated procedure and those obtained with a human supervisor. The locally running ASR system achieves a good performance overall, with only 0.6 % deletion errors and 3.0 % insertion errors on the scoring of the listener's responses. Furthermore, the system is robust with respect to the tested speech disorders and noisy recordings. The evaluation of the screening application for the smart speaker showed higher error rates from the commercially available system (6.1 % deletion and 1.9 % insertion errors on the scoring), and varying biases were observed for different subject groups, which cannot be easily compensated for. However, when making a binary decision (*Is the hearing-in-noise capability reduced or not*), a high precision accuracy (area under the receiver operating characteristic curve of 0.95) is observed. Hence, the system developed in this thesis should provide an accessible and easy-to-use screening procedure with a good reliability.

Kurzfassung

Sprachliche Kommunikation ist eines der Kernelemente sozialer Interaktion. Doch gerade das Hören und Verstehen von Sprache, insbesondere in lauter Umgebung, ist für Hörgeschädigte oft deutlich eingeschränkt. Mit weltweit 1,5 Milliarden betroffenen Menschen ist dies ein gravierendes globales Problem. Um Betroffenen zu helfen, braucht es zunächst Diagnoseinstrumente, wie die Sprachaudiometrie, welche zunächst erfasst, inwiefern die Fähigkeit Sprache im Störgeräusch zu verstehen, eingeschränkt ist. In der Sprachaudiometrie gibt es hierbei zwar mehrere gut etablierte klinische Messverfahren, welche jedoch meist mit einem hohen Messaufwands verbunden sind, da ein Spezialist den sprachaudiometrischen Test durchführen muss. Dieser hohe Messaufwand kann ein limitierender Faktor für die klinische Durchführung solcher Tests und deren allgemeine Zugänglichkeit sein. Eine Selbstmessanwendung, welche automatische Spracherkennung (ASR) nutzt, kann diese Einschränkung von sprachaudiometrischen Tests überwinden und sie sogar auf Screening Anwendungen ausweiten. Allerdings produzieren selbst die besten ASR-Systeme einige Fehler bei der Transkription der Antworten der Probanden, was die Messgenauigkeit der sprach- audiometrischen Selbstmessung potenziell verringert.
In der vorliegenden Arbeit wird der deutsche Matrixsatztest als Entwicklungsprototyp für ASR basierte, unüberwacht durchgeführte Messungen genutzt. Es wird unter anderem der Einfluss von ASR-Fehlern auf die Messgenauigkeit mit Hilfe von Monte-Carlo-Simulationsmethoden untersucht und es werden Selbstmessverfahren mit ASR für zwei verschiedene Anwendungsszenarien vorgeschlagen. Im ersten Anwendungsszenario wird in einer gut kontrollierten Laborumgebung mit einem lokalen ASR-System, welches speziell für diesen Zweck entwickelt wurde, gemessen. Das zweite Szenario ist eine Screening-Anwendung, bei der die ASR-Komponente

eines Smart Speakers verwendet wird, also eines hochwertigen, kommerziell erhältlichen Lautsprechers, der mit einem virtuellen Assistenten verbunden ist. In diesem Anwendungsszenario gibt es eine geringere Kontrolle über die experimentellen Parameter. Die beiden oben vorgeschlagenen Systeme werden mit insgesamt 139 Probanden evaluiert, welche ein breites Spektrum an Hörfähigkeiten abdecken: Normalhörende, leicht-, mittel- und hochgradig hörgeschädigte Personen sowie Personen mit Cochlea-Implantaten. Für das Anwendungsszenario in kontrollierter Laborumgebung wird einerseits mit unversorgten Messungen mit Kopfhörerpräsentation der Stimuli gemessen, was zu störgeräuschfreien Audioaufnahmen führt und andererseits mit versorgten Messungen mit Lautsprecherpräsentation, was zu verrauschten Aufnahmen führt. Bei dem zweiten Anwendungsszenario, die Screening Anwendung, werden drei verschiedene Nachhallszenarien betrachtet, da die akustischen Bedingungen bei der Selbstmessung zu Hause unklar sind.
Die Monte-Carlo-Simulationsmethoden, welche das adaptive Messverfahren nachbilden, zeigen, dass das adaptive Verfahren des Matrix-Satz-Tests generell sehr robust gegenüber Fehlern des ASR-Systems ist, was wiederum zeigt, dass ein automatisiertes Messverfahren möglich ist, ohne die Messgenauigkeit zu beeinträchtigen. Die Simulationsergebnisse werden durch Messungen validiert, bei denen kein Unterschied zwischen den erzielten Messergebnissen des automatisierten Verfahrens und des Verfahrens mit einem menschlichen Versuchsleiter, festgestellt wurde. Das lokal laufende ASR-System erreicht insgesamt eine gute Leistung mit nur $0,6\,\%$ Löschfehlern und $3,0\,\%$ Einfügungsfehlern bei der Bewertung der Probandenantworten. Zudem ist das System robust gegenüber den getesteten Sprachstörungen und verrauschten Aufnahmen. Die Evaluation der Screening Anwendung für den Smart Speaker zeigt allerdings höhere Fehlerraten des kommerziell erhältlichen Systems ($6,1\,\%$ Lösch- und $1,9\,\%$ Einfügungsfehler). Die systematische Abweichung zwischen klinischen Messergebnissen und Messergebnissen der Screening Anwendung variiert für verschiedene Probandengruppen und kann daher nicht einfach kompensiert werden. Bei einer binären Entscheidung (ist das Hörvermögen im Störgeräusch reduziert oder nicht) wird jedoch eine hohe Genauigkeit (Fläche unter der receiver operating characteristic Kurve von 0,95) beobachtet. Diese Ergebnisse lassen erkennen, dass das in

dieser Arbeit entwickelte Verfahren ein zugängliches und einfach zu bedienendes Screening-Verfahren mit einer guten Zuverlässigkeit darstellt.

List of Publications

- **Ooster, J., Tuschen, L., Meyer, B.T. (n.d.) "Self-conducted speech audiometry for users of hearing aids and cochlear implants", submitted to Computer, Speech & Language.**

- **Ooster, J., Krüger, M., Bach, J.B., Wagener, K.C., Kollmeier, B., Meyer, B.T. (2020). "Speech audiometry at home: Automated listening tests via smart speakers with normal-hearing and hearing-impaired listeners", Trends in Hearing,** `https://doi.org/10.1177/2331216520970011` **.**

- Ooster, J., Wagener, K. C., Krueger, M., Bach, J.-H., and Meyer, B.T. (2020). "Potential of self-conducted speech audiometry with smart speakers", Proceedings of the ISAAR 2019, Nyborg, Denmark, `https://proceedings.isaar.eu/index.php/isaarproc/article/view/2019-43` .

- Ooster, J., Porysek Moreta, P. N., Bach, J.-H., Holube, I., Meyer, B.T. (2019). "Computer, test my hearing: Accurate speech audiometry with smart speakers", in Proc. Interspeech, `https://doi.org/10.21437/Interspeech.2019-2118` .

- Castro Martínez, A.M., Gerlach, L., Payá-Vayá, G., Hermansky, H., Ooster, J., Meyer, B.T. (2019). "DNN-based performance measures for predicting error rates in automatic speech recognition and optimizing hearing aid parameters", Speech Communication, `https://doi.org/10.1016/j.specom.2018.11.006` .

- Ooster, J. and Meyer, B.T. (2019). "Improving Deep Models of Speech Quality Prediction through Voice Activity Detection and

Entropy-based Measures", in Proc. ICASSP, https://doi.org/10.1109/ICASSP.2019.8682754 .

- Ooster, J., Huber, R., Meyer, B.T. (2018). "Prediction of Perceived Speech Quality Using Deep Machine Listening", Interspeech 2018 https://doi.org/10.21437/Interspeech.2018-1374 .
- Huber, R., Ooster, J., and Meyer, B.T. (2018) “Single-ended speech quality prediction based on automatic speech recognition”, Journal of the Audio Engineering Society 66 (10), pp. 759-769, https://doi.org/10.17743/jaes.2018.0041 .
- **Ooster, J., Huber, R., Kollmeier, B., Meyer, B.T. (2018). "Evaluation of an automated speech-controlled listening test with spontaneous and read responses", Speech Communication,** https://doi.org/10.1016/j.specom.2018.01.005 .
- Meyer, B.T., Kollmeier, B., and Ooster, J. (2015). "Autonomous measurement of speech intelligibility utilizing automatic speech recognition", in Proc. Interspeech, https://www.isca-speech.org/archive/interspeech_2015/i15_2982.html .

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	automatic speech recognition
AUC	area under the curve
CI	cochlea implant
CVC	consonant-vowel-consonant
dB HL	decibel hearing level
DIN	digits-in-noise
DNN	deep neural network
DTW	dynamic time warp
fMLLR	feature-space maximum likelihood linear regression
GMM	gaussian mixture model
GUI	graphical user interface
HI	hearing-impaired
HINT	hearing in noise test
HL	hearing loss
HMM	hidden Markov model
KWS	keyword spotting
LDA	linear discriminant analysis
LIST	Leuven intelligibility sentences test
LVCS	large vocabulary continuous speech

MFCC	Mel-frequency cepstral coefficient
mHealth	mobile health
MLLT	maximum likelihood linear transform
NH	normal-hearing
OLSA	Oldenburger Satztest (english: Oldenburg sentence test)
OOV	out of vocabulary
PTA	pure tone average
RBM	restricted boltzmann machine
SAMT	speech-controlled automated matrix test
SDR	score deletion rate
SIR	score insertion rate
SNR	signal-to-noise ratio
SRT	speech recognition threshold
ROC	receiver operating characteristic
TDNN	time-delay neural network
TDNN-F	factorized time-delay neural network
VAD	voice activity detection
VCV	vowel-consonant-vowel
WER	word error rate
wFST	weighted finite-state transducers

CHAPTER 1

General introduction

Speech communication is one of the key elements of participating in social interactions. While human auditory processing is highly optimized to recognize and process speech, this ability is often reduced for hearing-impaired (HI) listeners. This is a major issue; since 1.5 billion people are affected by some form of hearing loss (HL), the number of which could grow to 2.5 billion by the year 2050 (World Health Organization 2021). In Germany, only 36.9 % of people who self-identify as HI have a hearing aid; and for 30 % of these hearing aid owners, it took three years or more to become aware of their hearing loss before they purchased a hearing aid (ANOVUM 2018). As the numbers are based on a questionnaire, ANOVUM (2018) did not account for people who are not aware of their hearing loss. "Uncorrected hearing loss gives rise to a poorer quality of life, related to isolation, reduced social activity, a feeling of being excluded, and increased symptoms of depression" (Arlinger 2003) and should therefore be addressed as soon as possible. The recent COVID-19 pandemic has caused a significant decrease in the amount of socialization for people with a self-reported HL (Littlejohn et al. 2021), causing this issue to become even more severe.

Speech intelligibility tests are the most direct way to measure and quantify the capability of a listener to process speech. Speech audiometric tests measure the psychometric function of listeners, i.e., the relation of the stimulus level to the intelligibility, with speech-based stimuli presented either with or without additional noise. This relation can be described with a logistic sigmoid function (Wichmann and Hill 2001).

Since the psychometric function is steepest at 50 % intelligibility, this point, referred to as speech recognition threshold (SRT), can be measured very precisely and is therefore often used as a single measure to estimate the speech recognition capability of a listener.
One of the biggest drawbacks of speech audiometric tests is the requirement for a human expert to be present during the measurement to evaluate the listener's performance. This limits the accessibility of speech audiometric test procedures, takes up a lot of resources in clinical diagnostics, and requires that the supervisor speaks the same language as the subject. The use of automatic speech recognition (ASR) however, can overcome these limitations by an automatic evaluation of the listener's performance based on a transcript of the responses during the measurement.

1.1 Speech in noise tests for clinical diagnostics

The need to quantify a listener's hearing loss already emerged 200 years ago with the idea of compensating for a hearing loss (Feldmann 2004). In its early years, the distance at which a listener could hear the ticking of a pocket watch or whispered speech was used as a measure of hearing loss (Schmalz 1846). Helmholtz' finding that vocals are composed of pure tones (von Helmholtz 1863) laid the foundation of pure tone audiometry. As Hahlbrock (1953) quotes Friedrich Bezold:

> "Die Sprache enthält eine so vollkommene Zusammenstellung aller möglichen Lautkomplexe, daß wir sie für unsere Hörproben erfinden müßten, wenn wir sie nicht schon hätten."

> "Speech contains such a complete collection of all possible complex sounds, that we would need to invent it, if we did not already have it."

Pure tone audiometry cannot cover all aspects of human speech processing and direct speech audiometry. Therefore, it seems likely that listening tests using speech signals will remain necessary. To this end, the first approaches worked on standardizing the word material for speech-based listening tests (Hudgins et al. 1947; Hirsh et al. 1952). In Germany, the speech audiometry was founded with the *Freiburger* speech

test (Hahlbrock 1953). While the *Freiburger* test is still in use in clinical practice, it has some severe limitations in its measurement accuracy (Baljić et al. 2016; Winkler and Holube 2016) as well as in its selection of the stimulus words (Steffens 2016). In addition to the need for hearing research, more advanced hearing aids with more options to fit an individual listener require more precise measurements. Today, there is a large variety of different speech intelligibility tests. In order of complexity, they range from: simple logatome tests (Mühler et al. 2009), consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) combinations (A. Paglialonga et al. 2011); to mono- or duo-syllabic rhyme tests (Hahlbrock 1953; Wallenberg and Kollmeier 1989; Kliem and Kollmeier 1994); to more complex methods utilizing full sentences with a fixed structure (Hagerman 1982; Wagener, Kühnel, et al. 1999a; Kollmeier, Warzybok, et al. 2015), as well as lists of daily life sentences with varying structure (Plomp and Mimpen 1979; Kollmeier and Wesselkamp 1997; Hochmair-Desoyer et al. 1997); to complex grammatical sentences (Uslar et al. 2013). As the complexity of the stimulus material increases and more and more cortical elements of the auditory processing are assessed, the subject's tested ability develops through the following stages: simple audibility, reception and recognition, up to comprehension (T. Steffens 2017). Other dimensions of the complexity of speech audiometric tests are spatial scenes and additional noises, which are used to distort the target speech. These elements are utilized to either assess specific elements of the auditory processing, or to increase the ecological validity of the test result, i.e., the correlation with the experience in daily life communication situations. Nevertheless, when exploring methods of automating of speech intelligibility tests, modification of the stimuli only plays a minor role - the listener's response behavior most likely only changes marginally when the target speech gets distorted.

There are several approaches for unsupervised speech intelligibility tests, i.e., tests without the need of a human expert to be present during the measurement:
Brand, Wittkop, et al. (2004) present a graphical user interface (GUI)-based approach for the matrix sentence test, which provides the same test results as the conduction with a human expert. Francart et al.

(2009) propose a system for automatic typo correction of written feedback for the Leuven intelligibility sentences test (LIST) and achieve the same measurement accuracy as with a human supervisor. Nogueira et al. (2010) evaluate a system for an ASR-based unsupervised pronunciation training for cochlea implant (CI) patients. Deprez et al. (2013) firstly proposed to use ASR for an automated scoring in a sentence speech intelligibility test. This system is evaluated with 17 normal-hearing (NH) subjects and achieves a false alarm rate of 9.3 % and a detection rate of 90.7 % on binary sentence scoring. They conclude from simulations that the errors would lead to a bias of 1.2 dB, and that the intra-subject standard deviation would increase from 1.2 dB to 1.8 dB. Venail et al. (2016) use dynamic time warp (DTW) on self-recorded audio to measure the perception of CVC word lists. When validated with 77 NH and 13 mildly to moderately HI listeners, they achieved equivalent results of traditional speech audiometry which utilizes prerecorded lists and manual scoring.
Approaches that use text- or GUI-based interfaces exclude subjects who are visually impaired, motorically restricted, or illiterate. In the year 2018, Grotlüschen, Buddeberg, et al. (2018) found that 12.1% of the German adult population (18 to 64 years) were functionally illiterate, i.e., they cannot read and process a full sentence. 4.0% were illiterate in the strict sense of the word. Therefore, this thesis pursues interfaces that purely use our natural means of communication - speech.

This thesis focuses on sentence-based speech intelligibility tests, since they can cover the full variance of speech which can occur during speech communication in daily life. There are two major approaches to sentence-based speech intelligibility tests: those where the target speech utilizes randomly generated sentences without any semantic meaning, and those where the target speech consists of meaningful sentences. Tests with meaningful target sentences (e.g. Kollmeier and Wesselkamp (1997)) are based on predefined lists and have the advantage that they utilize realistic sentence material. This can increase the ecological validity, as it includes linguistic elements representative of daily life communication. Nevertheless, the varying sentence structure makes it difficult to achieve the same intelligibility between all test lists. Furthermore, the words in

meaningful sentences are not independent of each other. For this reason, usually sentence scoring, as opposed to word scoring, is used in list-based sentence tests. Furthermore, since it is possible to remember these sentences, each measurement list can only be measured once with each subject.
In random sentence tests, the target sentences are based on a random walk through a five-by-ten word matrix (e.g. Wagener, Kühnel, et al. (1999a)). Since the words are not connected by their meaning, word scoring can be used for random sentences. Therefore, multiple data points are assessed in each presented sentence. Furthermore, since it is possible to fine-tune single words during the construction of the test (Kollmeier, Warzybok, et al. 2015), the intelligibility of the test lists can be better balanced. The highly-balanced stimulus material together with the increased number of test tokens, result in a higher measurement accuracy for the random sentence test in comparison to semantically meaningful tests. One drawback of the matrix-based target material test is that a significant training effect occurs during the first two measurements due to the subjects learning the words.
The ASR-based speech interfaces developed in this thesis also target screening applications. For screening purposes, meaningful sentences of tests that are used in clinical practice cannot be used. Otherwise, subjects could remember the sentences in which case the test would not be applicable in a clinical context anymore. The German matrix sentence test - the Oldenburger Satztest (english: Oldenburg sentence test) (OLSA) - is therefore selected as a development prototype for unsupervised speech intelligibility testing.

1.2 Speech tests for hearing screening

An early diagnosis of hearing loss and treatment is important to lower the burden of the disability (Arlinger 2003). In Germany, 55% of the people surveyed did not have a hearing test for at least 5 years - 31% never went for a hearing test at all (ANOVUM 2018). Most of the hearing tests are done by a specialist. Only 2% did an online/smartphone test (ANOVUM 2018). Online hearing screening tests, due to their easy accessibility, could reduce the rate of untested people. In turn, this could

reduce the rate of uncorrected hearing losses by increasing awareness and providing guidance to the users on whether or not to seek advice from a specialist.

In the simplest version of the speech-in-noise tests, the target speech and the distorting noise are presented over the same channel. Since the target speech and the distorting noise are affected by channel distortions in the same way, the signal-to-noise ratio (SNR) of the signal and thus the measurand, i.e. the intelligibility corresponding to the SNR, does not require a perfectly calibrated absolute presentation level nor a perfectly flat equalizer. This makes speech-in-noise tests good candidates for hearing screening, resulting in the development of the digits-in-noise (DIN) (Melanie A. Zokoll et al. 2012; Smits, Theo Goverts, et al. 2013; De Sousa et al. 2020). While the limitation to digits allows for a simple user interface for self-measurements, that also works with landline telephones, this restriction also limits the phonological variability of the target material. More complex target material is not suitable for self-conducted screening tests with such a simple user interface. The use of ASR can overcome this limitation as it does not require one to visually present all response possibilities to the user, which is not feasible on a smartphone screen. Additionally, the recent popularity of smart speakers, i.e., loudspeakers connected to a virtual assistant, provide a loudspeaker with good audio quality, as well as an ASR component that is distributed to end users. Therefore, smart speakers are prime candidates to provide accurate hearing screening with sentence-based listening tests at home.

1.3 Automatic speech recognition for hearing tests

Recognizing 50 words in a clean acoustic environment is generally a solved task for automatic speech recognition as state-of-the-art systems achieve parity with humans in transcribing large vocabulary continuous speech (LVCS) (Xiong et al. 2017). A lot of recent high-performance ASR systems are end-to-end approaches (Wang et al. 2019). These models learn the full representation from the features to a grapheme or word level with recurrent neural network structures such as long short-term memory networks and learn long dependencies in sequential data. Hence, end-to-end models do not require a separate language model, whereas the more

traditional deep neural network (DNN)-hidden Markov model (HMM) hybrid models do (Hinton et al. 2012). Therefore, the recurrent neural networks for ASR require training data that matches the target data up to the level of the sentences. Furthermore, these recurrent structures perform best when trained on at least several hundreds hours of training data (e.g. (Parthasarathi and Strom 2019)). To handle long-term dependencies of the speech, the hybrid models use a HMM structure to incorporate phonetic knowledge in the form of a word lexicon with phoneme transcriptions, as well as grammatical knowledge in the form of an N-gram graph trained separately on large text corpora. In such a separate language model, it is possible to manually incorporate explicit knowledge about the target data without the need for a training database. A hybrid approach is more applicable for automating speech intelligibility tests, since a large training data corpus with fully realistic spontaneous responses during a speech intelligibility test is not available. Nevertheless, the task of automating a speech intelligibility test differs from traditional LVCS ASR, since the full transcript is not required during the measurement. For correct scoring, only a recognition of the potential words from the target sentence in the subject's response is needed. The task also differs from a traditional keyword spotting (KWS) approach as these approaches are usually built for unknown keywords (e.g. (Chen, Khudanpur, et al. 2013)). Since the target speech is known when conducting speech audiometry, it is possible to highly optimize a state-of-the-art ASR system for the use in speech audiometric tests and to achieve error rates even lower than the benchmarks for conversational speech. The biggest challenge is to find the correct balance of optimization, otherwise an over-tailored system could result in a lot of errors.
Regardless of the quality of the ASR system, such an automatic recognizer will always make some errors. One of the main objectives of this thesis is therefore to investigate the response behavior of listeners in such speech audiometric tests and to evaluate how errors influence the measurement accuracy.

1.4 Structure of this thesis

Summarizing the points addressed above, the aim of this thesis is to propose a speech-based interface for speech audiometric tests to increase the accessibility and availability of speech audiometric tests for hearing screening as well as for clinical diagnostics. For this purpose, the matrix sentence test is chosen as a development prototype and the interaction of errors from the ASR system on SRT measurement accuracy is investigated.

Chapter 2 investigates the general feasibility of performing the matrix sentence test automatically and defines the evaluation metrics for the following chapters. A prototype of the unsupervised measurement procedure is developed, which included the recordings of a training database for the ASR system. The prototype is used to record a first database of realistic responses during the matrix sentence test with NH and mildly HI subjects. Furthermore, the general relation of the errors from the ASR system is investigated with Monte Carlo simulations.

Chapter 3 describes the extension of the evaluation database to a broader range of HI subjects, which includes aided measurements using a loudspeaker setup and measurements with subjects supplied with a CI. The use of loudspeakers for the stimulus presentation results in noisy recordings of the subjects' responses. To handle these more challenging acoustic conditions and the potentially worse pronunciation of severely HI or deaf subjects, a more sophisticated ASR system is developed and compared with the model of Chapter 2.

Chapter 4 describes the development and evaluation of a so-called "skill" of the matrix sentence test for a smart speaker, i.e., a loudspeaker with a far-field ASR component connected to a virtual assistant. Such an application can dramatically increase the accessibility of speech-in-noise tests but comes with the drawback of several uncontrolled factors that can potentially decrease the measurement accuracy.

Chapter 5 concludes this thesis by summarizing and discussing the overarching results and findings. Furthermore, the work is placed in the

context of recent trends in digitizing of healthcare and suggestions for future research are developed.

CHAPTER 2

Evaluation of an automated speech-controlled listening test with spontaneous and read responses

Abstract

A method for an automated system for speech audiometry is introduced and evaluated using pre-recorded responses as well as spontaneous utterances produced by listeners during a real measurement. A hearing test is performed under the use of automatic speech recognition (ASR) based on the matrix sentence test, which is used clinically for diagnostics and fitting of hearing devices as well as in psychoacoustic research. The test measures the speech reception threshold (SRT), i.e., the signal-to-noise ratio at which the subject achieves 50% word recognition rate. A major disadvantage of current testing procedures is the requirement of a hu-

This chapter is a formatted reprint of
***Jasper Ooster**, Rainer Huber, Birger Kollmeier, and Bernd T. Meyer,*
"Evaluation of an automated speech-controlled listening test with spontaneous and read responses,"
Speech Communication, vol. 98, pp. 85–94, Apr. 2018, `https://doi.org/10.1016/j.specom.2018.01.005` .
Author contributions: JO developed and implemented the prototype of the unsupervised measurement system, developed the simulation methods, performed the measurements, prepared the figures. All authors co-wrote the manuscript (with the main contribution coming from JO).

man expert supervising the test and logging the listener's responses. An automated system reduces the required resources and therefore provides a tool for frequent assessment of the SRT, which can contribute to an early diagnosis of hearing loss. The accuracy of the ASR-based SRT measurement is compared to results obtained with a human supervisor. To this end, two databases are used that contain either well-controlled read utterances that resemble typical responses during SRT measurements produced by 17 speakers, or spontaneous responses collected during real SRT measurements using ASR. Twenty normal-hearing and seven slightly to moderate hearing-impaired subjects participated in the collection of this spontaneous speech. In order to assess the SRT accuracy for read speech, two simulation schemes are proposed that employ Monte Carlo tests to simulate a listener's profile and corresponding responses, which are validated with the real measurement data. We show that ASR deletion rates of 0.9% and insertion rates of 2.9% for matrix text words are sufficiently low to obtain accurate SRT measurements in the range of 0.5 dB SNR. This is comparable to the test-retest accuracy obtained by human supervisors. While ASR errors are overestimated when using the controlled speech material in comparison to spontaneous speech, this error type has minimal effect on SRT estimation. Hence, the use of prerecorded, read speech material is sufficient when evaluating the accuracy of speech-controlled, automated listening tests.

2.1 Introduction

Speech audiometry is an important tool for diagnosing hearing impairments by determining the individual SRT of listeners, i.e., the SNR at which 50% of words are correctly recognized. Matrix sentence tests as first proposed in Hagerman (1982) are a successful implementation of speech audiometry; for these tests, random sentences are generated by walks through a matrix that contains the test words. Noisy sentences are presented to a listener, who responds with the words he or she recognized. A proposed method for recording and splicing words to assemble natural, random sentences (Wagener, Kühnel, et al. 1999a; Wagener, Kühnel, et al. 1999b; Wagener, Brand, and Kollmeier 1999), resulting in the Oldenburg Sentence Test (German: Oldenburger Satztest, OLSA), features

sentences that cannot be predicted by the listener. The matrix structure of the sentences results in a high measurement accuracy and in comparable measurements across different languages (Melanie A Zokoll et al. 2013), with currently 14 available languages (e.g., English (Hewitt 2008), Dutch (Houben et al. 2014), Italian (Puglisi et al. 2014), and Spanish (Hochmuth et al. 2012)). A review of the international matrix test with a refined recipe for designing a matrix test in a new language is presented in (Kollmeier, Warzybok, et al. 2015). A major disadvantage of current testing procedures is the requirement of a human expert supervising the test and logging the listener's responses. Secondly, it introduces a subjective element to the testing procedure, due to the clinician's potential lack of comprehension for a non-native language or momentarily inattentiveness. In Francart et al. (2009) a typing-based response system for a open-set sentence test is presented, which allows test designs that do not require the presence of a supervisor. Nevertheless, such a system requires proper writing skills and computer experience which cannot be assumed for all patients. Another approach suitable for the closed-set matrix test is the use of a GUI (Brand, Wittkop, et al. 2004). However, this requires the patient to read lists of all possible response alternatives and find the right words in a reasonable time. Even though this GUI is highly optimized during years of clinical usage, for tests with 50 different words such as the matrix test, this is a challenging procedure for regular users and excludes people who cannot operate such an interface. For instance, in Germany 10.0% of the population that is between 18 and 64 years old is functionally illiterate and additional 4.5% illiterate in the strict sense of the word (Grotlüschen and Riekmann 2012). This is in contrast to relatively simple GUIs for digit triplet tests, which can be easily represented and clearly arranged even on smaller displays.

Testing subjects in another language than their native language results in unreliable test outcomes (Warzybok et al. 2015). In those cases for which a visual response system (GUI/writing) is required, or supervisor and test subject speak different languages, a different solution is required. Hence, we propose a system that relies on our natural means of communication, i.e., spoken language by using ASR. Such a system enables an easy to use telephone- or app-based self-testing and can potentially be applied for open-set tests as well. While it can be assumed that an audiometrist

produces almost error-free response logs in a regular testing setup, the same is not guaranteed for ASR, despite the fact that matrix tests exhibit a small vocabulary (50 words in this case). While error rates are of lesser interest in the context of SRT testing, the effect of ASR errors on the SRT and the accuracy of the SRT are key measures associated with a system. In a related study (Deprez et al. 2013), this accuracy was investigated for an ASR-based listening test that was targeted at users of CIs, which was based on the Dutch LIST test (Van Wieringen and Wouters 2008). In this case, a test-retest reliability of 1.8 dB was obtained, which is not sufficient for SRT measurements when the goal is an accuracy that is on par with clinical measurements. To accurately differentiate between normal-hearing and hearing-impaired listeners, the test-retest reliability should be in the same range as for human test conductors of 0.5 dB (Brand and Kollmeier 2002) for this clinical application setting.

In own previous work (Meyer et al. 2015), a speech-controlled SRT measurement was proposed based on the OLSA matrix test described above. For that study, speech was collected from 20 speakers who read responses that are typical for matrix tests to establish a well-controlled data collection, which was used for training an ASR system. For speech without OOV words, ASR error rates were sufficiently low to obtain an unsupervised measurement accuracy of 0.5 dB. However, for the most difficult scenario (many out of vocabulary (OOV) words), the ASR performance was found to be insufficient (i.e., it resulted in a test-retest accuracy above the accuracy achieved with human supervisors (0.5 dB)). While this previous result for *read* speech is promising, it is unclear if an accurate, automated SRT testing can also be obtained with *spontaneous* utterances produced during a real test. Further, the work presented in Meyer et al. (2015) relied on listener's data from the literature only, while actual listening experiments were not performed. In the present work, we therefore investigate the accuracy of the ASR-based system for speech that was elicited during actual measurements from 20 NH and seven HI subjects. For these experiments, the original matrix test using 50 words is chosen, since it has a high measurement accuracy, the results are comparable between the different international matrix test, and since it is well-established and clinically used on a day-to-day basis. The accuracy

is determined by comparing the SRT_A (which is the SRT determined by our automated system) to the SRT_H, i.e., the SRT determined by a human supervisor. The collection of this dataset in combination with the data collected earlier (Meyer et al. 2015) enables us to address an important question in speech research, i.e., differences encountered between well-controlled speech material collected in the lab (which in this study corresponds to the read utterances described above) and realistic speech data with a higher variance that occurs in real-world scenarios (which in this study is given by the data collected during actual testing). This issue was addressed earlier: For instance, the influence of real and simulated training and test data was investigated in the CHiME 3 challenge (Barker et al. 2015), with the result that artificially added noise is beneficial during training, but results in very different error rates during testing, i.e., results of the simulated test set are not transferable to the realistic counterpart. In the context of reverberation, the REVERB challenge (Kinoshita et al. 2016) investigated the effect of artificially reverberated data (for which pre-recorded impulse responses were exploited) in comparison to data obtained in a real-world setting (Lincoln et al. 2005). Although results with simulated test sets from large rooms are correlated with results from real recordings, important differences between the data sets were observed, i.e., it is still more challenging to cope with real reverberation for speech enhancement and speech recognition when only simulated data is available for fine-tuning the dereverberation algorithms or for ASR training. However, in the context of automated speech audiometry, the SRT accuracy that can be measured or simulated has not been under investigation for well-controlled and real-world datasets. This study therefore addresses the question how the SRT accuracy obtained in realistic measurements relates to SRT accuracy estimated with read speech, and whether it is sufficient to measure the test reliability with data gathered in the lab. To this end, we propose a simulation framework based on Monte Carlo tests for simulating listener properties, corresponding responses, and potential ASR errors. To put the results in a broader context, ASR errors are simulated independently of a specific test-data set, which has implications for matrix tests in other languages. Furthermore, transcripts obtained from ASR experiments are taken into account to analyze which data is required to measure the performance

of the measurement system. The simulation method is validated with the realistic data set and then applied to the well-controlled data sets. The remainder of this paper is structured as follows: In Section 2.2, the data collection and the ASR system used for SRT measurements are introduced. Results for actual measurements and two types of simulations for the accuracy of the proposed SRT measurement are presented in Section 2.3. The discussion and the summary are presented in Sections 2.4 and 2.5, respectively.

2.2 Methods

2.2.1 Matrix sentence test

Speech intelligibility tests are an important and successful approach in audiometry, since the measurement of the intelligibility of words and sentences provides more relevant and crucial information about the everyday communication restrictions caused by a hearing impairment than the audibility threshold of pure tones as measured by the audiogram. The goal of speech-in-noise tests in particular is to estimate the parameters of the individual psychometric function of a subject, i.e., the intelligibility as a function of the SNR of the presented speech signal. This function can be described with a logistic sigmoid function:

$$p(L, SRT, s, p_{max}) = \frac{p_{max}}{1 + \exp(4 \cdot s \cdot (SRT - L))}, \tag{2.1}$$

with the slope s, the speech presentation level L (with a fixed level of background noise) and the maximum performance level p_{max} for high SNRs. During testing, the subject listens to noisy sentences. In a matrix test, all sentences exhibit the same structure (*name verb numeral adjective noun*) with ten alternatives for each position (cf. Table 2.1). Hence, the sentences are syntactically correct, but semantically unpredictable since they are produced from random left-to-right walks through the matrix. The general testing procedure as well as the extension to an ASR-based system (which will be described in the next section) is illustrated in Figure 2.1: After the presentation of a matrix test sentence (Figure 2.1 **A**), the subject responds with the recognized words (Figure 2.1 **B**). During supervised measurements, the audiometrist compares

Name	Verb	Numeral	Adjective	Noun
Peter	**bekommt**	**drei**	**große**	**Blumen**
	gets	*three*	*large*	*flowers*
Kerstin	**sieht**	**neun**	**kleine**	**Tassen**
	sees	*nine*	*small*	*cups*
Tanja	**kauft**	**sieben**	**alte**	**Autos**
	buys	*seven*	*old*	*cars*
Ulrich	**gibt**	**acht**	**nasse**	**Bilder**
	gives	*eight*	*wet*	*pictures*
Britta	**schenkt**	**vier**	**schwere**	**Dosen**
	donates	*four*	*heavy*	*cans*
Wolfgang	**verleiht**	**fünf**	**grüne**	**Sessel**
	loans	*five*	*green*	*armchairs*
Stefan	**hat**	**zwei**	**teure**	**Messer**
	has	*two*	*expensive*	*knives*
Thomas	**gewann**	**achtzehn**	**schöne**	**Schuhe**
	won	*eighteen*	*beautiful*	*shoes*
Doris	**nahm**	**zwölf**	**rote**	**Steine**
	took	*twelve*	*red*	*stones*
Nina	**malt**	**elf**	**weiße**	**Ringe**
	paints	*eleven*	*white*	*rings*

Table 2.1: The matrix of the German OLSA matrix test (from (Wagener, Kühnel, et al. 1999a)). Each sentence is generated by a random walk through the matrix. The underlined words are an example for a randomly chosen sentence: **'Nina bekommt vier rote Blumen'** (*Nina gets four red flowers*).

the subject's response to the presented matrix test sentence and calculates the recognition score N (Figure 2.1 **S**). Based on the recognition score, the change of the SNR ΔL is calculated as follows:

$$\Delta L = -f(i)\frac{(N/5 - target)}{s}, \tag{2.2}$$

as shown in Figure 2.1 **E**. For a recognition rate above the $target$ recognition rate (50%), the SNR is decreased and increased otherwise. s is again the slope of the psychometric function at the SRT. The function f controls the SNR changes which are gradually decreased, depending on the number of reversals i (i.e., continuous segments in which the SNR is always increased *or* decreased):

$$f(i) = \max(1.5 \cdot 1.41^{-i}, 0.25). \tag{2.3}$$

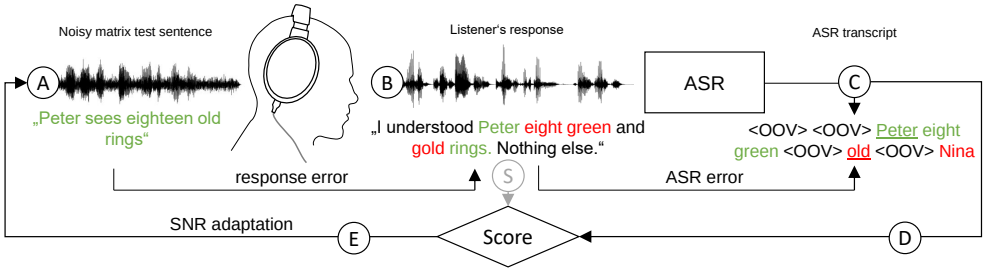


Figure 2.1: Flow chart of the measurement system. The exemplary subject's response and ASR transcript illustrate the potential unsupervised measurement error types. **A:** Presentation of a noisy matrix test sentence. **B:** The subject responds to a noisy matrix test sentence by repeating the recognized words; the corresponding error is the *response error* (red). **C:** Decoding of the recording of the subject's response with an ASR system. This introduces the *ASR error* (red). **D:** Scoring by comparing the ASR transcript with the matrix test sentence (underlined words). **E:** Adaptation of the SNR for the next noisy matrix sentence presentation based on the score. **S:** (Only in supervised measurement) Scoring from a human supervisor by comparing the subject's response with the matrix test sentence (green words).

The factors 1.5 and 1.41 were empirically determined (Brand and Kollmeier 2002), while the threshold of 0.25 ensures a minimal SNR adaptation for late trials in the test. During one measurement, 20 sentences are presented while the SNR is continuously adapted; the decrement of SNR changes result in a convergence of the SNR to the SRT, more specifically the SRT_{50}, since the criterion in this work is a word recognition rate of 50%. Details of the measurement procedure and the empirical values are described in Brand and Kollmeier (2002). To increase the robustness of the SRT measure, a maximum likelihood criterion is used to fit a psychometric function to the data points (Brand 2000) (which takes into account that the listener could have been inattentive during the last trials, which would overestimate the SRT). When ASR-based, automated measurements are performed, the recognition scoring is done by comparing the ASR transcript (obtained from the listener's response) with the matrix test sentence (cf. Section 2.2.4). Hence, the step sequence for each measurement loop in Figure 2.1 is (**A-B-C-D-E**). The resulting SRTs with the automated procedure are denoted as SRT_A and the SRTs obtained by a human supervisor as SRT_H.

The small vocabulary of the matrix test results in a training effect, since listeners learn the matrix words during the first tests, which results in

Name	No. of speakers	No. of utterances	Example sentence
read-gap	20	5998	**Peter acht Ringe**. (Peter eight Rings.)
read-HH	17	1675	Der Name war **Kerstin**. (The name was Kerstin.)
spontaneous	20	2400	**Ulrich** und **Bilder** könnten das gewesen sein. (That could have been **Ulrich** and **pictures**.)
spontaneous-HI	7	840	Ja was der **Peter** macht habe ich auch nur halb verstanden (Well, I only half understood what Peter is doing.)

Table 2.2: Overview of the recorded speech data that is used for testing.

an SRT improvement of 1-2 dB attributed to vocabulary training. To account for this effect, a presentation of two training lists, each with 20 matrix sentences, is recommended, which also is useful for familiarizing the patient with the testing procedure (cf. (Wagener, Brand, and Kollmeier 1999)). Normal-hearing, trained listeners typically achieve a mean SRT of -7.1 dB with a between-subject standard deviation of 1.1 dB for the German matrix test (Wagener, Brand, and Kollmeier 1999). In contrast, the test-to-retest or within-subject standard deviation is 0.5 dB for normal-hearing listeners (Brand and Kollmeier 2002) and 0.9 dB for hearing-impaired listeners (Wagener and Brand 2005). This test-to-retest standard deviation is the most important characteristic to describe the accuracy of the measurement procedure since it is independent of the individual spread of hearing capabilities in the population.

2.2.2 Collection of speech data

2.2.2.1 Pre-recorded controlled speech

The task-specific speech data were collected at our lab and first described in (Meyer et al. 2015). In this study, the data is used to train the acoustic model for ASR, and also to investigate if read speech (which does not originate from an actual measurement) can be exploited to deter-

mine the accuracy of an SRT measurement system through simulations. The corpus contains 27,000 utterances (23 hours) collected from 20 different speakers (10 female, 10 male). The subject's age ranged from 21 to 60 years with an average age of 35.8 years. It covers complete matrix sentences with all five words, incomplete sentences with gaps, as well sentences with OOV words (i.e. words that are not part of the 50 words vocabulary of the matrix test) that resemble interaction between listeners and an human supervisor (human-human interaction). For the material without OOV words, the subjects read randomly generated matrix sentences from a screen. Complete matrix sentences are not sufficient to investigate the accuracy of ASR-based SRT testing (since in every SRT measurement, there should be unrecognized/missing words), but are required for high-SNR responses (when the listener recognizes all words) and are used here to obtain representative acoustic models for ASR, i.e., the complete sentences are later used for training (see below). Incomplete sentences without OOV words were recorded to represent the ideal case when the tested person is aware he or she is interacting with an ASR system that is limited to the matrix test vocabulary. These utterances contain one to five words with the same functional order as the original matrix sentences, and are referred to as ***read-gap***. To simulate human-human interaction, we first analyzed responses from participants directed at a human supervisor during traditional SRT measurements. A total of 1,400 responses from seven subjects who participated in ten measurements on average were collected and transcribed. The original recordings were not used for ASR experiments, since they originated from clinical measurements for which permanent storage is not line with test regulations. 100 of these transcribed sentences were selected to represent a broad range of typical responses with OOV words and read by 17 speakers, which resulted in well-controlled stimuli. The corresponding recordings are referred to as ***read-HH***. The recordings took place in an isolated sound booth using a Neumann KM 184 microphone and a Fireface USB UC soundcard with 44.1 kHz sampling frequency and 32 bit resolution. A simple graphical user interface in Matlab was used to visually present sentences to speakers. The recordings were performed with SoundMexPro, a toolbox for studio quality recordings in Matlab[1].

[1] `www.soundmexpro.de`

Speakers were paid for their participation, and were instructed both verbally and in writing. On average, we collected 2200 recordings for each word in the matrix test.

2.2.2.2 Spontaneous responses during automated speech audiometry

The first goal of the recordings described in the following is to create a database of responses from subjects under realistic conditions. A second goal is to analyze differences in accuracy of the unsupervised measurement in comparison to the measurement with a human supervisor. A prototype of the speech-controlled system was used for recording. Measurements took place in an isolated sound booth with a high-quality microphone (Neumann KM 184) at a distance of approximately 0.5 m. For the stimulus presentation calibrated, Sennheiser HDA 200 headphones were used. The measurements were conducted with a continuous noise presentation at 65 dB sound pressure level. This represents a typical environment for listening experiments conducted by a professional audiologist. Since users interact with an ASR system (in contrast to *reading* responses that were previously observed), the users' behavior should closely resemble later application scenarios. Subjects completed the tests without supervision, but were able to use a talk-back system to a human supervisor outside the booth in case they require help. They were instructed in writing with the standard manual of the matrix sentence test (HörTech 2019) with added instructions for the speech-controlled interface (i.e., the information of interacting with an automated system and the fact they could use control commands such as *'start measurement','pause','resume'*). Since these commands were not implemented in the measurement prototype, a human supervisor outside the booth executed the commands without the subject's knowledge in a 'Wizard of Oz' setup (Fraser and Gilbert 1991). The rest of the measurement was completely automated. The ASR transcript of the response was used for scoring (cf. Section 2.2.4) and for calculating the SNR of the next listening item. The supervisor outside the booth also logged the subjects answers, i.e., a raw version of the labels was created online, which was later checked for errors to obtain accurate transcriptions of the responses for evaluating the ASR performance.

Following this data collection protocol, SRT measurements of 20 subjects were performed, which results in an evaluation database of 2400 user responses with spontaneous speech. In total, 154.3 minutes of speech were collected, resulting in an average duration of (3.83 ± 0.95)s per utterance. Subjects were compensated for participation. All participants (except for one) were untrained listeners who participated in the matrix sentence test for the first time. The subject's age ranged from 17 to 50 years with an average age of 28.5 years. In order to have a reference value for the hearing status, an air conduction tone audiogram for each subject was measured. According to the WHO grading (Mathers et al. 2000), the subjects were all normal-hearing listeners. With the same protocol, 7 hearing-impaired subjects were measured. The subject's age ranged from 53 to 77 years with an average age of 68.3 years. This resulted in a database of 840 user responses with an average duration of (2.32 ± 0.68)s per utterance or a total duration of 52.3 minutes. The hearing impaired subjects had a pure tone average (PTA) HL (at [0.5 1 2 4]kHz with air conduction) from 26.3 dB HL to 42.5 dB HL with an average of 35.2 dB HL.

2.2.3 ASR system

The data described above were used to train a standard DNN - HMM system following the procedure for the DNN-baseline described in (Vesely et al. 2013). However, discriminant training was not applied in our study since it resulted in a negligible performance increase and considerable increase of the computational cost. The specific implementation (or recipe) as outlined below is part of the open source ASR software Kaldi (Povey, Ghoshal, et al. 2011) and uses the *nnet* configuration. The input features to this system areMel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein 1980) with 13 components per frame, to which the first and second numerical derivatives (delta and double delta features) are appended. The MFCC features are adapted to each speaker with a feature-space maximum likelihood linear regression (fMLLR) (Gales 1998) on top of a linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) (Gopinath 1998). This results in 40-dimensional features that are spliced with ± 5 frames and used as input for the fully-connected DNN, resulting in an input dimension of 440 (cf. Type-III

features in (Rath et al. 2013)). The DNN had three hidden layers, each with 1024 neurons, a sigmoid nonlinearity and a softmax function applied to the output layer. It was initialized with a layer-wise restricted boltzmann machine (RBM) pre-training (Mohamed et al. 2012) and fine-tuned with the cross entropy between output and the alignments for 1977 triphones from a gaussian mixture model (GMM) - HMM system, which used the same fMLLR speaker adapted MFCC features. During training, the full lexicon with all phonetic transcriptions was provided to train the best possible acoustical models. To handle OOV words during testing, a phone-level 4-gram language model was trained using the lexicon of the general corpora. This model handles phoneme sequences that do not fit the matrix test vocabulary. The keyword spotting is performed on grammar level of the decoding graph with a 0-gram of the 50 keywords from the matrix test and the OOV-phone-language-model. Since it was not possible to make a reasonable assumption for the ratio of vocabulary and OOV words, the same insertion cost of $-\log(1/51)$ was used for all 51 entries in the grammar. The language model weight was set to 20 with an acoustic weight of 0.1. The insertion penalty was not found to be a very sensitive parameter and was set to 3.0 for all experiments. In order to provide a sufficient amount of training material for OOV words, two out-of-domain, commercial German corpora (King-ASR-L-092 and King-ASR-L-182, see `kingline.speechocean.com`) with 16,000 utterances (18 hours) from 40 speakers were additionally used for training. Far-field microphone data also contained in the corpora were excluded from training to resemble the testing scenario.
To test data from one of the speakers from the task-specific training-corpus, a leave-one-out procedure was used, i.e., an ASR system was trained with the speech material from the remaining 19 speakers and the 40 speakers from the out-of-domain corpus.

2.2.4 Scoring of the subject responses

One straight-forward way to score the subject's response is to cover the complete vocabulary that is likely to be encountered in a real test. However, since the response format in a natural interaction is very open with responses such as *'I've understood Peter got four ... oh, no, the name was Thomas.'*, this implies that (a) the vocabulary size would be con-

siderably larger than for the small 50-word corpus, and (b) the decoding complexity would heavily increase. We circumvent these problems by ignoring all words in the response that are not part of the matrix test (and which therefore become OOV words), and decoding matrix words only. An additional advantage of this approach is that a confusion of a matrix word with a command word is not possible, which makes the scoring of the response (which control the SNR estimation of the testing procedure) very straight-forward.

2.2.5 Evaluation metrics

ASR studies typically employ the word error rate (WER) as an evaluation metric. The WER is of limited importance in this study, since the main goal is to obtain an accurate SRT estimate (i.e., to correctly score the subject's response). This accuracy is not influenced by insertion and deletion of OOV words, since these are discarded during the scoring procedure (cf. Figure 2.1). Note that - since the system effectively operates as a keyword spotting algorithm for matrix sentence words in arbitrary order - the error rates reported here are limited to deletion and insertion errors. This leads to the following main ASR evaluations metrics based

stimulus	response	transcript	
✓	✓	-	response score
✓	X	-	response deletion
X	✓	-	response insertion
-	X	✓	ASR insertion
-	✓	X	ASR deletion
✓	X	✓	score insertion
✓	✓	X	score deletion

Table 2.3: An overview of the potential error types of the measurement. Important factors are the presence (✓) or absence (X) of a word in the **stimulus** sentence, the **response** of the subject and the ASR **transcript** of this response. The dash (-) denotes that for this error type it is not relevant if the word is present or not.

on the error types defined in Table 2.3 and illustrated in Figure 2.1. The score insertion rate (***SIR***) is defined as

$$SIR = \frac{N_{\text{score insertions}}}{N_{\text{response score}}}, \tag{2.4}$$

and the score deletion rate **(*SDR*)** is given by

$$SDR = \frac{N_{\text{score deletions}}}{N_{\text{response score}}}. \tag{2.5}$$

In contrast, the ASR deletion error rate **(*DR*)**, the ASR insertion error rate **(*IR*)** and the OOV word rate are normalized with the total number of keywords. All other error types (e.g., when the subjects erroneously inserts a keyword in his or her response) are not considered since they do not influence the measurement.

2.2.6 Simulation methods for accessing the accuracy of automated SRT measurements

The accuracy that can be expected from unsupervised SRT measurements with an ASR system is simulated based on two simulation schemes for two reasons: First, it is used to determine the theoretical effect of ASR insertion and deletion errors on SRT measurements. Secondly, if the simulation is found to be accurate, it could replace an evaluation with data collected during actual measurements (which require more resources than the controlled, read material from other corpora, and which would be an important finding for unsupervised SRT tests in other languages). An illustration of the steps involved is shown in Figure 2.2. The two paths (I and II) through the flow chart symbolize two types of simulations that serve two different goals: (I) To cover a wide range of ASR insertion and deletion rates and evaluate their influence on the SRT, independently of specific speech material (based on (Brand and Kollmeier 2002)), and (II) to obtain an SRT scoring error for pre-recorded utterances, which should answer the question if read speech is sufficient for a system evaluation, how results deviate from real data, and how robust our specific ASR system generally is when encountering read speech. Both simulations are based on Monte Carlo tests.

2.2.6.1 Simulating the general effect of score deletion and score insertion errors on SRT prediction

To simulate an SRT measurement, we first determine a listener's profile by drawing random parameters for the psychometric function (Equation 2.1). The required parameters are sampled from the normal pa-

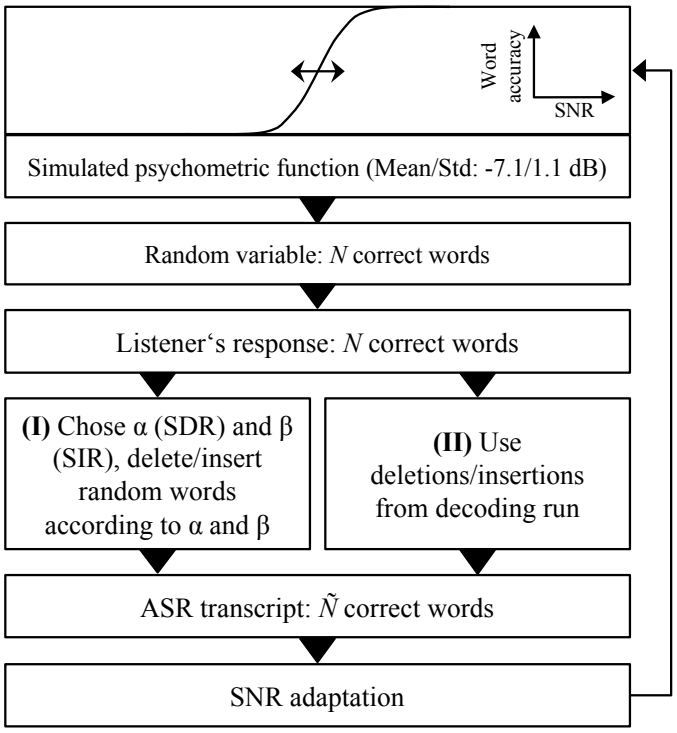


Figure 2.2: Illustration for simulating the accuracy of SRT predictions.

rameter distributions of normal-hearing listeners (Wagener, Brand, and Kollmeier 1999) (p_{max} = 1, SRT: mean: -7.1 dB, standard deviation: 1.1 dB; s: mean: 0.171 dB^{-1}, standard deviation: 0.016 dB^{-1}). The standard value of 0 dB is chosen as initial presentation SNR; it is constantly adapted during the simulated measurement. For each simulated measurement with 20 simulated sentences and responses, a psychometric function is defined by drawing from a normal distribution that corresponds to the standard normal-hearing subject. To assess the influence of ASR errors, a specific score deletion and score insertion rate was chosen for each simulated measurement. 2000 measurements were simulated for each combination of score insertion and deletion error rate. Error rates were chosen from 0.0% to 10% (for score deletions) and to 20% (for score insertions) with a step size of 0.25%. Since it is not possible to assume a direct relation between ASR errors and score errors, this simulation directly simulates the score related errors. According to these rates, words from the listener's response with N correct words are deleted, or additional words are added. Based on this estimated number of correct words $\tilde{N}$, the SNR is adapted according to Equation 2.2. This

process is repeated for the 20 sentences (one complete measurement). The SRT is determined by fitting an (estimated) psychometric function to the measurement values obtained by the described procedure.

2.2.6.2 Simulating the effect of ASR errors for specific data sets

In this section, we will introduce a method to predict the outcome of the SRT measurement with simulated normal-hearing listeners, taking into account the actual insertion and deletion errors from ASR (in contrast to *assuming* pairs of score insertion/deletion errors). To simulate SRTs, we first follow the approach described above: We draw parameters for a psychometric function, which results in N correctly classified words (given the presentation level L). Subsequently, a random, read utterance with N matrix words is first chosen from the specific speech data base and then decoded, and the effect of deletions/insertions for this example is analyzed (Figure 2.2.II). To this end, a random stimulus sentence is selected that matches the recorded utterance. For example, when the subject's response in the utterance was *'**Doris gives** ... and maybe **flowers**'*, random words are chosen for the missing categories *Numeral* and *Adjective*, e.g., *'Doris gives seven small flowers'* in this example. If we now analyze the ASR transcript of this potential stimulus sentence, it is possible to calculate $\tilde{N}$, and to simulate score errors for the read databases. After 20 repetitions, an SRT value is obtained.

2.3 Results

In this section, we first present ASR error rates for the speech corpora introduced in Section 2.2.2, since they are likely to influence an automated SRT measurement. Subsequently, a comparison of ASR-based SRT measurements and tests conducted by a supervisor are presented. The *simulation* of SRT accuracy (which is first validated with the results from real measurements) is presented to quantify the general effect of insertion and deletion errors on SRT measurements, and to explore if the controlled speech material is suitable for estimating the accuracy that can be expected from the ASR-based measurements. Finally, the effects of training, speakers' age, and rate of OOV words are shown.
The evaluation focuses on the normal-hearing subjects, since this data

test set	*OOV rate*	*DR*	*SDR*	*IR*	*SIR*
	(A) training with task-specific and out-of-domain data				
read - gap	0.0	0.69 ± 0.66		1.76 ± 0.95	
read - HH	28.0	3.7 ± 1.5		32.5 ± 5.2	
spontaneous	10.6 ± 6.6	0.95 ± 0.89	0.85 ± 0.73	12.6 ± 5.8	2.9 ± 1.5
spontaneous-HI	16.4 ± 11.1	0.74 ± 0.41	0.60 ± 0.47	17.2 ± 7.9	3.2 ± 1.7
	(B) training only with out-of-domain data				
spontaneous		1.46 ± 0.90	1.34 ± 0.81	7.5 ± 3.1	2.1 ± 1.2
spontaneous-HI		0.73 ± 0.51	0.60 ± 0.41	7.1 ± 4.8	2.1 ± 1.5

Table 2.4: Performance of the ASR system in percent and the between-subject standard deviation. Training was performed either with task-specific data and out-of-domain data (A) or only out-of-domain (B). The evaluation metrics are defined in Section 2.2.5.

can be compared with the prerecorded read speech and is analyzed in detail in the subsequent simulation part. Data from hearing-impaired subjects is used to explore if accurate results can be obtained from this patient group as well.

2.3.1 ASR performance for controlled and spontaneous utterances

The ASR system performance for the different data sets is shown in Table 2.4. For the ideal data *read-gap* (read speech, no OOV words), the system reaches a very high accuracy with a nearly perfect deletion rate and a small insertion rate. For the controlled speech with a high OOV percentage, the insertion rate is very high, because many OOV words are incorrectly classified as keywords. However, for this database, the high amount of OOV words also interferes with the recognition of the presented keywords, resulting in a higher deletion rate.

For spontaneous speech the OOV rate is not controlled and - depending on the speaker - ranges from 2.6% to 25.5% for normal-hearing subjects and from 8.3% to 37.9% for hearing-impaired subjects. Detailed ASR results for this data set can be found in Figure 2.3. The ASR insertion rate is similar or higher than the OOV word rate for all subjects, which shows that OOV words tend to result in insertion errors. Nevertheless, just a small fraction of these insertion errors result in a score insertion error. The deletion error rate is constantly low across all subjects. Response

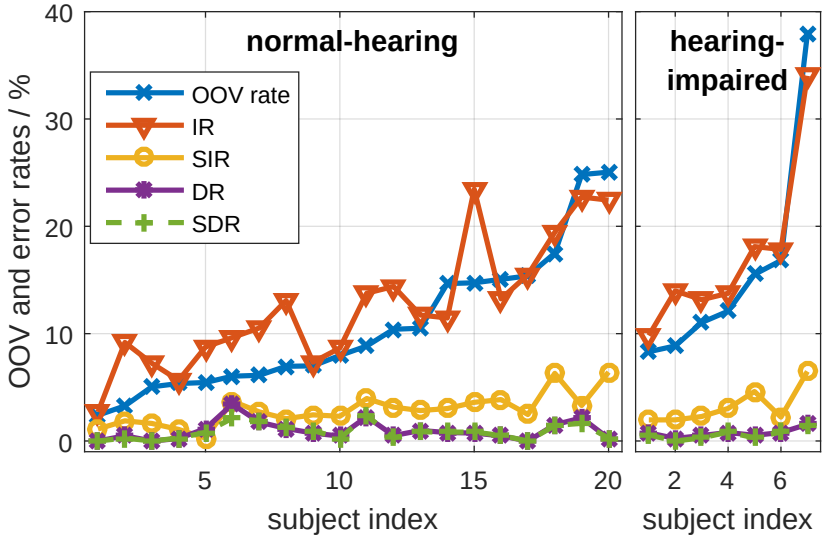


Figure 2.3: The performance of the ASR system for spontaneous speech during evaluation measurements. The OOV word rate, insertion rate (IR), deletion rate (DR), score insertion rate (SIR), and score deletion rate (SDR) as introduced in Section 2.2.5. Results show mean values over six unsupervised measurements.

insertions from the subjects are rare, i.e., if the ASR system misses a keyword, this will also lead to a score deletion error in most cases. When a subject utter a word partially a human supervisor scores this as a response deletion. For the ASR system this is not always possible and partially uttered words are often decoded as the whole word (since the language model does not cover these partially uttered words). Such errors are likely to result in score insertion errors.

Table 2.4 also shows the performance of the ASR system when it is trained with out-of-domain data (18 h of speech data that does not cover matrix sentences collected from 40 speakers). While deletion and insertion rates are different when comparing task-specific and out-of-domain training (with the tendency of lower deletion rates for task-specific, and lower insertion rates for out-of-domain training), the more important errors that affect the *scoring* are rather similar: These scoring error however follow the same trend: On average, scoring deletion errors are reduced with the task-specific set (by 0.5% on average), while scoring insertion errors are smaller with the general training data (by 0.9% on

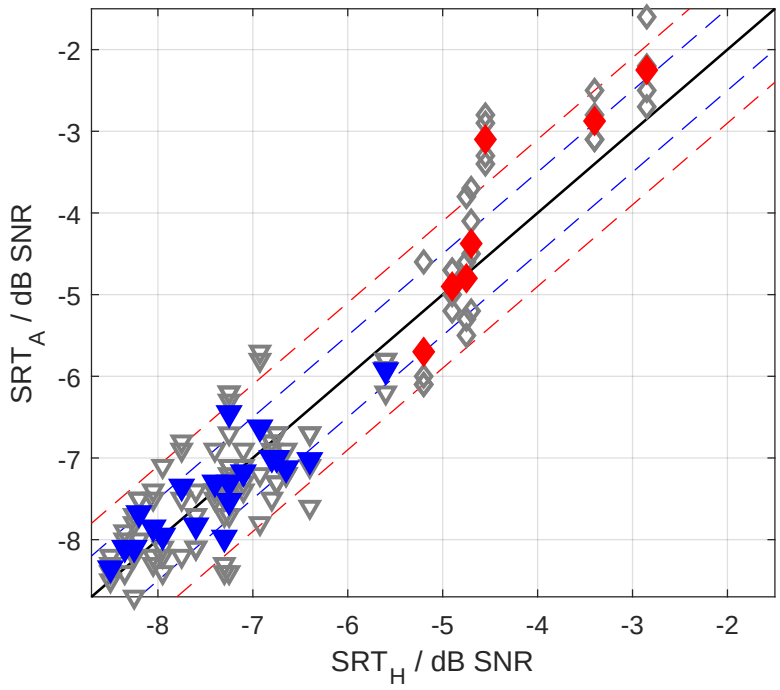


Figure 2.4: Accuracy of the SRT measurement system. The dashed lines indicate the test-to-retest standard deviation of the measurement with a human supervisor. The results from the normal-hearing subjects are represented with the blue filled triangles that represent the mean values estimated from four single measurements, which are indicated by gray, open triangles. Analogously, the results for hearing-impaired subjects are shown with red and gray diamond markers.

average). For both training sets, the ASR system achieves very similar results for hearing-impaired and normal-hearing subjects (Table 2.4 and Figure 2.3), despite the different average age of the two groups.

2.3.2 Accuracy of automated SRT measurement for spontaneous speech (measurement)

Figure 2.4 shows a comparison of individual listeners who were tested with the automated system (SRT_A) and by a human supervisor (SRT_H). The majority of data points from the normal-hearing subjects is within a 0.5 dB interval, which indicates a high accuracy for automated SRT measurements. No significant difference was found between the SRT values measured with the unsupervised and supervised system. Nevertheless, since rejecting significant differences does not imply equality, this is

further analyzed with two one-sided t-tests as proposed in (Schuirmann 1987). This method tests two hypotheses with an upper and lower boundary for the interval of statistical equality. The predetermined values for the boundaries are the test-to-retest standard deviation of the manual, conventional system with $\Theta_{1,2} = \pm 0.5$ dB SNR for normal-hearing subjects. With $\alpha_{1/2} = 0.005$, the two one-sided t-tests show the equality of the unsupervised and the supervised system with $p_1 = 2.8 \cdot 10^{-13}$ and $p_2 = 5.7 \cdot 10^{-14}$ for the normal-hearing subjects. While keeping the p-value below 0.005, the boundaries can be reduced to $\Theta_1 = -0.141$ and $\Theta_2 = 0.161$. The test-to-retest standard deviation of 0.50 dB therefore match the accuracy of a measurement with a human supervisor. Hence, there is no difference in the SRT obtained by an unsupervised or a supervised measurement for normal hearing-subjects that exceeds the test-retest variance. For hearing-impaired subjects, the test-to-retest standard deviation is 0.8 dB and therefore below the reference value of 0.9 dB. The result of the two one-sided t-tests shows that the measurement values are within this 0.9 dB margin (with p values below $2 \cdot 10^{-4}$). The deviation between SRT_A and SRT_H was especially high for one subject with a difference of 1.45 dB. This mismatch is not explainable with the errors introduced by the ASR system. This subject has an SIR of 1.96% and a deletion rate of 0.56% (hearing-impaired subject index 1 in Figure 2.3). When this particular subject is excluded from the analysis, the test-to-retest standard deviation decreases to 0.63 dB.

2.3.3 Accuracy of automated SRT measurement for spontaneous speech (simulation)

Although accuracy was already *measured* for the real data (see Figure 2.4), it is simulated here since a direct measurement of SRT accuracy is not possible for read utterances (for which a simulation is performed later): These utterances were not produced during an adaptive testing procedure, hence we have to make reasonable assumptions about the listener's perception and the corresponding stimuli that resulted in this perception.

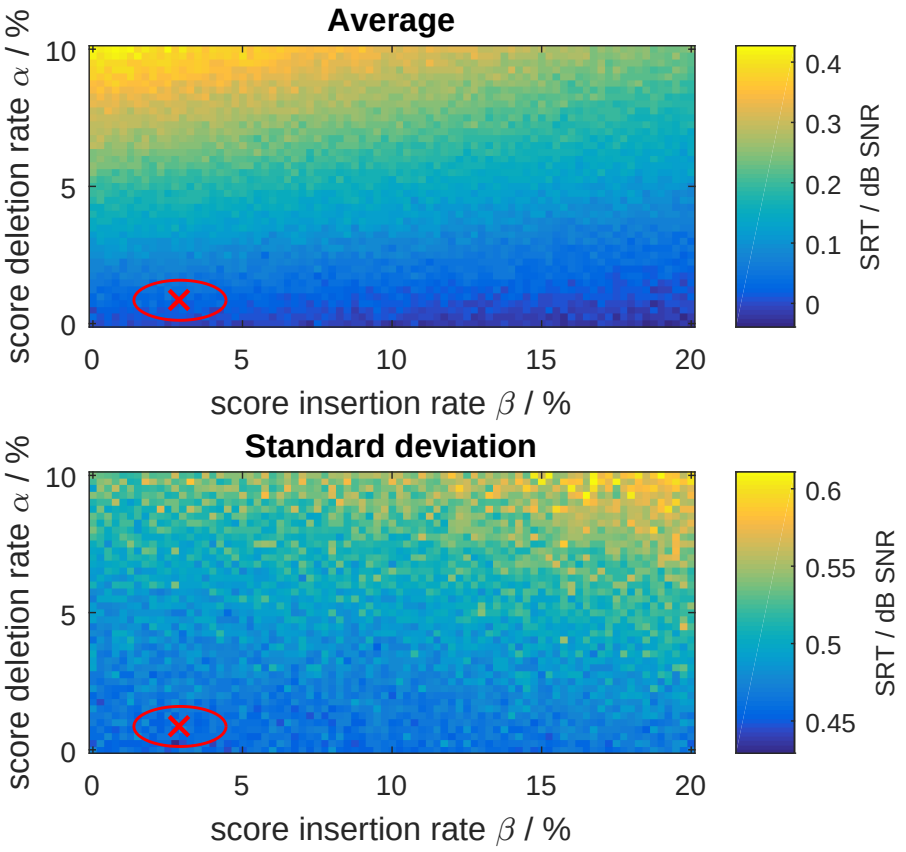


Figure 2.5: The simulation results over 2000 Monte Carlo simulation runs. The red cross indicates the score insertion/deletion rates we obtained with our ASR system. The red oval indicates the standard deviation of these errors.

2.3.3.1 General effect of score deletion and score insertion errors on SRT prediction

Figure 2.5 shows the results of the simulations. When the score deletion rate is increased to 10%, a bias of at least 0.4 dB is introduced. The score insertion rate has a smaller effect: Even high insertion rates of 20% result in a minor negative bias. However, a combination of score insertion and deletion errors with a bias-free measurement can always be obtained since the errors cancel each other out at a specific ratio. The downside of this is shown in Figure 2.5 in the lower panel. For higher error rates, the standard deviation increases, so the test accuracy decreases.
The simulation shows that score deletion errors have a higher impact on the SRT than score insertion errors. This effect originates from the

fact that the common procedure is to start the measurement at an SNR with high intelligibility (which prevents listeners from being frustrated). For normal-hearing subjects, the initial presentation level is 0 dB SNR (HörTech 2019), i.e., the initial presentation level is 7.1 dB above the actual SRT on average. Hence, to decrease the SNR towards the actual SRT, it is more important to not miss any correctly repeated words from the matrix test (i.e., to avoid score deletion errors) than to avoid score insertion errors.

2.3.3.2 Effect of ASR errors for specific data sets

Two of the datasets investigated in this paper were collected by recording controlled speech (sets *read-gap* and *read-HH*). This has the advantage of control over vocabulary and OOV words, but at the same time it prevents a direct assessment of how accurate a measurement is during an actual test. This issue is addressed with this simulation, where the subjects' behavior is simulated similar to the first simulation, but instead actual ASR transcripts are used for the simulation of the influence of the ASR system.

The results of this simulation are shown in Figure 2.6. To investigate if the proposed simulation produces accurate SRT predictions, the results from the ASR-based real measurements with subjects are compared to the simulated results using the identical utterances presented during the measurements. Figure 2.6 shows both curves to be nearly identical. The means of the curves differ only by 0.049 dB. This indicates that the simulation is suitable to estimate the possible SRT accuracy for a given data set. As expected, the curve from the simulation based on the *read-gap* database is shifted to negative SRTs with a bias of -0.10 dB since this database contains no OOV words, resulting in a low insertion rate. On the other hand, *read-HH* exhibits a high OOV rate and therefore high insertion error rate. This curve has an average bias of $+0.12$ dB. All SRTs distributions simulated with the different data sets have a standard deviation in the same range. Hence, the ASR errors on the controlled data sets just have a minor influence on the test accuracy.

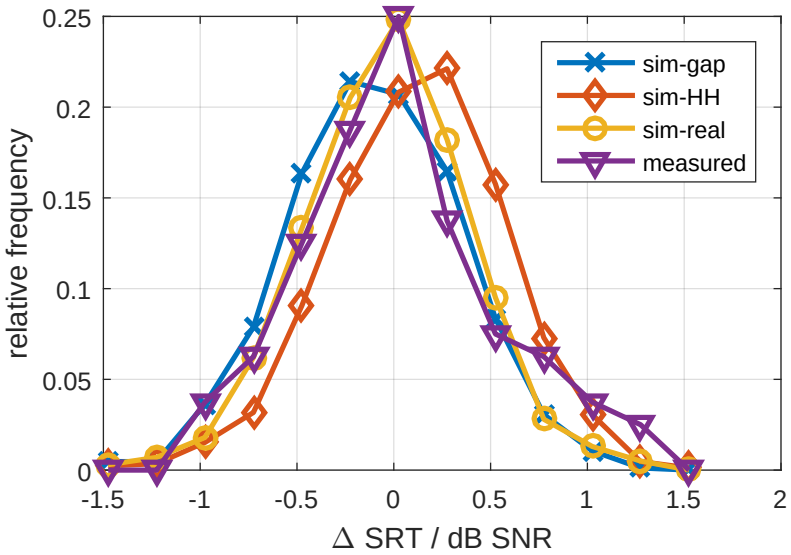


Figure 2.6: Normalized histogram of the simulation results obtained by the procedure described in Figure 2.2.II. *sim-gap* and *sim-HH* are based on the two controlled speech databases. On the other hand, $sim - real$ uses the recorded utterances from the realistic measurements to simulate the ASR errors influence on the SRT. Data referenced as *measured* represents the actual measured results from the realistic test conduction.

2.3.4 Analysis of controlled and spontaneous speech data

This section analyzes data set differences that are likely to have an influence on ASR error rates and hence indirectly on the score error that relates to the predicted accuracy of SRT measurements. Since the number of words per utterance was known, it was possible to determine the average speaking rate in words per second for the different data sets by using a simple energy-based VAD. As shown in Table 2.5, the speaking rate of spontaneous speech is significantly lower than the speaking rate of read speech.

When analyzing the relation of error rates (sum of insertion and deletion errors) and speaking rate for spontaneous speech, a clear trend was not observed. The linear correlation between both factors was not found to be significant ($r = 0.17$, $p = 0.18$). Furthermore there was no observable difference in the speaking rate between the normal-hearing and the hearing-impaired subjects. Hence, the effect of speaking rate was not subject to further analysis. Similarly, a significant relation between

	speaking rate / words/s
Stimulus	2.28
Read - gap	2.19 ± 0.45
Read - HH	2.39 ± 0.38
Spontaneous	1.74 ± 0.18
Spontaneous-HI	1.76 ± 0.27

Table 2.5: Speaking rates for the different datasets with between-subject standard deviations

age and error rate was not found for the 20 normal-hearing and the 7 hearing-impaired subjects. In this case, the linear correlation coefficient was $r = -0.23$ (with $p = 0.26$). Nevertheless, it can be assumed that listeners learn the existing words since the matrix test vocabulary is relatively small (50 words). Hence, the amount of the subjects' training does play a role and it was therefore analyzed if this presumed training effect interacts with the OOV word rate. Figure 2.7 shows the OOV rate in dependency of the measurement number (where each measurement corresponds to the presentation of 20 sentences to determine the SRT). In the first measurement, the OOV rate is very low, since the stimuli are presented at an SNR with high intelligibility (+5 dB SNR), at which errors do not occur. When the adaptive level control is used for the measurement, the subjects misunderstand some words or miss some sentences completely; at the same time, the OOV word rate strongly increases. It decreases again for the following measurements since the subjects adapt to the procedure and learn the vocabulary of the matrix test. In the last measurements the OOV word rate increases for several subjects, presumably due to fatigue effects after approximately one hour of measurement. The amount of OOV words for the controlled speech is fixed, since the words are determined by the predetermined text, and is comparatively high (27.98%).

Figure 2.8 shows the overall statistics from Figure 2.7 in comparison to the statistics in supervised measurements. The data from the supervised measurements are manually corrected transcripts created from conventional matrix test measurements. These data points show that on average the subjects produce a significantly lower number of OOV words during the interaction with an unsupervised computer system (two sample t-test, $\alpha = 0.01$, $p = 0.0026$, equal variances are not assumed).

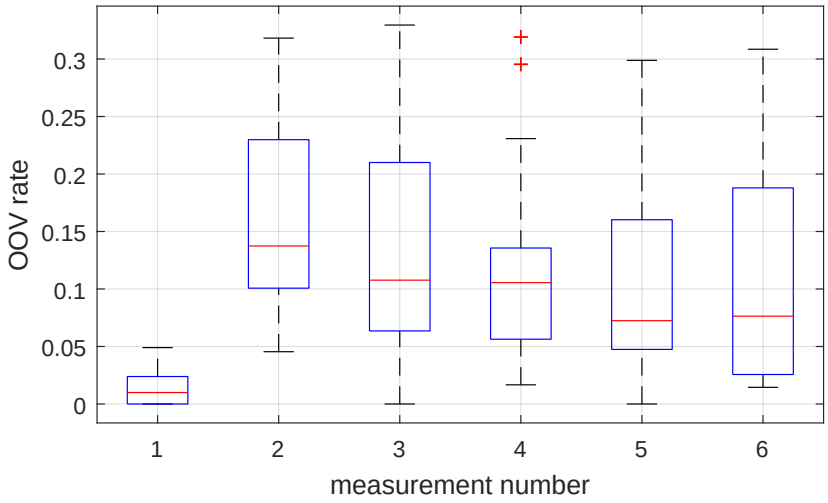


Figure 2.7: Dependency of the rate of OOV-words on the training of the 20 normal-hearing subjects with the test-specific word material. All measurements were performed with an adaptive SNR except the first, which was performed at an SNR with high intelligibility (+5 dB SNR) for training purposes.

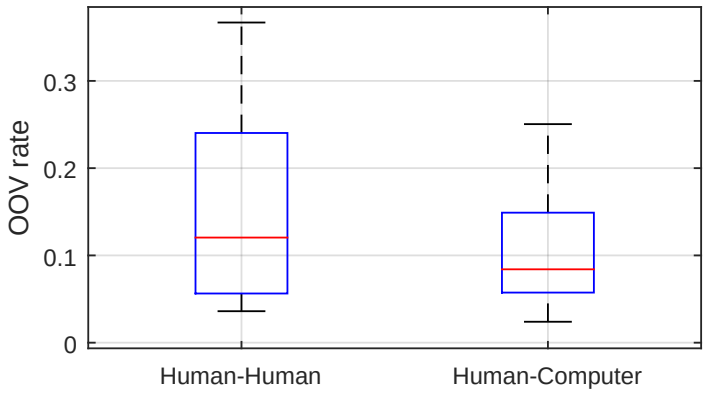


Figure 2.8: The speaker-related distribution of OOV word frequency. *Human-Human* corresponds to 1400 responses from seven subjects in measurements with a human supervisor. *Human-Computer* covers 2400 responses from 20 subjects in unsupervised measurements with normal-hearing subjects.

2.4 Discussion

For several fields in speech research and signal processing, the use of data obtained in real-world measurements was found to produce different results than controlled data, e.g., read speech with artificially added noise in ASR (Barker et al. 2015), or speech that is artificially reverberated

with prerecorded room impulse responses (Kinoshita et al. 2016). In this paper, we also observed differences between the controlled read corpora and the speech produced during actual measurements:
Although the databases containing read and spontaneous speech were recorded under the same acoustical conditions, they differ in several aspects such as the speaking rate, the amount of OOV words, and the HL of the subjects. The lower speaking rate of spontaneous responses might be explained with a higher cognitive load required for listening to, understanding and repeating noisy speech. Although ASR training was performed with the read sentences, we didn't observe a significant effect of speaking rate with any corpus tested in this study, which indicates that the variance in the training data was still sufficient for the ASR system for generalizing to spontaneous utterances. Similarly, the subject's age did not have a significant effect on error rates, again, presumably since this factor is covered by the training data and consequently by the acoustic model. Hence, both parameters as covered by the different test databases in this study are not critical for the accuracy of SRT_A for the proposed system.
On the other hand, the number of out-of-vocabulary words had a major impact on speech recognition, especially on the ASR insertion error rate. The OOV rate ranged from 0% (*read-gap*) over 10.6% (*spontaneous*) and 16.4% ($spontaneous - HI$) to 28.0% (*read-HH*), which resulted in vastly different insertion error rates (1.8, 12.6, 17.2 and 32.0%, respectively). At the same time, a moderate increase for the deletion rate was observed (from 0.7 over 1.0 and 0.7 to 3.7%).
The first simulation of measuring SRT_A (Section 2.3.3.1) has shown that a low deletion rate is more important than a low insertion rate. This is presumably caused by the structure of the matrix test, since a match of erroneously inserted and presented matrix words has a small probability. A fraction of just 10% that affect SRT_A can be expected for random insertion errors. Consequently, a special instruction set for subjects tested by an automated test (such as completely avoiding OOV words) seems not to be required, as long as the deletion rate is kept low (even at the cost of high insertion error rates, which have a negligible effect on SRT scoring). The second simulation has shown that these requirements are met during automated test conduction, and reliable SRT

measurements can be obtained. This is true for the controlled and spontaneous measurements: The largest deviations from the measure SRT_H obtained from a human supervisor was in the range of 0.1 dB, and therefore well within the test-retest accuracy that can be expected from SRT measurements with matrix sentence tests. Therefore, it seems that the differences between controlled and spontaneous speech are less relevant in automated speech audiometry than for speech recognition or dealing with room reverberation. An additional error type was not considered in the simulation, i.e., confusion errors produced by subjects (which are very likely to introduce score insertion errors, e.g. "Boris" instead of "Doris"). Despite the fact that a considerable fraction of the OOV words ($\approx 14\%$) results from this error type, the good fit between the measured and simulated SRT values with the ASR error rates for the spontaneous speech indicates that confusion errors play a minor role for the accuracy estimation of SRT_A. Further, partially produced words could result in insertion errors, but in our measurements, the percentage of partial words is below 2% among OOV words, indicating that partial word production does not play a major role in our experiments. In contrast to our own previous work (Meyer et al. 2015) as well as other groups (Deprez et al. 2013), the proposed system can be used to accurately determine SRT_A. This study investigates the main target group of matrix sentence tests, which are NH or slightly to moderate HI patients. While we didn't find any difference between these two subject groups, future research needs to address the interaction with severely impaired hearing, which can result in impaired speech (Arlinger 2003). This could require ASR adaptation to account for a larger variance of speech features. Furthermore, the assumption of a maximum performance level of $p_{max} = 1$ can be incorrect for this patient group, which should be considered in future research.

2.5 Summary

This study investigated a system for an automated, ASR-controlled conduction of a listening test, which is based on the Oldenburg matrix sentence test and used to determine the SRT of a listener. Two ASR experiments were performed using either pre-recorded, controlled and read utterances produced by 17 speakers, or data collected during actual test

conduction with realistic data collected from 20 NH and 7 HI subjects. The test accuracy for spontaneous speech was compared to SRT values measured by a human supervisor, and found to be dominated by the test-retest variance which is in the range of 0.5 dB for NH subjects and 0.8 dB for HI subjects. Hence, no significant differences between automated and human-controlled test conduction were observed. To explore if the controlled speech data is suitable to determine the reliability of the automatically determined SRT_A or if data collected during real measurements is required, a simulation was proposed that predicts the outcome of SRT measurements given the pre-recorded utterances based on Monte Carlo tests. Despite the fact that the speaking rate was found to be lower for spontaneous speech (presumably since it was collected during test time, imposing additional cognitive load on the subjects), and the rate of out-of-vocabulary words was lower (despite the fact that listeners were not instructed to avoid OOV words), it was shown that the simulation accurately predicts the automated measured SRT_A. Hence, it is possible to predict the test outcome with controlled speech material alone, which is attributed to the adaptive SRT measurement procedure that was shown to be relatively robust against ASR errors.

Acknowledgments

This work was funded by the Cluster of Excellence 1077/1‘ “Hearing4all” and the SFB/TRR 31 ’The active auditory system’ funded by the DFG. The authors would like to thank Franz Kunze for the help with the conduction of the measurements and the labeling of the data, and Constantin Spille for valuable contributions and discussions to this work.

CHAPTER 3

Self-conducted speech audiometry for users of hearing aids and cochlear implants

Abstract

Speech-in-noise tests are an important tool for assessing hearing impairment, the successful fitting of hearing aids, as well as for research in psychoacoustics. An important drawback of many speech-based tests is the requirement of an expert to be present during the measurement, in order to assess the listener's performance. This drawback may be largely overcome through the use of automatic speech recognition (ASR), which utilizes automatic response logging. However, such an unsupervised system may reduce the accuracy due to the introduction of potential errors. In this study, two different ASR systems are compared for automated testing: A system with a feed-forward deep neural network (DNN) from a previous study (Ooster, Huber, et al. 2018), as well as a state-of-the-art system utilizing a time-delay neural network (TDNN). The dynamic

This chapter is a formatted reprint of
***Jasper Ooster**, Laura Tuschen and Bernd T. Meyer,*
"Self-conducted speech audiometry for users of hearing aids and cochlear implants",
Submitted to Computer, Speech & Language.
Author contributions: JO developed and implemented the proposed ASR systems, performed the measurements with the NH and the severely HI listeners, prepared the figures and wrote the manuscript. LT conducted the measurements with the CI listeners. All authors co-wrote the manuscript (with the main contribution coming from JO).

measurement procedure of the speech intelligibility test was simulated considering the subjects' hearing loss and selecting from real recordings of test participants. The ASR systems' performance is investigated based on responses of 74 listeners, ranging from normal-hearing to severely hearing-impaired as well as cochlear implant listeners. The feed-forward DNN produced accurate testing results for NH and unaided HI listeners but a decreased measurement accuracy was found in the simulation of the adaptive measurement procedure when considering aided severely HI listeners, recorded in noisy environments when measuring with a loudspeaker setup. The TDNN system produces error rates of 0.6 % and 3.0 % for deletion and insertion errors, respectively. We estimate that the SRT deviation with this system is below 1.38 dB for 95% of the users. This result indicates that a robust unsupervised conduction of the matrix sentence test is possible with a similar accuracy as with a human supervisor even when considering noisy conditions and altered or disordered speech from elderly severely HI listeners and listeners with a CI.

3.1 Introduction

Speech intelligibility in noisy conditions is a key element in daily social interaction and communication, which is often reduced for hearing-impaired (HI) listeners. Speech-in-noise tests are a common method to evaluate this capability of a listener by measuring the speech recognition threshold (SRT) which is the signal-to-noise ratio (SNR) corresponding to 50 % average intelligibility. Speech-in-noise tests can give insights into aspects of hearing impairment that are not reflected in the audiogram but play an important role in the perceived strength of a HL or the performance of a hearing aid. However, an important drawback of such tests is the effort to conduct these in a clinical test environment, since a human supervisor needs to be present during the measurement to log the responses of test subjects. An automated procedure could reduce this effort and thereby increase the testing rate of speech-in-noise tests within a clinical context. For screening purposes, there are several approaches to conduct speech-in-noise tests, from which the most prominent representative is the digit triplet test, also known as the digit-in-noise test (Smits, Kapteyn, et al. 2004; Smits, Merkus, et al. 2006; Vlaming et al.

2011; Melanie A Zokoll et al. 2013; Potgieter et al. 2016). While these tests can be easily conducted without supervision by using a keypad to capture the listeners' responses, they have the disadvantage of limited vocabulary and phonetic variance. Therefore, they are only used for screening rather than for clinical diagnostics.

Francart et al. (2009) proposed a framework with a written feedback system including automated typo correction for the automated conduction of the Leuven intelligibility sentences test (LIST) (Van Wieringen and Wouters 2008) which achieved the same measurement accuracy as a human supervisor. Deprez et al. (2013) firstly proposed using automatic speech recognition (ASR) for automated conduction of the LIST and achieved a 9.3 % false alarm rate and a 90.7 % keyword detection rate, which resulted in a bias of 0.2 dB and an increase of the test results standard deviation from 1.2 dB to 1.8 dB, for 17 normal-hearing (NH) listeners. Due to the limited vocabulary, matrix sentence tests provide the possibility to use a graphical user interface for capturing the listeners' performance in a clinical context by providing all possible responses (Kollmeier, Warzybok, et al. 2015). However, this limited vocabulary also allows for a specialized ASR system to be built, which is able to capture the listeners' responses using a purely speech-based interface without the drawbacks of a graphical user interface: In previous work, an ASR-based prototype for an unsupervised measurement of the matrix sentence test was developed and evaluated (Ooster, Huber, et al. 2018). This prototype features an ASR component that was trained to recognize only the words from a speech-in-noise test for automatic scoring of spoken responses produced by test users. The transcript from the ASR system is used to dynamically adapt the SNR (as described below). When testing the system with 20 NH and seven mildly HI, either producing well-controlled (read) as well as spontaneous responses, we did not observe a influence on the SRT measurement accuracy (with 0.9 % deletion and 2.9 % insertion errors on the scoring, respectively). To address the influence of ASR errors on the SRT measurement accuracy with the matrix sentence test, two simulations using Monte-Carlo methods were proposed. The first method addresses the general influence of these errors by taking the insertion and deletion errors as free

parameters. This simulation suggested a high robustness of the adaptive procedure against ASR errors with a stronger sensitivity to deletion errors. The second method utilizes actual transcripts from an ASR system that uses real speech responses as input to estimate the performance for a specific target group.

While the results from Ooster, Huber, et al. (2018) seem to be promising, all measurements were carried out using a headphone-based setup. This results in clean speech recordings as the masker noise does not get captured by the microphone. When using loudspeakers in free-field setups, which is required for testing users with hearing aids or cochlear implants (CI), the masker noise needs to be continuously presented to allow for an adaptation of the hearing aid, which consequently results in noisy speech recordings. Furthermore, speech production can change with the age of a speaker (Mortensen et al. 2006) and severe HL and profound deafness can result in disordered speech (Leder and Spitzer 1990). In CI users, the speech production quality depends on the onset and duration of deafness (Ruff et al. 2017). This can influence the ASR systems performance when testing with elderly and severely hearing impaired listeners (Vipperla et al. 2008; Moore et al. 2018).

In this study, we investigate if current ASR technology can be used for an automatic conduction of speech-in-noise tests for test users with very different hearing profiles, and if the ASR robustness is sufficient to conduct such tests in free-field environments with a high level of background noise. To this end, data collected from measurements with 73 subjects is analyzed (ranging from NH over unaided mild & moderate HL, aided severe HL, to listeners with a CI). The ASR system from Ooster, Huber, et al. (2018) serves as a baseline system. It follows the classical structure as described by Hinton et al. (2012) utilizing a fully-connected, feed-forward, deep neural network (DNN) hybrid system with a hidden Markov model (HMM) language model. The baseline system's performance is compared to a implementation that reflects the current state of the art, using a factorized time-delay neural network structure (TDNN-f) trained with a lattice-free maximum mutual information cost function (LF-MMI), also combined with a HMM language model (Povey, Cheng, et al. 2018). This system represents one of the most advanced hybrid systems available in the open source speech recognition toolkit *kaldi* (Povey,

Ghoshal, et al. 2011). The TDNN topology uses temporal sub-sampling to incorporate a wider temporal context with fewer parameters in comparison to a fully-connected DNN (Waibel et al. 1989). The LF-MMI cost function takes the whole utterance into account, rather than working on a per-frame level as the previously used cross entropy cost function (Povey, Peddinti, et al. 2016). An end-to-end system based on a recurrent neural network structure was not considered for this study. The end-to-end models learn the full representation from the feature level up to a grapheme or word level. This implies that the training and test data need to match on sentence level, so that the model can learn long-term dependencies. Even though there is training data available containing matrix sentences, this data did not cover the full variability in the spontaneous responses that occur during realistic measurements. With the hybrid models investigated here, it is possible to incorporate knowledge about the test data in the language model independently, and the additional variability in the spontaneous responses is covered by a garbage model. Training data containing both filler words and matrix words are not required for this method.

The ASR systems used in this study are evaluated by using speech responses that were recorded from the corresponding listener groups. To take into account the dynamic nature of matrix tests (i.e., the adaption based on the number of correct/incorrect response words), we use the second simulation scheme from our previous work (Ooster, Huber, et al. 2018) as described above. Our aim is to quantify the reliability for ASR technology for different patient groups and noisy scenarios for automated speech audiometry.
The remainder of this paper is structured as follows: Section 3.2 describes the speech audiometric measurement procedure, the implementation of the ASR system and the simulation methods to infer the resulting SRT measurement accuracy. Section 3.3 describes the ASR results for the two compared ASR system as well as the simulated influence on the SRT measurement accuracy. Section 3.4 discusses the measurement results and Section 3.5 provides a summary of the main findings.

3.2 Methods

3.2.1 Matrix sentence test

Speech intelligibility tests try to capture the capability of individuals to understand speech in difficult conditions and are an important measure in diagnostics. The matrix sentence tests (Hagerman 1982; Wagener, Kühnel, et al. 1999a; Kollmeier, Warzybok, et al. 2015) are an efficient tool to measure parameters of the psychometric function, i.e., the relation of SNR to intelligibility, which can be described with a logistic-sigmoid function. The most important target value of the matrix tests is the SRT, i.e., the SNR corresponding to 50 % speech intelligibility. Matrix sentence tests considered in this study are constructed from a five-by-ten matrix so that the sentences are syntactically fixed but semantically unpredictable (Wagener, Kühnel, et al. 1999a). The matrix sentence tests exist in more than 20 languages with a similar structure which should enable measurements that can be compared across languages.

During the measurement, the matrix sentences are presented to the subject with a speech-shaped stationary noise, which is generated by multiple overlaps of the stimulus sentences. This ensures the noise to exhibit the same long-term spectrum as the sentences, while at the same time the noise does not contain any audible speech segments. During the measurement, 20 sentences in noise are typically presented, and the SNR is dynamically adjusted with the aim of approaching 50 % recognition rate: When more than 50 % of words are correctly identified by the listener (i.e., three or more words), the SNR is decreased. Otherwise, the SNR is increased and the task becomes easier for the listener. The SNR step size is gradually decreased during the measurement after each SNR reversal to support convergence towards the SRT (Wagener, Brand, and Kollmeier 1999). The SRT measurement outcome is estimated by a maximum likelihood fit of a psychometric function to the data points (Brand and Kollmeier 2002).
For an (unsupervised) conduction of the matrix sentence test, it is necessary to estimate the score, i.e., the number of correctly recognized words from the stimulus, so that the SNR can be adapted for the next presentation.

3.2.2 Automatic speech recognizer

The ASR system is the core element of the unsupervised measurement system and both ASR systems which are analyzed in this study are realized as a deep neural network - hidden Markov model (DNN-HMM) hybrid system (Hinton et al. 2012), which has a separate acoustic and language model. The first system is the exact ASR system which was used in (Ooster, Huber, et al. 2018) to which we refer to as **f**ully-**c**onnected **d**eep **n**eural **n**etwork with a **l**anguage **m**odel for all **50** matrix test words (*FC-DNN-LM50*). The second ASR system is realized as a **t**ime-delay **n**eural **n**etwork with a sentence specific **l**anguage **m**odel for the respective **5** words (*TDNN-LM5*). Both ASR systems are implemented using the speech recognition toolkit *Kaldi* (Povey, Ghoshal, et al. 2011) and are based on publicly available training recipes.

3.2.2.1 Acoustic model

The acoustic model infers posterior probabilities for triphones, i.e., phoneme classes in which the neighbor phonemes are considered, from the acoustical data, which are later combined to words and sentences by the language model. For both ASR systems the acoustic model is trained on an in-house database (Meyer et al. 2015), which contains German matrix sentences as well as two commercially available databases: King-ASR-L-092 and King-ASR-L-182 (see `kingline.speechocean.com`) consisting of 16,000 utterances (18 hours) from 40 different speakers. The self-recorded data base consists of 27,000 utterances (23 hours) from 20 different speakers (10 female, 10 male). The speakers' ages ranged from 21 to 60 years with an average age of 35.8 years. The training targets for both DNNs (triphones, see above) are generated with a Gaussian mixture model (GMM)-HMM system trained on the clean audio with 13 dimensional Mel-frequency cepstral coefficients (MFCC) features (plus the first and second temporal derivative - delta and double delta) with speaker adaptive training as described in (Vesely et al. 2013).
The *FC-DNN-LM50* ASR system is trained on clean data sampled at 44.1 kHz and uses the same speaker adaptation on the features as the GMM system used to align the training data together with a splicing of ± 5 frames. It is a fully-connected, feed-forward, deep neural network with three hidden layers - each of which consists of 1024 neurons and

a sigmoid nonlinearity, which combines to a total of 5.6 million parameters. It follows the implementation of the DNN-baseline as described in (Vesely et al. 2013), further details of which can be found in Ooster, Huber, et al. (2018, Section 2.3).

For the *TDNN-LM5* ASR system which combines recent advances in speech technology, all audio data is down-sampled to a 16 kHz sampling frequency with a 16-bit resolution. Data augmentation is implemented by adding copies of the training data with artificially decreased or increased speed (with factors or 0.9 and 1.1, respectively), which should increase the variability of the training data with respect to speaking rates. This extended data set was again doubled, and noise was added to the copy using noise signals from the *MUSAN* corpus (Snyder et al. 2015). The *MUSAN* corpus contains technical noises (dialtones, fax machine noises and more) as well as ambient sounds (footsteps, paper rustling, car idling, crowd noises with indistinct voices and more), which were added at random SNRs selected from (+5, +10, +15, +20) dB. The resulting training set is six times larger compared to the original set and contains clean and noisy signals at different speaking rates.
For each frame, 40-dimensional MFCCs (without any speaker adaptation) are grouped with 100-dimensional speaker identity vectors (i-vectors) (Saon et al. 2013) serve as input to the DNN. The acoustic classification is done with a factorized time-delay deep neural network (TDNN-F) and follows the implementation of Povey, Cheng, et al. (2018). Overall, the TDNN-F has 13 hidden layers - each of which consists of 1024 neurons, a 128 dimensional bottleneck, and a rectified linear unit (ReLU) nonlinearity. The first three TDNN-F layers have a temporal context of -1,0,1, the fourth layer has no temporal slicing and the remaining higher layers have a slicing of -3,0,3 (following the notation of (Peddinti et al. 2015)).
Moreover, the network has residual neural network (ResNet) skip connections between the hidden layers. This results in 8.4 million parameters in the neural network and an overall temporal context of 29 frames. The network is trained with a lattice-free maximum mutual information (LF-MMI) cost function (Povey, Peddinti, et al. 2016). To prevent overfitting, four different regularization methods are used: l_2-norm reg-

ularization, batch-norm regularization of all layers, regularization with a second output layer which is trained with a cross-entropy loss function, and leaky HMM states Povey, Cheng, et al. (2018, Section 2.7.). Furthermore, linear dropout is used to increase the redundancy within the neural network: 0 % for the first 20 % of the training epochs, which linearly increases to 50 % dropout probability at 50 % of the epochs, and then linearly decreases back to 0 at the end of the training.

3.2.2.2 Language model

The language model combines the outputs of the DNN, which correspond to posterior probabilities of triphones, to the most likely word sequence using weighted finite-state transducers (WFST) (Mohri et al. 2008). The lexicon of the language model of both ASR systems contains phonetic transcription for 19k words extracted from the *MaryTTS* system (Schröder and Trouvain 2003). The full lexicon is only used while training the model to translate the word labels into phonetic labels. During testing, only the phonetic transcriptions of the German matrix sentence words are required. To handle words observed during test time that are not part of the matrix test ('out-of-vocabulary (OOV)' words), a phone-level 4-gram garbage model was trained on the 19k words of the training lexicon. This ensures that OOV words do not disturb the alignments and recognition of the matrix test words and that non-matrix test words are not falsely transcribed as matrix test words, which would result in insertion errors.

The two ASR systems differ on the grammar level, which defines word transition probabilities of the language model. As noted above, the *FC-DNN-LM50* ASR system uses a single unigram grammar for all test sentences, which includes all 50 German matrix test words with the same probability and the garbage model for OOV words.
The *TDNN-LM5* system uses a unigram grammar that is specific for each stimulus sentence. This is possible since - in the application of the unsupervised measurement system - it is always known which stimulus sentence is presented. Therefore, we can construct a sentence-specific decoding graph, which increases the probabilities for the words of the stimulus sentence to occur, as the five words of the stimulus sentence are

more likely to be in the response of the subject than the other 45 words of the matrix test.
Furthermore, in the grammar of *TDNN-LM5*, the 20 most frequent OOV words from the NH subjects data in Ooster, Huber, et al. (2018) were included with a low probability in the sentence specific unigram model. Besides common filler words such as *nichts* (nothing) and *irgendwas* (something), common confusion words such as *Rosen* (roses) instead of *Dosen* (cans) and *Boris* instead of *Doris* are included. This NH data was also used as a development set to fine-tune hyper parameters such as the probabilities in the unigram, and the weighting between the acoustic model and the language model. It was, however, not used in the training of the acoustic model. Furthermore, the pronunciation probability is adjusted for the *TDNN-LM5* system, based on the occurrence in the training data (Chen, Xu, et al. 2015). This also includes phoneme-specific probabilities for the silence class.

3.2.2.3 Evaluation metric

The performance of the ASR system of the unsupervised measurement prototype is evaluated by the errors that the system makes in the scoring of the subjects' responses, rather than by all the errors within the full transcript, since the ASR system is built and optimized for this purpose. These errors in the scoring are measured by the *score deletion rate* (SDR) and the *score insertion rate* (SIR):

$$SIR = \frac{N_{score\ insertions}}{N_{subject\ score}}, SDR = \frac{N_{score\ deletions}}{N_{subject\ score}} \tag{3.1}$$

$N_{subject\ score}$ is the number of correctly recognized and uttered words of the subject in response to the stimuli. This metric is calculated for each list, i.e., for 20 sentences. Since the adaptive procedure of the matrix test aims at achieving 50% intelligibility, this results in $N_{subject\ score} \approx 50$. Note that the order and repetition of words are neglected by this metric - just as a human supervisor would do.

3.2.3 Evaluation data

In this study, differences between the ASR-based measurement and the supervised clinical measurement are estimated with a simulation scheme

	N (f/m)	Age [years]	PTA [dB HL]
NH	20 (10/10)	28 ± 3	3 ± 3
Unaided HI	26 (17/9)	69 ± 7	30 ± 6
Aided, severely HI	13 (5/8)	72 ± 8	66 ± 11
CI	14 (10/4)	49 ± 19	> 60

Table 3.1: Statistics of the four subject groups who participated in the evaluation.

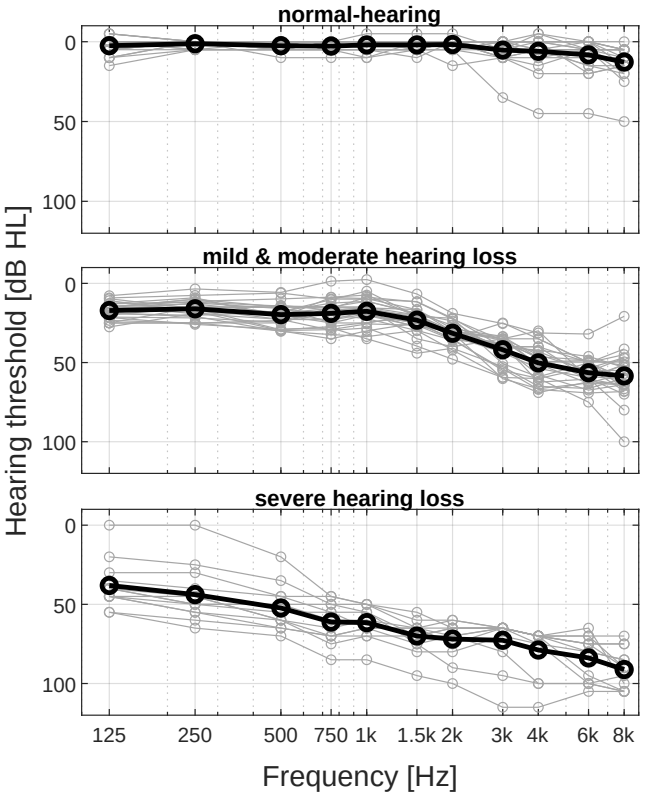


Figure 3.1: Individual audiograms of the better-hearing ear of our subjects (gray lines) together with the average audiogram for the respective subject group (black lines).

that uses real recordings obtained during actual SRT measurements (with the exception of the CI user group, for which read responses were used). This simulation scheme is described in detail in Section 3.2.4. The recorded responses are collected with four different subject groups as described in Table 3.1, Figure 3.1, and the following subsections.

Figure 3.2 shows an overview of the measurement and the recording of the spontaneous evaluation data. Two different approaches are used for audio recordings:

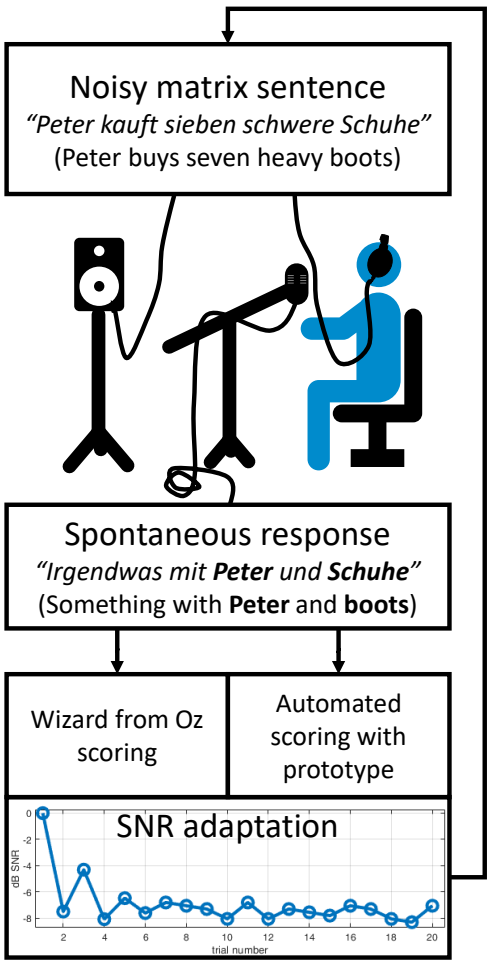


Figure 3.2: To record spontaneous responses, matrix sentence test measurements are either conducted with a loudspeaker setup and a hidden human supervisor ("Wizard of Oz") or in a headphone setup controlled by an ASR system, i.e., a prototype of the unsupervised measurement system.

When headphone presentation is used for testing (NH and unaided HI listeners), it is possible to record the listeners' responses as clean audio. In this case, a prototype of the unsupervised measurement system is used for a fully automated test conduction: First, an energy-based speech activity detection (SAD) controls the duration of the recording which is then fed into the ASR system, which produces a transcript of the response. The score is estimated based on the recognition results from the ASR system, after which the SNR is adapted automatically for the next presentation as specified in the test procedure. The prototype uses the *FC-DNN-LM50* ASR system. The audio recordings of the subjects' responses were manually transcribed on word level after the measurements

to generate human reference labels for the ASR system's evaluation.
For measurements with a loudspeaker (used for aided, severely HI participants), the continuous masking noise is also picked up by the microphone, and the recorded speech signal is noisy. Before conducting the study, it was not clear if these conditions allow for a regular conduction of the test, i.e., if the ASR system is sufficiently robust to the masking noise. Hence, these measurements were conducted in a "Wizard of Oz" setup, where the subjects were told that they were talking to an automated system, while a hidden human supervisor controlled the actual measurements. The scoring of the human supervisor were also used as labels to evaluate the ASR system's performance later.
The subjects of the spontaneous responses all received standardized written instructions - stating that they are interacting with an automated system. No limitations were applied to their response behavior. The subjects with a CI did not record spontaneous responses, but read a list of 30 matrix sentences recorded in a clean environment. All subjects were compensated for their participation in this study.

3.2.3.1 Normal-hearing listeners

This data set contains spontaneous responses collected from 20 young NH listeners. Their responses were recorded using a prototype of the unsupervised measurement system and originally published in (Ooster, Huber, et al. 2018). In this study, they serve as a reference and are used for optimizing parameters of the *TDNN-LM5* ASR system. The subjects were selected based on the $\mathrm{PTA}_{0.5,1,2,4} < 20\,\mathrm{dB\,HL}$ NH criterion (Mathers et al. 2000). Each subject conducted six measurement lists (including two training lists) with the standard speech-shaped noise of the matrix test (so-called *olnoise*) at a fixed level of 65 dB through headphones. The measurements were automated using the ASR prototype mentioned above and conducted in a sound-isolated listening booth. Subjects' responses were recorded with a small membrane condenser microphone with a cardioid characteristic (Neumann KM 184), in conjunction with a RME Fireface USB UC soundcard with 44.1 kHz sampling frequency and 32 bit resolution, at a distance of approximately 0.5 m to the subject. The measurement software was implemented in MATLAB.

3.2.3.2 Unaided hearing-impaired listeners

The unaided HI subjects had mild to moderate hearing losses (25 dB HL $< \mathrm{PTA}_{0.5,1,2,4} <$ 45 dB HL). Eight of the 26 subjects used a hearing aid on a regular basis but did not wear it during the measurements. The data from this group was collected using the prototype system utilizing two measurement setups: For seven subjects, the same setup as for the NH group was used with six measurement lists, with stationary speech-shaped noise being added to the presented signals. The remaining 19 subjects performed twelve measurement lists (including two training lists) with different noise maskers - ranging from stationary speech-shaped noise to a single talker interferer. The stimuli were presented monaurally through headphones (Sennheiser HDA200) to the subjects, the responses of which were captured using a close-talk condenser microphone with a large membrane (Neumann TLM 103). The presentation over headphones implies that the ASR system is not affected by the different noise types: The speech from both setups described in this subsection was recorded without background noise. An RME Fireface UCX soundcard was used for stimulus presentation and recording of the subjects responses, and the measurement was controlled by the Oldenburg Measurement Applications (R&D version 2.0) software.

3.2.3.3 Aided, severely hearing-impaired listeners

Subjects from this group are supplied with a hearing aid for both ears and have a severe or close-to-severe HL ($\mathrm{PTA}_{0.5,1,2,4} >$ 55 dB HL) in both ears. They were all experienced users of their hearing aids ($>$ 10 years). They were not tested for speech disorders related to HL and therefore were not selected based on this criterion. These subjects cannot partake in the matrix sentence test without their hearing aids: To reach 50% intelligibility, it would require high sound pressure levels of the target speech, when the fixed noise level is high enough for them to hear the noise. Nevertheless, since this subject group might be measured with the matrix sentence test for a hearing aid fitting, the unsupervised measurement system should be robust for this group of listeners. Therefore, this subject group was measured using a loudspeaker setup. To allow for a potential adaptation of the hearing aid algorithms, a continuous speech-shaped noise was presented during the measurements.

Before the measurement, a human supervisor introduced the subjects to the measurements with standardized written instructions. During the measurement, the supervisor was outside the booth, not visible to the subjects, and listened to the subjects' responses to score the results for the next stimulus SNR adaptation. If required, the human supervisor could guide the subjects through the measurements with pre-generated, short, synthesized speech instructions.
For the stimulus presentation a Genelec 8030B was placed at 1.5 m distance in front (0°) of the subjects and calibrated to a noise level of 65 dB(A). The measurement was conducted with the Oldenburg Measurement Applications (R&D version 2.0) software and a Focusrite 2i2 soundcard. For the synthesized instruction, another loudspeaker (Genelec 8330A) was placed at an azimuth angle of 20° with the same 1.5 m distance to the subject. For the instruction, the loudspeaker was controlled by an RME Fireface UCX soundcard, which prevents interference with the audio drivers of Oldenburg Measurement Applications. This soundcard was also used to record the subjects' responses with a large membrane condenser microphone (Neumann TLM 103) at a distance of approx 30 cm to the subject.

3.2.3.4 Noisy versions NH and unaided HI

A noisy version of the clean data was generated to disentangle the differences between the acoustic condition and potential pathological speech. These noisy data versions of the data sets should recreate the acoustic conditions of the aided, severely HI subjects as accurately as possible. All three data sets were recorded using a high-quality cardioid microphone in the same isolated listening booth. Therefore, the transfer functions of the speech to the recording microphone are very similar. Due to the higher distance and the off-axis positioning, the transfer functions from the loudspeaker (noise source) to the microphone might be more dependent on the exact positioning. To capture this potentially higher variability, the transfer function from the loudspeaker (noise source) to the microphone was measured before each measurement and after every break in each measurement (and therefore potential repositioning of the microphone and subject) with an exponential frequency sweep. The exponential frequency sweep had a range from 75 Hz up to 22.05 kHz over

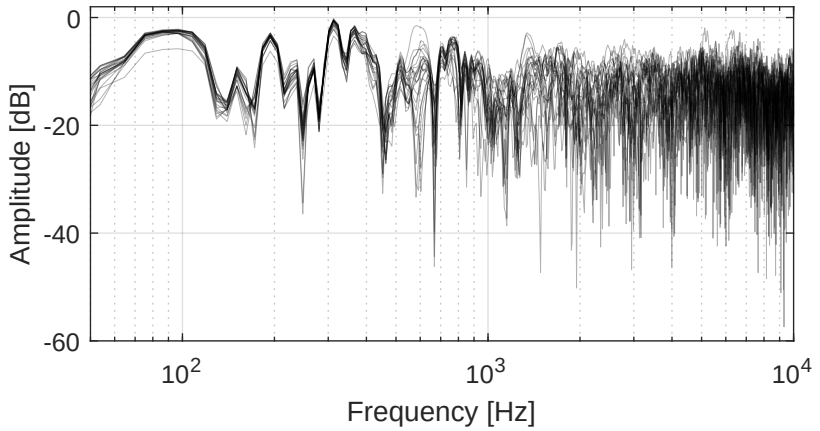


Figure 3.3: 24 different frequency responses from the stimulus loudspeaker to the speech recording microphone. The frequency responses are measured with an exponential frequency sweep with 13 different subjects.

a duration of 5 s. It was measured with the subject and the microphone already positioned. The measured exponential frequency sweep was deconvolved with the original sweep signal, in order to obtain the transfer function (Farina 2000). Figure 3.3 shows 24 different transfer functions which were measured with this procedure with the 13 different subjects (not all subjects took a break during the course of the measurement and therefore, the microphone was not re-positioned in every measurement session). To find the correct SNR values to mix the convolved speech-shaped noise to the recorded clean responses, an energy-based SAD was used to separate the audio recordings of the whole measurements lists recording into the non-speech parts, where only the continuous speech shaped-noise is present, and the speech parts. The root mean square (RMS) value of all these speech/non-speech parts was used to estimate the average SNR over a whole measurement list. Figure 3.4 shows a box plot over the six measurements for each of the 13 severely HI subjects. For every (clean) utterance from the NH and unaided HI subjects, one of these SNRs as well as one of the measured transfer functions were randomly selected to create a noisy version.

3.2.3.5 Cochlear implant listeners

This group consists of 14 subjects who participated in the measurements of the THERESIAH project which aims at developing a new therapy sys-

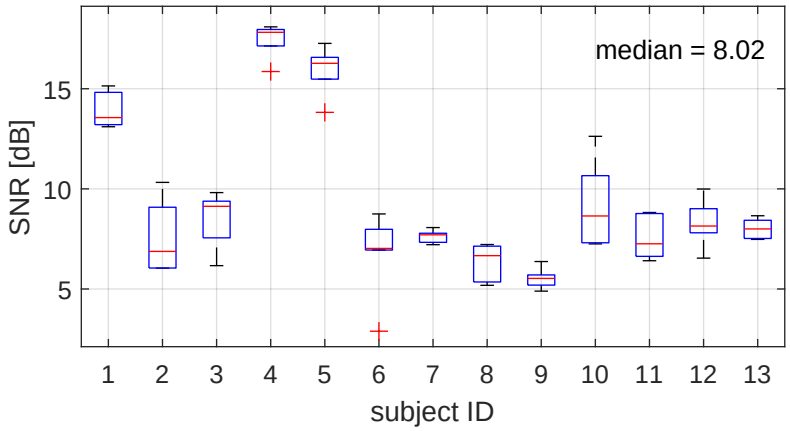


Figure 3.4: SNR of the speaker's responses to the recorded background noise during loudspeaker measurements with hearing aids for each of the 13 subjects of the severely HI group. One SNR is calculated per measurement list, so there are six data points per subject in the box plot.

tem to train hearing and articulation of highly hearing-impaired persons. These are listeners who either have two CIs or are bimodally supplied, i.e., they have one hearing aid and one CI. For two of the subjects, the onset of hearing impairment was post-lingual. The other 12 subjects had their hearing impairment before the onset or during language development. Hearing impairment-related speech sound disorders were detected in 13 out of 14 subjects.

This group did not conduct SRT measurements, but read a list of 30 complete matrix sentences in a quiet acoustic environment, recorded with a Tascam DR-100 MKII and an AKG C520 professional head-worn condenser microphone.

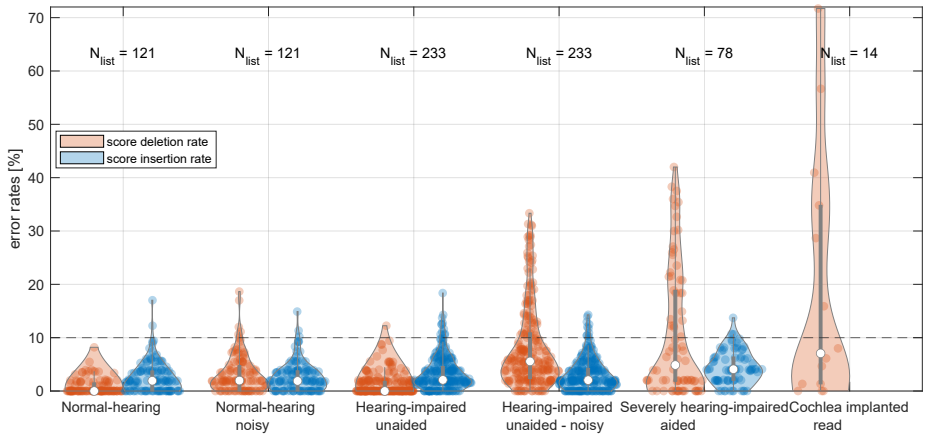
3.2.4 Simulation on the ASR error's influence on the SRT measurement accuracy

In this study, it is not possible to directly compare the SRT measurement outcome SRT_{ASR} obtained autonomously based on the ASR decoding to the $STR_{Reference}$ obtained with a human supervisor, since all ASR performance evaluation is done as post-processing and the adaptive measurement procedure relies on the ASR decoding. Nevertheless, based on Monte-Carlo simulation methods, it is possible to simulate the adaptive measurement procedure to infer the influence the ASR errors have

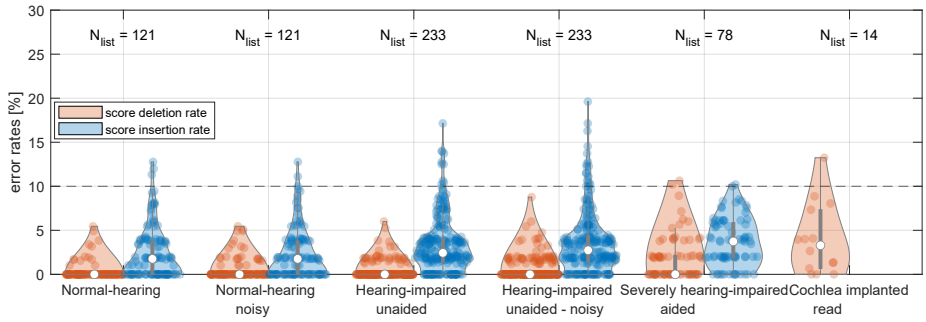
on the measurement accuracy. This simulation method was proposed in Ooster, Huber, et al. (2018, Section 2.6.2) and takes the errors that an ASR system makes on a specific data set into account. It accurately reproduced the results obtained from real adaptive measurements with the prototype system Ooster, Huber, et al. (2018, Section 3.3.2). The ASR errors utilized in the simulation are measured for the real recordings, as described above.

For this simulation, the probability of a subject to recognize a word from the stimulus sentence correctly is taken from a psychometric function. This psychometric function is defined before the simulation run randomly from the distribution of the respective subject group. The recognition of each word is carried out as a Bernoulli trial, while the number of independent elements per stimulus sentence (between 3.18 and 4.29) is also considered (Bronkhorst et al. 2002; Brand and Kollmeier 2002). These Bernoulli trials give a simulated response score, i.e., the number of correctly recognized words by the subject (e.g. three out of five words in a sentence recognized correctly). This number is used to randomly select an audio file of the evaluation database which corresponds to this correct response score (which is known from the manually generated labels). The SNR is adapted based on the ASR output for this audio signal, which includes the errors of the ASR system.

The full standard measurement procedure is simulated with this scheme; starting at 0 dB SNR followed by the adaptive procedure with 20 sentences and the respective simulation of ASR errors. The parameters of the psychometric function are randomly selected from the SRT distribution of the respective subject group to account for the higher variability in the (severe) HI subjects. For each simulated measurement list, only responses from one single subject are selected and for each subject of the evaluation database. 200 measurement lists are simulated with different psychometric functions. To create a reference value, a second measurement run is simulated with the same psychometric function, but without any errors from an ASR system, to account for the test-to-retest variability of the measurement procedure itself. This should be a conservative approximation of the test with human supervisors, during which usually no errors occur when logging responses.



(a) FC-DNN-LM50



(b) TDNN-LM5

Figure 3.5: Violin plot of the ASR system's error rates on the utterances from the four different subject groups. The individual data points denote error rates of N_{list} single measurement lists; the width of the violin denotes the estimated probability density of the error rates. The median value and the interquartile range are denoted by white dots and the bold grey lines, respectively. The whiskers (thin grey lines) have a maximum length of 1.5 times the interquartile range. All values above the whiskers are treated as outliers. The horizontal-dashed line serves as orientation for better comparing the results of Figure a) and b) with the two different ASR systems.

3.3 Results

3.3.1 Errors of the ASR system

Overall 447 measurement lists from the four subject groups are recorded and analyzed. This results in 8640 unique spontaneous responses to noisy matrix sentences and 420 read sentences from the CI-subjects. When including the noisy versions of the audio, a total of 15k utterances are analyzed. Figure 3.5 shows the statistics of the ASR system's errors for each of the four subject groups. Since the error rates are

not normally distributed (Kolmogorov-Smirnov test, $p < 10^{-4}$ for all groups), the differences between subject groups are analyzed with the Mann–Whitney–Wilcoxon test, with a Bonferroni correction. Comparisons within a subject group are evaluated with the Wilcoxon signed rank test. A $p < 0.05$ was considered as significant.
Generally, the error rates are not equally distributed but clustered around multiples of 2%. In our analysis, the errors are evaluated for each measurement list, each of which contains 100 stimulus words. In a successful measurement run, the subject correctly recognized and uttered ≈ 50 words. Since the error rates are normalized with this number of correctly recognized words, a single false word corresponds to an error of $\approx 2\%$, which results in the clustering of the error rates.
For all data sets, it was possible to significantly reduce the SDR with the *TDNN-LM5* ASR system in comparison to the *FC-DNN-LM50* ASR system ($p < 0.002$). The differences in SIR was only significant in the noisy version of the NH data ($p = 0.02$). In all other data sets, the SIR difference was not significant ($p > 0.09$). The presentation of the results will focus on the results with the *TDNN-LM5* and only account the *FC-DNN-LM50* results partially as a reference.

3.3.1.1 Normal-hearing and unaided hearing-impaired subjects

The median SDR with the *TDNN-LM5* system is 0.0 % for the NH as well as the unaided HI subjects. The corresponding arithmetic means are 0.27 % and 0.33 %, respectively. All data points above 0.0% are treated as outliers. The median SIR is slightly elevated in comparison and reaches a value of 1.8% (NH) 2.4% (unaided HI). The data from the NH subjects was used as a development set to fine-tune parameters of the language model. While we did not find a significant difference in the SDR, the difference between these two data sets in SIR is significant (mean: 2.3 %(NH) / 3.3 %(unaided HI), $p < 0.002$).

3.3.1.2 Aided, severely hearing-impaired subjects

The severely HI subjects participated while they were wearing their hearing aid; hence, a free-field measurement with a loudspeaker setup was used, and the recordings of the subjects' responses are consequently noisy. Compared to the previously discussed listener group, the SDR strongly

increases to a median value of 4.9 % and maximum error 42 % per measurement list with the *FC-DNN-LM50* system. The SIR increases to a median of 4.1 % (mean 4.4 %), which is significantly higher in comparison to the NH and unaided HI data ($p < 0.01$). For the *TDNN-LM5* system, the SDR median value is still at 0.0 %, although the arithmetic mean (1.8 %) is significantly increased in comparison to the NH and unaided HI data ($p < 0.002$). The mean SIR of 3.7 % (median 3.7 %) with the *TDNN-LM5* is significantly higher than the NH data ($p < 0.002$), however no significant difference was found when compared to the unaided HI subjects ($p = 0.32$).

To disentangle the influence of potential speech distortions from the more difficult acoustic scenario, the clean audio from the NH and the unaided HI subjects was distorted to recreate the acoustic conditions of the aided, severely HI subject, as described in the methods section. The masker noise added to recordings of the NH and unaided HI subjects only slightly increases the error rates for the *TDNN-LM5* system as all data points above 0.0 % are rated as outliers (as mentioned above) and the median SIR stays at 1.8 for the NH data and increases from 2.4% to 2.7% for the unaided HI. Despite the median SDR of 0.0% in the noisy versions of the NH and unaided HI data, the differences in the arithmetic mean SDR (clean NH: 0.28 %, noisy NH 0.52 %, clean unaided HI: 0.33 % noisy unaided HI: 0.56 %) are found to be significant for the NH as well as the unaided HI ($p < 0.01$ for both). The difference in mean SIR was only found to be significant for the unaided HI data ($p < 0.01$) but not for the NH data ($p = 0.52$) when comparing the clean and noisy versions.

For the *TDNN-LM5* system, the worse acoustic condition cannot fully explain the higher SDR for the data from the aided, severely HI subjects, since the noisy versions of the NH and unaided HI data are also still showing a significant difference in the mean SDR ($p < 0.002$ for both). The mean SIR is only significantly different for the noisy version of the NH data ($p < 0.002$), but not for the noisy version of the unaided HI data ($p = 0.74$).

3.3.1.3 Cochlear-implanted subjects

The CI subjects read full matrix sentences, and therefore, the ASR system cannot make score insertion errors and no SIR is shown in Figure 3.5. Furthermore, since each subject read only one measurement list, each with 30 sentences in this data set, the number of lists corresponds to the number of speakers. The data from the CI subject group is the only group that showed a non-zero median $SDR = 3.3\%$ (mean 4.0 %) with the *TDNN-LM5* system. This is significantly higher than all other groups ($p < 0.002$), but not in comparison to the data from the aided, severely HI data ($p = 0.10$).

3.3.2 Simulated influence on the SRT measurement accuracy

The results of the simulations are shown in Figure 3.6. Since the simulations required recordings of incomplete sentences, it is not possible to conduct the simulations with the recordings of the CI subjects.
The upper panel in Figure 3.6 describes the expected outcome for the *FC-DNN-LM50* ASR system. The NH and unaided HI subjects have very similar error rates and thus the resulting simulated SRT shows a nearly perfect overlap. Both have a small negative bias, as the SDR's are lower than the SIR's. The aided, severely HI subjects showed error rates of up to 42%SDR. This results in positive bias, since the performance of the subjects is underestimated in this case. The spread of the measurement results is strongly increased and the 97.5 % percentile is almost doubled to a +2.05 dB mismatch.
For the simulation shown in the lower panel of Figure 3.6, the error rates of the ASR system were lower with the *TDNN-LM5* system, hence all three subject groups show an overlap.

3.4 Discussion

This study explored the unsupervised conduction of speech intelligibility tests. To this end, the ASR performance of two different ASR systems is compared utilizing recorded matrix test responses from four different subject groups. The two ASR approaches explored in this study differ in several design decisions, e.g., the temporal context taken into account, the number of parameters, the sampling rate, the cost function as well as

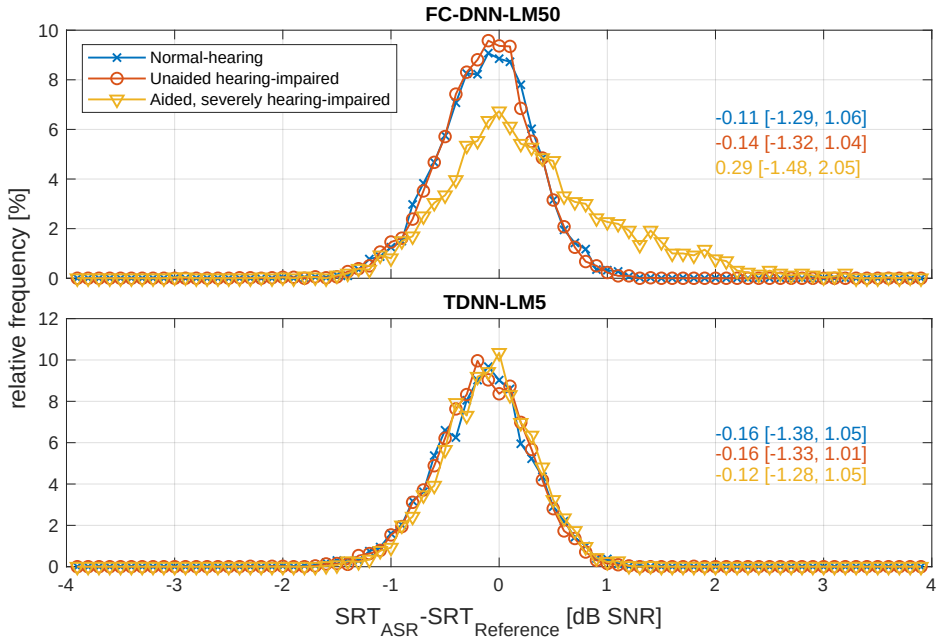


Figure 3.6: Normalized histogram of the difference between simulated SRT outcomes with (SRT_{ASR}) and without ($SRT_{Reference}$) errors from the ASR system. For each point, the subjects' response behavior is simulated based on a psychometric function; and an audio file is selected based on the simulated response to include the errors from the ASR system. It is not possible to conduct these simulations with the data from the CI-subjects, since they require recordings of incomplete sentences. The numbers on the right are the mean and the [2.5, 97.5] percentile points.

regularization methods during training, overall topology of the acoustic model, data augmentation and the structure of the language model. It would be interesting to explore the effect of each parameter separately, but this would result in a comparison of at least eight systems, which is out of scope of the current study. The reason we selected the TDNN system (*TDNN-LM5*), trained on the same speech data, is that it represents the the state-of-the-art which is available as open-source software which should foster reproduction of our approach. The reason the DNN-based system (*FC-DNN-LM50*) was chosen is to establish a direct comparability with previous work.

For NH subjects under good acoustic conditions, the previous study showed that it is possible to measure the matrix sentence test autonomously without a loss in SRT measurement accuracy (Ooster, Huber, et al. 2018). Even though the *FC-DNN-LM50* ASR system of that study worked fine

for HI subjects, when including loudspeaker presentation and testing the system with severely aided HI subjects, the error rates increased strongly. The simulated influence on the measurement accuracy shows that a reliable measurement using the *FC-DNN-LM50* ASR system is no longer ensured.

Therefore, the *TDNN-LM5* ASR system was introduced. The error rates for the aided, severely HI can be reduced down to the range of NH and unaided HI listeners, with improvements on the acoustical model side as well as on the language model side. The simulations barely showed an influence on the SRT measurement accuracy with this improved system. Nevertheless, there was still a significant difference in SDR.

The higher SDR for both ASR systems with the aided, severely HI subjects is not fully explainable by the more challenging acoustic conditions - as the noisy versions of the clean data sets are still showing a significant difference in the SDR.

Even though there is a mismatch in the age of the speakers between the training data (on average 36 years old) and the aided, severely HI subjects (on average 72 years old), this seemed not to be a crucial factor, since the unaided HI subjects (on average 69 years old) have a similar age mismatch, just without this increase in error rates. Besides the 65 dB constant noise which was present during the responses of the aided, severely HI subjects, good SNR values could be achieved with the close-talker, cardioid microphone. The masker noise captured in free-field recordings is handled with data augmentation methods during the training of the acoustic model. This made additional noise suppression unnecessary, as barely any differences in SDR between the clean and noisy version of the NH and unaided HI subjects can be observed.

The remaining difference between the noisy data and aided, severely HI subjects might be introduced by pathological factors in the speech from the severely HI subjects or from a potential Lombard speech, as these subjects spoke in a noisy environment. This is in line with (Marxer et al. 2018; Uma Maheswari et al. 2020), where it was shown that even when accounting for the correct gain of the speech (which was done here by mixing at the correct SNR), this cannot cover all aspects of Lombard speech and there is still an increased error rate.

The SIR only showed small differences across different data sets. The

previous study (Ooster, Huber, et al. 2018) showed a generally stronger influence of deletion errors on the SRT measurement accuracy. Therefore, the *SDR* was prioritized in the parameter optimization of the *TDNN-LM5* systems language model. Generally, this prioritization to reduce deletion errors also helped to keep the *TDNN-LM5* system robust for the aided, severely HI data.
The responses from the CI subjects are recorded under clean conditions, which, unlike the other data sets, are not spontaneous speech. Even though, in a real application with CI subjects, the measurements need to be conducted with loudspeaker measurements - the small increase in error rates between the clean and noisy versions of the NH and unaided HI data indicate that a similar performance can be achieved under realistic acoustic conditions. Furthermore, (Ooster, Huber, et al. 2018) showed that in the context of the matrix sentence test, it is possible to estimate the ASR performance for spontaneous speech with read speech. Future research should investigate if this is also true for CI users.
Even though (Ooster, Huber, et al. 2018) showed a generally high robustness of the adaptive measurement procedure to errors from the ASR system, the *FC-DNN-LM50* ASR system cannot be transferred using the this specific ASR approach to open-set (list-based) speech-in-noise tests (e.g. (Kollmeier and Wesselkamp 1997; Nilsson et al. 1994; Van Wieringen and Wouters 2008)), since the overall vocabulary is highly increased in comparison to the matrix sentence tests. Nevertheless, the reduction to a sentence-specific decoding graph in the language model of the ASR system, presented in this study, suggests that it is possible to build an automated test procedure using any sentence-based listening test with an approach such as *TDNN-LM5*. For a new test it is only required to add the words to the lexicon and to change the grammar according to the respective target sentences. Test specific training data is not necessarily required.

3.5 Summary

This study explored a system using automatic speech recognition (ASR) for automated conduction of speech audiometry and listening tests with headphones or in free-field conditions. Our analysis included speech data

from listening tests conducted with NH listeners and HI listeners with different degrees of HL. The experiments were performed by using the recordings collected during speech audiometry from listeners, using these as input to ASR systems, and measuring the errors of the transcript. The dynamic measurement procedure was reproduced with Monte Carlo simulations that take into account the HL of subjects and the stochastic elements in responses. Two different ASR systems were considered: First, we used a hybrid recognizer that combines a simple feed-forward deep neural network with a hidden Markov model, which we used in a previous study. This system produced relatively accurate measurement results compared to the supervised system for NH and unaided HI listeners (average bias of -0.11 and -0.14 dB, respectively, and 95 % of listeners with a deviation of 1.32 dB or below). However, in noisy conditions as well as for aided, severely HI listeners and CI listeners, the ASR-based scoring error increased, resulting in a higher bias (0.29 dB) and 95 % confidence interval, exceeding 2 dB. Therefore, a second ASR system was trained that represents the current state-of-the-art, using a time delay neural network (TDNN) and combining it with a current cost function (latice-free maximum mutual information) and a language model tailored to the target sentence. This second system resulted in error rates that are relatively low (with the highest score deletion rate of 3.3%), and we estimate that the SRT deviation with this system is below 1.38 dB for 95% of the users, with an average bias of -0.16 dB or lower.

Acknowledgements

The authors would like to thank Dirk Eike Hoffner and Anita Georges for assisting with conducting the measurements, as well as Muriel Schaber for transcribing the data of the unaided hearing-impaired subjects.
This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 (Hearing4all)- Project ID 390895286 and the CRC TRR 31, Transfer Project T01.

CHAPTER 4

Speech audiometry at home: Automated listening tests via smart speakers with normal-hearing and hearing-impaired listeners

Abstract

Speech audiometry in noise based on sentence tests is an important diagnostic tool to assess listeners' speech recognition threshold (SRT), i.e., the signal-to-noise ratio corresponding to 50% intelligibility. The clinical standard measurement procedure requires a professional experimenter to record and evaluate the response (expert-conducted speech audiometry). The use of automatic speech recognition (ASR) enables self-conducted measurements with an easy-to-use speech-based interface. This paper

This chapter is a formatted reprint of
***Jasper Ooster**, Melanie Krueger, Jörg-Hendrik Bach, Kirsten C. Wagener, Birger Kollmeier, and Bernd T. Meyer,*
"Speech Audiometry at Home: Automated Listening Tests via Smart Speakers With Normal-Hearing and Hearing-Impaired Listeners,"
Trends in Hearing, vol. 24, Jan. 2020, `https://doi.org/10.1177/2331216520970011` .
Author contributions: JO developed and implemented the prototype of the smart speaker application and the measurements, prepared the figures, and wrote the manuscript. MK conducted the measurements. All authors co-wrote the manuscript (with the main contribution coming from JO).

compares self-conducted SRT measurements using smart speakers with expert-conducted lab measurements. With smart speakers, there is no control over the absolute presentation level, potential errors from the automated response logging, and room acoustics. We investigate the differences between highly controlled measurements in the lab and smart speaker based tests for young normal-hearing listeners as well as for elderly normal-hearing, mildly and moderately hearing-impaired listeners in low, medium, and highly reverberant room acoustics. For the smart speaker setup, we observe an overall bias in the SRT result that depends on the hearing loss. The bias ranges from +0.7 dB for elderly moderately hearing-impaired listeners to +2.2 dB for young normal-hearing listeners. The intra-subject standard deviation is close to the clinical standard deviation (0.57/0.69 dB for the young/elderly NH compared to 0.5 dB observed for clinical tests and 0.93/1.09 dB for the mild/moderate hearing-impaired listeners compared to 0.9 dB). For detecting a clinically elevated SRT, the speech-based test achieves an area under the curve (AUC) value of 0.95 and therefore seems promising for complementing clinical measurements.

4.1 Introduction

Speech intelligibility, especially in noisy conditions, is a crucial factor of successful social interaction and is often limited for HI listeners, which potentially reduces their quality of life. Early supply with hearing aids can ease this limitation (Arlinger 2003), but requires an early and reliable diagnosis of HL. However, hearing aid adoption rates are low, especially for mild to moderate HL, and typically 1-7 years pass between becoming aware of a HL and being provided with hearing aids (ANOVUM 2018). A reliable measurement tool with a high accuracy for quantifying speech intelligibility in noise is internationally available through matrix sentence tests (Kollmeier, Warzybok, et al. 2015). Due to the closed-vocabulary construction of this test with a vocabulary size of 50 words, it allows for an unsupervised measurement with a graphical user interface (so-called closed response format). Nevertheless, a graphical interface excludes subjects who cannot read, such as children, visually-impaired, and illiterate people. As an example, 12% of the population between 18

and 64 in Germany is functionally illiterate (Grotlüschen, Buddeberg, et al. 2018). Furthermore, finding the correct words in a 50-word matrix might increase the effort for the subject to conduct the measurement, which could increase the influence of cognitive skills on the result. For elderly subjects, it is often not possible to conduct the measurement with a graphical user interface in a reasonable time (Brand, Wittkop, et al. 2004; Brand and Wagener 2017).

To overcome these limitations, we explore systems based on human-machine interaction using a speech interface. A similar approach was used for the automatically conducting the Dutch LIST sentence test with the goal of quantifying intelligibility of patients with cochlear implants (Deprez et al. 2013). In our own previous work, we developed a speech-controlled automated matrix test (SAMT) that uses ASR for response logging. The system has been evaluated in a clinical setting with well-controlled acoustics settings in the lab (Ooster, Huber, et al. 2018): This system was built to be used in an sound-isolated hearing booth where it is possible to record clean audio from the subjects' responses. With an ASR system fine-tuned for the vocabulary of the matrix sentence test, very low error rates below 1% deletion errors and 3% insertion errors could be achieved using responses from 20 NH subjects and 7 mildly HI subjects. With such low error rates, the measurement reliability using ASR was not significantly different from the clinical test. Furthermore, detailed Monte-Carlo simulations of the measurement procedure and potential ASR errors showed that up to 7% deletion errors and 15% insertion are acceptable for an accurate measurement (i.e., in the range of the normal test-retest accuracy of the test when it is conducted by a human supervisor).

SAMT uses an ASR-based setup for clinical environments, but it has not been designed for use at home. One example of a speech-in-noise test which already has been successfully implemented for screening purposes via telephone or headphones is the digit triplet test (Smits, Merkus, et al. 2006; Vlaming et al. 2011; Melanie A. Zokoll et al. 2012; Smits, Theo Goverts, et al. 2013; De Sousa et al. 2020), which is also available as a smartphone-based measurement (Potgieter et al. 2016). The limitation to digits enables automated telephone testing since the subjects' responses can be logged using the keypad of the telephone. However, it

also limits the ecological validity of the test results as the words are not phonetically balanced and the linguistic variety is small.

Smart speakers, i.e., voice-controlled audio devices connected to a virtual assistant such as Amazon's *Echo*, Apple's *HomePod* or *Google Home* also have the potential of increasing the accessibility of speech intelligibility tests by performing self-measurements at home, since they provide a good audio quality and have a built-in dialogue management system including an ASR component. There have been several approaches to use smart home systems for medical purposes, e.g., to provide acoustic cues to support dementia patients' memory (Boumpa et al. 2019) or to support elderly people in their physical therapy (Vora et al. 2017). The Apple ResearchKit (Apple Inc. 2018) features a speech-in-noise test that is similar to our approach, but has not been compared to a standard audiological test in a clinical setting.

In this article, we present a smart speaker application for measuring the SRT, i.e., the SNR ratio corresponding to 50% intelligibility, with the matrix sentence test[1]. Due to the increasing availability of smart speakers, an accurate screening procedure for hearing deficits could potentially lower the threshold for conducting tests for a large number of users and therefore have a positive effect on early provision of assistive hearing devices. In a previous pilot study, the smart speaker-based measurement was evaluated in a single office room with six young, NH listeners, where a similar performance to clinical lab results was found (Ooster, Moreta, et al. 2019). However, the reliability for HI subjects was not part of the previous study although this aspect is crucial for speech audiometry. Furthermore, in a real use case, the acoustic conditions in which the test is conducted can exhibit large variability which could also influence reliability.

To explore automated at-home hearing screening for such use cases, the current study therefore compares clinical SRT measurements with SRTs obtained with a smart speaker application. Specifically, we analyze the errors and the resulting measurement reliability from the ASR system of a smart speaker in comparison to a calibrated clinical setup conducted by an expert; we also investigate potential decision thresholds for providing

[1] A short description and a link to the actual application can be found at `ca.uol.de/alexa-testmyhearing`

simple feedback to the user. These analyses are conducted for users with different degrees of HL (ranging from young normal hearing to elderly, moderately HI listeners), performing the test in different acoustic conditions with the aim of quantifying the interaction of test accuracy, user group and environment. Room acoustics are taken into account by simulating three different acoustic environments with different reverberation times which is realized through a room acoustic simulator.

4.2 Methods

This section presents an overview of the underlying principles of the matrix sentence test that are used in the smart speaker measurement as well as in the clinical reference measurement. We describe the implementation of this test as an application for the smart speaker and conclude with a description of the evaluation measurements as well as the data analysis performed on the measurement results.

4.2.1 Matrix sentence test

The speech audiometric test used in this study is the German matrix sentence test OLSA (short for its German name: OLdenburger SAtztest, English: Oldenburg sentence test) (Wagener, Kühnel, et al. 1999a; Wagener, Kühnel, et al. 1999b; Wagener, Brand, and Kollmeier 1999). The words of the stimulus sentences are randomly selected from a five-by-ten word matrix in order to create sentences with the structure *Name Verb Numeral Adjective Object*. The stimulus material is arranged in lists of sentences with the aim of providing phonetically balanced listening tasks with similar intelligibility. This design choice results in a low test-to-retest standard deviation of 0.9 dB for HI subjects (Wagener and Brand 2005) and 0.5 dB for NH subjects (Brand and Kollmeier 2002). The NH reference values are (-7.1 ± 1.1) dB with the male stimulus speaker (Wagener, Brand, and Kollmeier 1999) and (-9.4 ± 1.0) dB with the female stimulus speaker (Wagener, Hochmuth, et al. 2014). The influence of a HL on the SRT value measured with the matrix sentence test can be found in (Wardenga et al. 2015). The standardization of how to construct, record, and optimize the test across languages (Akeroyd et al. 2015) also yields a high comparability across different languages

(Kollmeier, Warzybok, et al. 2015; Melanie A Zokoll et al. 2013). During testing, the subject repeats the words he or she has recognized from the noisy sentence. Based on this response, the number of correctly recognized words N (referred to as sentence score) is calculated. Since the target value is the SRT, the SNR is dynamically adapted: It is increased for a word error rate below 50% and decreased otherwise (Brand and Kollmeier 2002). The final measurement outcome is estimated by a maximum likelihood fit to all the data points from the complete measurement list with twenty sentences where the underlying data distribution is given by a psychometric function, i,e. a logistic sigmoid.

4.2.2 The smart speaker application

An overview of the elements of the smart speaker application for the automated SRT measurement is shown in Figure 4.1 (adopted from (Ooster, Moreta, et al. 2019)). The setup differs from established clinical setups in several ways: (1) It uses synthesized speech instead of the original speech test recordings (which are protected by copyright), (2) when used at home, the sound is presented via the speaker in a reverberant environment, (3) the presented audio files are stored with lossy audio formats, and (4) the listener's response is transcribed via ASR and not logged by an audiometrist. The application was implemented with the Alexa Skill

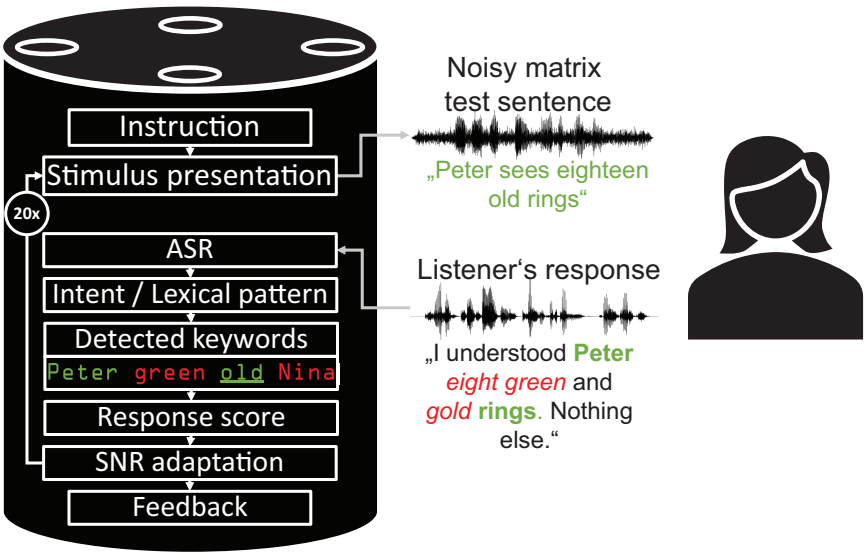


Figure 4.1: Overview of the smart speaker measurement application.

Developer Kit in Python (Amazon Inc. 2018, version 1.10.2) and executed on an Amazon Echo loudspeaker (2nd generation). When the measurement application is started, the listener hears an instruction about the general measurement procedure and the structure of the hearing test. These instructions are based on the guidelines for the clinical application of the matrix sentence test (HörTech 2019) . Since the subjects who participated in this study had no previous experience with smart speakers, they were told that the smart speaker is only listening when its optical indicator is active. During the measurement, the dialogue manager of the smart speaker uses so-called *intents*. These define the intended actions a user wants to take with their spoken command and are defined by lexical patterns within our application that are matched to the ASR transcript. These intents trigger the next action when they are detected by the ASR component. The core intent of the measurement application is the response to a matrix stimulus sentence. The lexical patterns to invoke this intent are based on real responses obtained in previous work (Ooster, Huber, et al. 2018), and the ASR engine of the smart speaker generalizes to variants of these responses. Based on this intent, the matrix test keywords in the subject's response are collected, and the SNR for the next presentation is adapted based on the resulting score.

We used a synthesized version of the sentences from the female German matrix sentence test, which was evaluated in a previous study (Nuesse et al. 2019). In that study, the same 150 sentences from the original female stimulus speaker (Wagener, Hochmuth, et al. 2014) were synthesized, and the commercial synthesis provided by the Acapela Group was found to produce the best results (in terms of naturalness as well as SRT results when compared to the original speaker). After checking the SNR of the smart speaker output by re-recording stimuli with known SNR and analyzing them, all synthesized sentences were premixed with the speech-shaped noise at steps of 0.1 dB and converted to the MP3 data format (MPEG version 2, 48 kbps, 16 kHz) as required for playback through the smart speaker. The noise signal was generated in the same way as for the original test by superimposing the synthesized speech material.

4.2.3 Evaluation measurements

The goal of the evaluation measurements is to investigate two major factors that could influence the measurement reliability of speech audiometry conducted with a smart speaker: The first is the degree of HL since the measurement application on the smart speaker should produce valid test results for all users. The second is the influence of different room acoustics that might considerably vary for at-home measurements.

4.2.3.1 Subject groups:

The listener groups differ by their degree of HL, and the listeners' age was also taken into account to explore age-related effects. In total, 4 subject groups were considered: Groups 1-3 covered age-matched elderly listeners categorized according to their PTA criterion from 0.5 to 4 kHz (Mathers et al. 2000): (1) NH (≤ 25 dB HL), (2) mildly HI ($26-40$ dB HL), and (3) moderately HI ($41-60$ dB HL). Group 4 consists of young NH listeners who satisfied a stricter definition of normal hearing, i.e., their HL did not exceed 15 dB at any frequency with one possible exception: a HL of 20 dB was allowed at one frequency from 250 Hz to 8 kHz. In total, 46 listeners participated in the study, with 9 to 16 subjects in each of the aforementioned groups (cf. Table 4.1).

	N (f/m)	Age [years]	PTA [dB HL]
Young normal-hearing max. one freq. at 20 dB HL	16 (12/4)	23 +/- 4	0 +/- 5
Normal-hearing PTA <= 25 dB HL	9 (5/4)	61 +/- 6	10 +/- 9
Mild hearing Loss PTA = 26 – 40 dB H	11 (5/6)	63 +/- 6	31 +/- 5
Moderate hearing loss PTA = 41 – 60 dB HL H	10 (3/7	62 +/- 10	46 +/- 6

Table 4.1: Statistics of the four subject groups who participated in the evaluation.

All subjects were paid for participating in this study. The audiogram of the better hearing ear for each of the subjects is shown in Figure 4.2. The HL was symmetrical (≤ 10 dB HL difference in the PTA) for 43 listeners. Three listeners had an asymmetric HL: One subject with a mild HL with 32.5 dB difference and two subjects with moderate HL with 12.5 dB and

15 dB difference, respectively. Three subjects in the elderly NH group showed HLs above 50 dB in the high frequencies, but still reached a PTA below 25 dB HL.

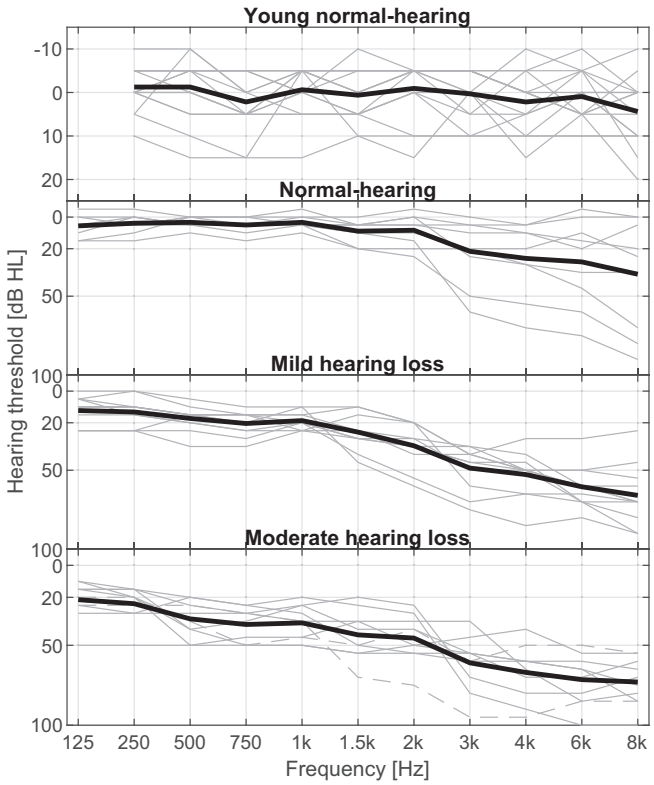


Figure 4.2: Individual audiograms of the better hearing ear of our subjects (gray lines) together with the average audiogram for the respective subject group (black lines). Note the different y-axis for the young NH listeners. The two dashed lines in the moderately HI panel describe the audiogram from two subjects which had to be discarded from data analysis as explained in the results section.

4.2.3.2 Test conditions:

The smart speaker measurements were conducted in a room referred to as Communication Acoustic Simulator (CAS) at the Hörzentrum Oldenburg. This room has a size of 12 m by 7 m by 2.8 m and uses 16 distributed microphones and 24 loudspeakers (12 flat wall-speakers and 12 two-way ceiling speakers) and a sound-regenerative Variable Room Acoustics System (VRAS) with a programmable microphone-to-speaker transfer function matrix to simulate different room acoustics. The sub-

jects were seated at the center of the CAS with the smart speaker in front of them at a distance of 2 m. Before the actual measurement, the subjects were asked to adjust the volume of the speaker to a comfortable level with high intelligibility of the speech assistant's voice; subjects were also allowed to change the volume of the speaker during the measurement.

To account for different acoustic conditions, rooms with different reverberation times T_{30} were simulated: *Living Room* (T_{30} = 0.51s), *Poor Classroom* (T_{30} = 1.12s) and *Concert Hall* (T_{30} = 1.52s). With the assumption of a spherical sound source, this results in critical distances of 0.88 m, 0.59 m and 0.51 m for the three simulated rooms, respectively. Hence, regardless of the simulated acoustic condition, the subjects were always in the far-field.

The clinical reference measurements were performed in a soundproof room at the Hörzentrum Oldenburg using speech signals from the original German female matrix sentence test speaker (Wagener, Hochmuth, et al. 2014). After D/A conversion (converter ADI-8 Pro by RME, Haimhausen, Germany), the speech and speech-shaped noise signals were amplified (HB7 by Tucker-Davis) and both presented to the subjects from the frontal direction via a loudspeaker (Mackie HR 824 by LOUD technologies). The distance between the subject and the loudspeaker was 1.4 m. The level of the speech-shaped noise was calibrated to 65 dB SPL in the absence of the listener using a measurement microphone (type 4189 by Brüel and Kjær) at the position of the listener and a sound level meter ("Modular Precision Sound Analyzer"; model 2260 by Brüel and Kjær). The speech test was performed using the Oldenburg Measurement Applications (HörTech; version 2.0.1.0). Responses of the speech intelligibility test were given orally by the listeners and marked by the (human) experimenter on a touch screen not visible to the listener.

Before conducting the main measurements, the audiogram was recorded with an audiometer (Aurical by Natus) with Sennheiser HDA200 headphones in a sound isolated booth using the ascending method.

4.2.3.3 Measurement procedure:

The subjects were invited for two measurement sessions each with nine SRT measurement lists with 20 presented matrix sentences, as described

in Table 4.2. The first two measurement lists from each session were used

Room settings A	Training list 1 Training list 2 Test list 1 Test list 2
Room settings B	Test list 3 Test list 4
Room settings C	Test list 5 Test list 6
Isolated booth	Reference

Table 4.2: Measurement sequence during one of the two sessions for each subject. While the reference measurement with the clinical setup was always performed at the end of each session, the order of the room characteristics of the CAS during the smart speaker measurement was randomly chosen for each subject, i.e. each setting (*Living Room* / *Poor Classroom* / *Concert Hall*) correspond to *A*, *B* or *C*.

as training due to the strong training effect of up to 2 dB in the first two measurement lists, which results from the limited vocabulary of the stimulus material of only 50 different words (Wagener, Brand, and Kollmeier 1999). To make training more efficient, the first ten sentences of the first list of each session were presented without additional noise, so that each subject heard and understood each possible word of the matrix test at least once before the adaptive procedure started. After the respective training, each subject conducted two measurement lists in each of the three room settings, resulting in twelve measurement lists in total with the smart speaker application, as well as two clinical reference measurement lists (one at the end of each of the two measurement sessions). Whenever the room acoustic settings were changed, the subjects heard four random sentences at different SNRs so they could adapt to the new room characteristics and could adjust the speaker volume if needed.
The young NH subjects only conducted one measurement session with four measurement lists (plus two training lists and the clinical reference) with the *Living Room* settings, since the results from the elderly subjects showed only a minor influence of the room acoustic settings on the measurement results.
At the end of each measurement session, all recorded audio files in the cloud of the smart speaker were deleted to avoid speaker adaptation of the ASR system. During the ASR-based measurements, a human su-

pervisor recorded the subjects' responses to obtain the ground truth of responses (assuming that the experienced human supervisor produced no errors when logging the reported words). This human transcript was later used to determine ASR errors as reported in the next section.

4.2.4 Data analysis

The data from the measurements are evaluated in three different ways: First — in order to evaluate the SRT measurement reliability of the smart speaker based measurement system — the SRT results from the measurements with the smart speaker ($SRT_{SmartSpeaker}$) are directly compared to the $SRT_{Reference}$ measured with the clinical setup at the end of the respective measurement session. Second, the ASR transcription errors of the smart speaker system are analyzed, and third, by an analysis of the collected data we find criteria for passing/failing the test.

4.2.4.1 SRT measurement accuracy:

The main measure of reliability in this study are the intra-subject standard deviation and the bias between $SRT_{SmartSpeaker}$ and $SRT_{Reference}$. The intra-subject standard deviation is obtained by calculating the standard deviation within each of the subjects and by averaging these standard deviations over the respective subject groups.

4.2.4.2 Performance of the ASR system:

The ASR transcription errors are estimated by comparing the transcript from the ASR system with the labels generated in parallel to the measurement by the human supervisor. Errors are quantified by two measures, the Score Insertion Rate (SIR) and the Score Deletion Rate (SDR), which only take into account the errors that could actually have an influence on the SRT scoring (either by inserting or deleting a matrix word). Out-of-vocabulary words are ignored with this metric (as they are in the clinical measurement). The SDR and the SIR that quantify the performance of the ASR system are defined by

$$SIR = \frac{N_{score\ insertions}}{N}, SDR = \frac{N_{score\ deletions}}{N}, \tag{4.1}$$

i.e., the number of errors $N_{score\ insertion}$ and $N_{score\ deletion}$, which are normalized by the number of correctly repeated matrix sentence test words in the subject's response N (ignoring non-matrix words). This metric is evaluated on the list level, i.e., using responses to 20 stimuli with an average of 50 correctly repeated matrix sentence test words. Note that the order of the words is neglected in this error metric, since the order of the words is also ignored during scoring in the clinical tests. The full error rates in the classical sense of an ASR system were not calculated since the full transcript (including words that are not relevant for the score) was not created in parallel to the measurements.

4.2.4.3 SRT decision threshold for providing user feedback:

To evaluate the performance of a decision threshold in terms of sensitivity and specificity, a potential value for a boundary is compared to three different reference decision criteria: **(a)** A deviation of the reference SRT measured with the clinical setup more than 1.96 standard deviations above the mean NH SRT, i.e., results outside the 95% percentile (which is the common approach for analyzing the result with the clinical setup) **(b)** a non-NH PTA ($>$ 25 dB HL) based on the WHO rules (Mathers et al. 2000) **(c)** an audiogram-based indication for a hearing aid, which is in Germany given by a HL of at least 30 dB in one of the audiogram frequencies between 500 Hz and 4 kHz (Gemeinsamer Bundesausschuss der Ärzte und Krankenkassen (G-BA) 2012). Additionally, the Youden-Index is derived which describes the ability of a decision threshold to separate the respective groups of data when sensitivity and specificity are equally weighted. To quantify how well the measure $SRT_{SmartSpeaker}$ is suited to determine one of the three reference criteria, we use the area under the curve (AUC) value, which describes the area under the receiver operating characteristic (ROC) curve which is obtained by plotting the sensitivity over (1 - specificity).

4.3 Results

4.3.1 SRT measurement accuracy

Figure 4.3 describes the SRT measurement accuracy with the smart speaker application compared to clinically-acquired reference estimates. This fig-

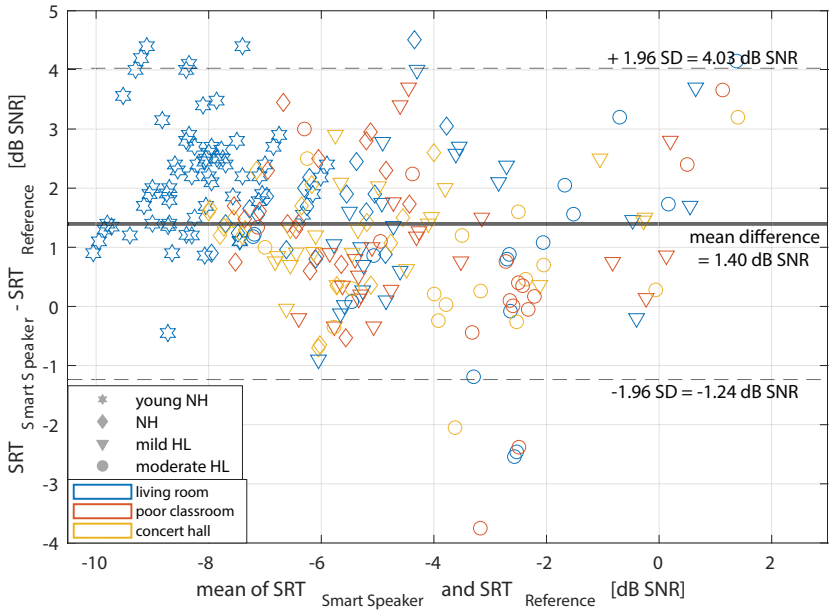


Figure 4.3: Bland-Altman plot for visualizing the agreement between automated and regular test conduction: The figure compares SRT differences between smart speaker and clinical measurement (measured in the same session) to the average of these two values. Data is shown for different subject groups (labeled with different shapes) and room characteristics (differentiated by color). Additionally, the average difference between the two measuring methods (grey solid line) and the 95% percentiles (grey dashed lines) are shown. The young-NH subjects only conducted experiments in the *living room* condition.

ure shows the difference in the SRT outcome obtained with the smart speaker application to the reference measurement against the *average* of these two values for all of the subjects. The data from two moderately HI subjects has been removed from the analysis since their spread of SRT results was exceptionally high (intra-subject standard deviation of 3.46 dB) and they reached SRTs above 10 dB. For the first excluded subject, the ASR errors were very high ($\approx 25\%$ on average); the second excluded listener spoke very softly during the first session, which resulted in several terminations of the measurement application. The second measurement sessions for the two subjects were performed both normally (presumably due to a certain familiarization to the speech interface of the speaker), but were also excluded from further analysis. While the zero-line in Figure 4.3 indicates a perfect match between the clinically measured value

and the value estimated with the smart speaker application, most of the data points are above this line. This bias is highly significant (paired-sample t-test, $p < 10^{-3}$) and amounts to 1.40 dB on average (± 2.63 dB 95% confidence interval, solid- and dashed-grey lines in Figure 4.3), i.e., the SRT measured with the clinical reference setup is lower than for the smart speaker condition. The different acoustic conditions are spread across a wide range of SRT differences, which is also reflected by the bias that is mostly constant over the three acoustic conditions (cf. Figure 4.4), and the only significant difference was the 0.38 dB between the *poor classroom* and the *concert hall* (paired-sample t-test, Bonferroni-adjusted, $p_{living\ room/poor\ classroom} = 0.91$, $p_{living\ room/concert\ hall} = 0.15$, $p_{poor\ classroom/concert\ hall} = 0.02$). The data analyzed in this section is measured after presenting two training lists which compensates most of the training effect (Wagener, Brand, and Kollmeier 1999). We did not observe any additional significant training effect over the course of the measurement sessions nor interactions with the room characteristics of the CAS with a two-way analysis of variance (ANOVA) (Interaction: $F(10, 318) = 1.50$, $p = 0.14$, training: $F(5, 318) = 0.81$, $p = 0.54$).

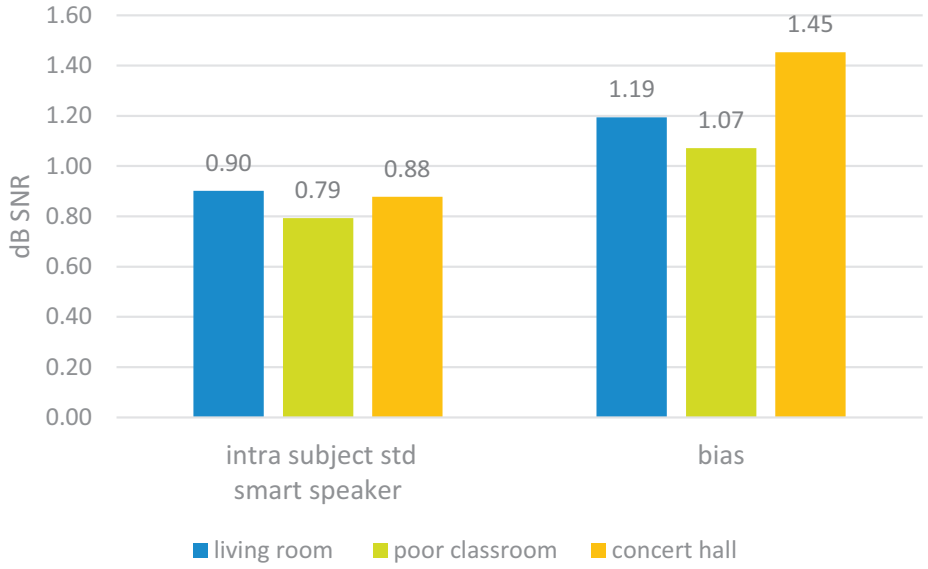


Figure 4.4: Bias and intra-subject standard deviations for all elderly, age-matched subjects for different room configurations; the bias relates to the difference between clinical and automated measurements. A positive bias refers to a lower (better) SRT in the clinical measurement.

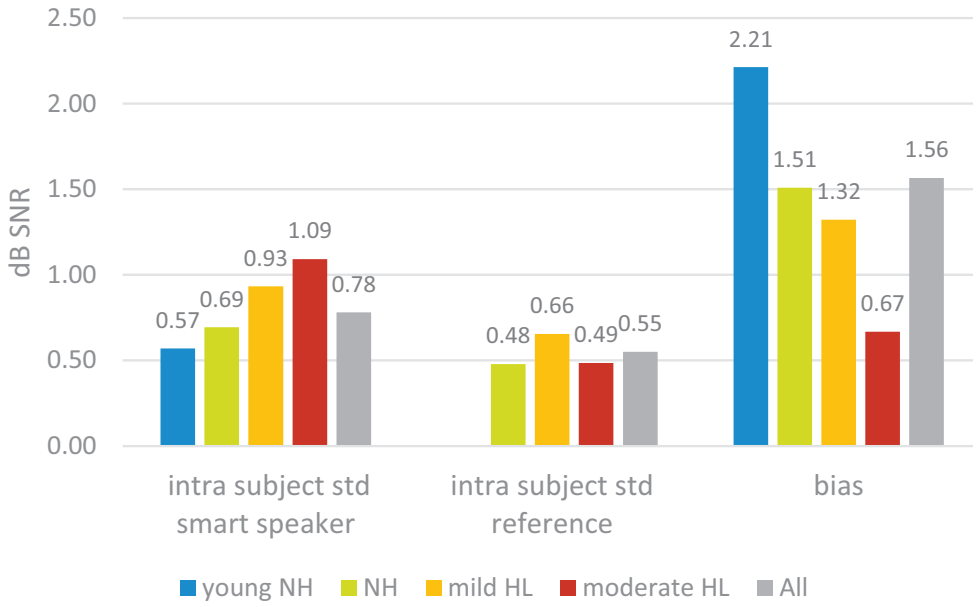


Figure 4.5: Bias and intra-subject standard deviations in the *living room*-settings for the different subject groups. Since the young NH subject group only conducted one measurement session with one reference measurement, the intra-subject standard deviation for this subject group in the reference setup cannot be estimated.

The differences between the subject groups are shown in Figure 4.5. The intra-subject standard deviation is increasing with increasing HL, reaching 1.09 dB for the elderly moderately HI listeners. On the other hand, the bias between the two measurement methods is decreasing for stronger HL from 2.21 dB for young NH listeners down to 0.67 dB for elderly moderately HI listeners. The bias from the young listeners is significantly higher than from the elderly subjects (two-sample t-test, $p < 10^{-6}$). Within the elderly listener group, the moderately HI listeners differ significantly from the other subgroups (two-sample t-test, Bonferroni-adjusted, $p_{NH/mild} = 0.78$, $p_{NH/moderate} < 0.01$, $p_{mild/moderate} = 0.02$).

4.3.2 Performance of the ASR system

The ASR performance of the smart speaker for all subjects is shown in Figure 4.6. The SIR is quite low with an overall average of 1.9% ($\pm$1.0% inter-subject standard-deviation) regardless of the subject group. We did not observe a significant difference between the elderly listener groups nor between the young and elderly listener groups (two-sample

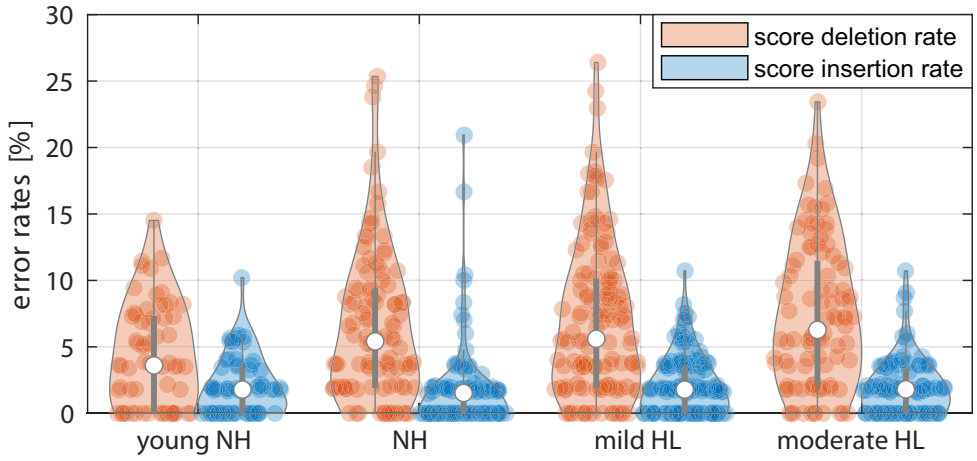


Figure 4.6: Violin plot of the ASR system's performance from the smart speaker. The individual data points denote error rates of single measurement lists, each with 20 presented sentences; the width of the violin denotes the normalized histogram of the error rates. The median value and the interquartile range are denoted by white dots and the grey line, respectively.

t-test, Bonferroni-adjusted, $p_{NH/mild} = p_{NH/moderate} = p_{mild/moderate} = p_{young/elderly} = 1.0$). Among 424 single measurement lists, only two SIR outliers are observed in the group of elderly NH listeners (with an SIR above 11%). The SDR is higher than SIR (with an average of 6.1% ($\pm$3.4% inter-subject standard-deviation)), and larger SDR differences between the subjects were observed, which was positively correlated to the subjects' age: The score deletion errors of elderly subjects (average $SDR = (6.7 \pm 5.6)\%$) are significantly elevated compared to SDRs of young NH subjects ($SDR = (4.2 \pm 3.8)\%$) based on a two-sample t-test and with the assumption of different variances ($p < 10^{-5}$). We did not find any significant differences between the different levels of HL within elderly listeners (two-sample t-test, Bonferroni-adjusted, $p_{NH/mild} = 1.0$, $p_{NH/moderate} = 0.683$, $p_{mild/moderate} = 1.00$).

Further, no significant difference was found in the SIR for different acoustic scenarios (two-sample t-test, Bonferroni-adjusted, $p_{living\ room/poor\ classroom} = p_{living\ room/concert\ hall} = p_{poor\ classroom/concert\ hall} = 1.00$). The only significant room-related difference for SDR was observed between *concert hall* and *poor classroom* with a difference of +2.1% (two-sample t-test, Bonferroni-adjusted, $p_{living\ room/poor\ classroom} = 0.08$, $p_{living\ room/concert\ hall} = 1.0$, $p_{poor\ classroom/concert\ hall} = 0.01$).

4.3.3 SRT decision thresholds for user feedback

Figure 4.7 compares a potential decision threshold based on the $SRT_{SmartSpeaker}$ for the three reference criteria in terms of sensitivity, specificity and the Youden-index. The 95% percentile decision threshold of −5.2 dB SNR,

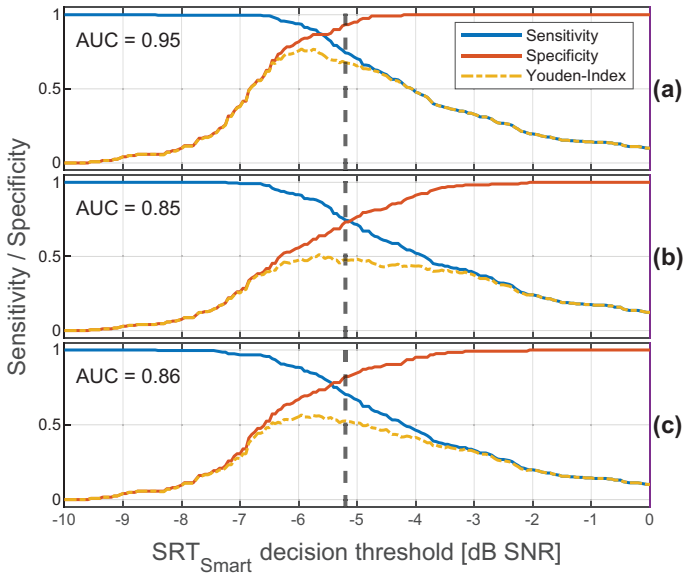


Figure 4.7: Sensitivity and specificity for analyzing how well a potential $SRT_{SmartSpeaker}$ decision threshold is suited for providing a binary screening decision. The curves shown here are derived from criteria that are used to quantifying HL:
(a) $SRT_{Reference} > -7.3$ dB SNR (the 95% percentile boundary from the young NH data measured with the reference setup);
(b) PTA > 25 dB HL.
(c) a HL of 30 dB or higher in at least one audiogram frequency between 500 Hz and 4 kHz (which is an indication for a hearing aid in Germany).
The dashed black line shows the 95% percentile boundary from the young NH data measured with the smart speaker.

directly calculated from the young NH subjects data measured with the smart speaker, is marked as the dashed black line. The maximum of the Youden-index, i.e., the statistical optimal decision boundary (when sensitivity and specificity are weighted equally) is always below this threshold ((a): −6.0 dB, (b): −5.7 dB and (c): −6.0 dB). At this threshold, the sensitivities for criteria (a) and (c) are higher than specificities (criterion (a): 0.93 vs. 0.74, criterion (c): 0.82 vs. 0.70). For criterion (b), sensitivity and specificity have similar values at the threshold (0.75 vs. 0.73).

4.4 Discussion

In this study, we investigated the SRT measurement reliability with a smart speaker-based application in three different acoustic conditions and with four different subject groups. First, we discuss the influence that errors made by the ASR system have on the measurement reliability. This is followed by the discussion of the overall SRT measurement reliability with the smart speaker application.

4.4.1 Effect of ASR errors

In our previous study that investigated automated tests in a clinical setting (Ooster, Huber, et al. 2018), we did not observe any significant decrease of the measurement reliability conditioned by the errors from the ASR system. Simulations regarding the overall influence of ASR errors on the SRT measurement accuracy in (Ooster, Huber, et al. 2018) matched well with the experimental findings. Even though these simulations are only based on NH SRT distributions, they are in principle valid for the smart speaker system explored in this study as well, since the tests are structurally identical and the same SNR adaptation scheme was used. On the one hand, the insertion errors are generally low (with an SIR of $(1.9 \pm 1.0)\%$) - and consequently it can be concluded that these errors barely influenced the measurement accuracy. On the other hand, the SDR is relatively high $(6.1 \pm 3.4)\%$ which results in an elevated intra-subject standard deviation and an SRT bias (Figure 4.5), which is in line with the simulations from ((Ooster, Huber, et al. 2018).
Even though extreme settings in the CAS were selected to evaluate the influence of different room acoustics, we did not observe strong differences in terms of ASR errors for different acoustic conditions (error rates for *concert hall* and *poor classroom* was the only significant difference in terms of SDR). We assume that the simulated acoustics resulted in a reduced diffuseness compared to real room acoustics, which could result in an underestimation of SRTs in that environment (i.e., SRT could appear as too good), although the simulated rooms appear to be very convincing.
For the best-performing group of young, NH listeners, average error rates up to 4% are obtained, which is quite high given the simple vocabulary.

This can presumably be attributed to the missing fine tuning of the Alexa-based ASR back-end and the overall challenging acoustic conditions with a distance of two meters between subject and smart speaker. Compared to these younger subjects, the error rates for elderly listeners are significantly elevated (Figure 4.6). However, the degree of hearing-impairment does not affect the ASR component, i.e., there is no significant difference in the ASR errors between the different groups of hearing-impairment within the elderly listeners in our data. We speculate the experience in human-machine interaction to be the underlying reason for the contrast between young and old participants and for the generally elevated intra-subject standard deviations: For instance, longer pauses in responses triggered the end-point detection of the smart speaker, and the remainder of the response will be ignored, which in turn could result in an undesired SNR increase for the next test item. Experienced users of speech-based interfaces would probably adapt to the machine listener (as we already observed to some extent during the measurements of this study), and therefore our scoring results could represent a conservative estimate with respect to errors through user behavior. Long-term measurements with the same group of listeners should be conducted to test this hypothesis. Furthermore, we observed a few terminations of the measurement application when the next action was not triggered although the subject responded with words from the matrix sentence test, since the ASR system from the smart speaker did not recognize any of the words in the subject's response correctly as matrix sentence words. This happened in about nine measurement lists from 6 different elderly listeners (one NH, four with mild HL, one with moderate HL) among the overall 424 conducted measurement lists with the smart speaker. In that case, the subjects had to restart the measurement from the beginning.
Major companies that develop and sell smart speakers use distributed data processing, i.e., the ASR is performed in the cloud, in this case on servers from Amazon. These companies regularly update the ASR back-end, over which developers and end users have no control. In theory, such an update could have a huge negative impact on the recognition of the matrix words, and the measurement accuracy described with the current setting could not be reached. However, changes to the back-end are generally driven by the desire to increase the robustness and reliabil-

ity of the system; it therefore seems unlikely to us that changes to the back-end would reduce recognition performance; it appears to be more likely that future optimizations will improve ASR and therefore reduce the test bias and its standard deviation. Similarly, the smart speaker's synthesized voice is used to describe the testing procedure to the listener, and developers also have no control over this component of the speech interface. For the actual test conduction described in this paper, this does however not play a role since we used synthesized clean and noisy speech samples that were uploaded as part of the application, so this part remains unaltered even when the server-side synthesis is changed.

4.4.2 Bias and reliability of SRT measurements

The young NH subjects average $SRT_{Reference}$ of (-9.3 ± 1.0) dB using a female natural voice matches well the results from the literature with the female speaker, which was reported to be (-9.4 ± 1.0) dB (Wagener, Hochmuth, et al. 2014) and -9.1 dB (Nuesse et al. 2019). When using synthetic speech in a clinical setup, a mean SRT of -8.5 dB was obtained for young NH subjects in a related study (Nuesse et al. 2019). This is on average 1.4 dB lower than the (-7.1 ± 1.0) dB SRT for the young NH subjects measured with the smart speaker using synthetic speech, i.e., a bias exists between the two measurement methods (clinical vs. automated) even when using the same stimulus material. This bias is varying with the different levels of HL from 2.2 dB (young NH) down to 0.7 dB (elderly moderate HL) and therefore this the main limitation to obtain accurate SRT results with the smart speaker. Nevertheless, the SRT results itself are reliable since the intra-subject standard deviation is independent from the simulated room acoustics and in the same range as with the clinical setup: The young NH intra-subject standard deviation with the smart speaker of 0.57 dB matches well with findings from other studies (0.5 dB in (Brand and Kollmeier 2002; Ooster, Huber, et al. 2018)). Compared to this, the elderly NH listeners have a slightly elevated intra-subject standard deviation (0.69 dB) when using the smart speaker. The mild/moderate HI listeners intra-subject standard deviation of 0.93 dB/1.09 dB is slightly higher than the 0.9 dB found with 10 HI subjects (Wagener and Brand 2005). The estimated intra-subject standard deviation values with the reference setup are smaller compared

to the smart speaker measurement, but they only rely on two values from each listener before averaging over all respective listeners so these values might not be very reliable.
For the smart speaker measurements, the presentation level is not calibrated to a specific level, and the individual noise and speech levels could not be controlled for at all (as explained in the methods section), which seems not to be crucial and is in line with previous studies: Wagener and colleagues (Wagener and Brand 2005) did not measure a significant influence of the presentation level on the SRT result (for levels that are clearly above the hearing threshold).
Note that all measurement reliability results are based on the measurement list after two training lists and there can be an training effect of up to 2 dB in the first two measurement lists (Wagener, Brand, and Kollmeier 1999). This training effect can be a drawback for a screening procedure, as it increases the required measurement time to reach the optimal result. However, training will always reduce the SRT, i.e., an SRT result below the decision threshold is already valid after the first measurement. If the result exceeds the threshold, the system will therefore recommend to repeat the measurement up two times, so test users will learn the matrix test vocabulary while they are performing the screening.

4.4.3 Deriving user recommendations from SRT values

The 95% percentile criterion derived from the young NH data is used to analyze the recommendation outcome of the smart speaker-based measurement. The corresponding threshold is compared to different reference criteria based on the reference SRT measured with the clinical setup or the audiogram (which is only indirectly related to the measured SRT).
Overall, the smart speaker measurement shows a good classification performance with an AUC of 0.85 for predicting a $PTA > 25\,\mathrm{dB\,HL}$ and an AUC value of 0.86 for predicting a hearing aid indication (cf. Figure 4.7 **(b)** and **(c)**). This is slightly higher than the observed prediction performance of the telephone-based German digit triplet test for which an AUC value of 0.82 was observed on 1903 listeners for predicting a $PTA > 25\,\mathrm{dB\,HL}$ and an AUC value of 0.76 for predicting a hearing aid indication (Gablenz et al. 2014). For the computer-based headphone

conduction of the English digit triplet test, an AUC value of 0.95 was found for 20 NH and 20 HI listeners (Folmer et al. 2017) for predicting a $PTA > 25\,\mathrm{dB\,HL}$, whereas (De Sousa et al. 2020) could increase the AUC from 0.78 with diotic stimuli in a smartphone-based headphone presentation to 0.94 with antiphasic stimuli on 145 listeners.
The prediction performance reported in this paper can also be compared to other speech audiometric tests: (Smits, Kapteyn, et al. 2004) found an AUC of 0.97 with the digit triplet test in comparison to the Plomp sentence test (Plomp and Mimpen 1979) with 38 subjects tested with headphones, and (Gablenz et al. 2014) measured an AUC of 0.70 when comparing the digit triplet test over telephone to the Göttinger sentence test (Kollmeier and Wesselkamp 1997) in a clinical setup. Therefore, the overall classification performance of the smart speaker-based measurement can be rated high with an AUC of 0.95 (cf. Figure 4.7 **(a)**) for detecting a clinical elevated SRT.

4.4.4 Limitations of this study

An important limitation of this study is that experiments were conducted in a lab environment with simulated room acoustics. Even though the small influence of the different simulated acoustics seems promising, future evaluations should take into account environments such as the private homes of the listeners with their respective real room acoustics for capturing the full variability of real-life scenarios. Furthermore, even though all the instructions for the test were given by the smart speaker application, a human supervisor was always present during the measurement which could have influenced the behavior of the listeners. The listeners who participated in this study where naive users of such a smart speaker. Experienced users potentially can profit from two training effects, i.e., adaptation of the speaker to its main user (which would decrease ASR errors) and secondly a training of the user, since long-term users of speech assistance presumably learn how to best interact with a virtual assistant (thereby avoiding incorrect end point detection). Finally, the threshold defined in this paper was obtained from measurements performed after training. For untrained listeners, the number of false positives should be higher compared to the presented values, and at the same time the number of false negatives should be lower. It is

therefore important to note that the decision boundaries derived from our experiments are valid for trained listeners only. In the smart speaker application, this is considered by recommending a repetition of the test (which effectively trains the users).

4.5 Conclusions

This paper introduced smart speaker-based speech audiometry with the matrix sentence test. We presented results for young and older listeners with different degrees of HL in three different simulated room acoustic conditions. For the different simulated room acoustics, we observe only small differences in the measured SRTs and the performance of the smart speaker's ASR system. The different listener groups showed a slightly decreased measurement reliability in terms of intra-subject standard deviation in comparison to results with the clinical version in the literature. The automatic speech recognition performance is significantly worse for elderly listeners, which appears to be the main source of the reduced reliability. However, the main limitation for obtaining accurate SRT results is a varying bias between the listener groups, which ranges from +0.7 dB for elderly moderately HI listeners up to +2.2 dB for young NH listeners. Nevertheless, the data presented in this paper supports the conclusion that smart speaker-based speech audiometry can reliably detect a deviating SRT in a self-guided manner at home since the ROC analysis showed an AUC of 0.95 for detecting a deviating clinical SRT, where the 95% percentile threshold based on the young NH data results in 93% sensitivity and 74% specificity. The smart speaker-based speech audiometric testing therefore seems promising for complementing clinical tests with the advantage of at-home screenings.

Acknowledgment

The authors would like to thank Inga Holube and Theresa Nüsse for providing the synthesized matrix test stimulus sentences, as well as Timmy Kröger and Dirk Eike Hoffner for their help with the conduction of the evaluation measurements.
Data from a pilot study (covering responses from 15 listeners compared to 46 listeners presented here) have been presented at the ISAAR sym-

posium 2019 in Nyborg (Ooster, Wagener, et al. 2020).
This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 (Hearing4all)- Project ID 390895286 and the CRC TRR 31, Transfer Project T01.

CHAPTER 5

General conclusions

This thesis explored the conduction of an autonomous conduction of a speech audiometric listening test based on ASR. The German matrix sentence test is used as a development prototype and two different application scenarios are explored: a locally running measurement in a well-controlled and calibrated laboratory environment on the one hand, and a screening test in an home environment with lots of uncontrolled parameters utilizing a smart speaker with a commercially ASR cloud-based back-end on the other. Overall, 139 subjects participated in the development and evaluation of the unsupervised measurement methods. The ASR performance is evaluated on 20.638 unique utterances from these subjects.

5.1 Summary of the Contributions of this Thesis

Chapter 2 introduced the concept of the self-measurement procedure utilizing ASR. A prototype of the unsupervised measurement system was used to record a database of spontaneous responses and to evaluate the measurement accuracy with the automated system in comparison to the conduction with an human supervisor. 20 NH and 7 mildly HI listeners participated in the measurements using the prototype system. This chapter introduced the error metric which is used to evaluate the performance of the ASR system. The error metric captures errors in the ASR system's transcript that are relevant for estimating the score correctly, i.e., if a target word from the stimulus was in the subject's response or

not. The order of the words as well as word repetitions are neglected. Therefore, the error metric consist of score deletion and score insertion errors. The error metric introduced differs from the WER, which is usually used in ASR, since estimating the full transcript is not necessary for the unsupervised conduction of speech intelligibility tests.
The ASR system's performance on the spontaneous responses was compared the the ASR performance on previously recorded read sentences. From this comparison it was concluded that in the context of the matrix sentence test it is possible to infer the performance with spontaneous speech from read speech.
Chapter 2 also addresses the question of how the errors from the ASR system interact with the SRT measurement accuracy. This is addressed using the two Monte Carlo simulation methods proposed: the first method uses the score insertion and score deletion error rate as free parameters and analyzes their influence on the SRT measurement outcome; the second method uses the decoding hypothesis from an actual ASR system on a specific data set to assess the SRT measurement accuracy. The findings of the Monte Carlo simulation are that deletion errors have a stronger influence on the measurement accuracy and that the adaptive measurement procedure is generally quite robust against errors from the ASR system.
The simulation methods were validated using the evaluation measurements with 20 NH and 7 mildly HI subjects participating in measurements using the prototype system. The test-to-retest standard deviation of 0.5 dB with the automated measurement system matched well with the 0.5 dB test-to-retest standard deviation found in the literature for NH listeners (Brand and Kollmeier 2002) and no significant difference was found between SRTs obtained by a human supervisor and SRTs measured by the ASR-based automated system.

Chapter 3 extended the evaluation of the unsupervised measurement system to a broad range of hearing losses. To this end two potentially major factors on the ASR system's performance were considered: firstly, disordered speech from severely HI subjects as well as CI subjects; secondly, noisy recordings which occur during aided measurements when a loudspeaker is used for stimulus presentation. To handle these more challeng-

ing conditions, an improved ASR system is proposed. This ASR system reduces the error rates in the noisy recordings of severely HI subjects and for disordered speech from CI subjects. The error rates are reduced down to the range of the NH and mildly/moderately HI subjects, which ensures a reliable measurement for these subject groups.
Besides a more robust acoustic model, the improved ASR system uses a language model which is specific to the respective target sentence, which is of course known in advance in any speech-based listening test. This implies that this system is able to serve as a direct development prototype for most of the sentence-based speech intelligibility tests, and the approach could also be applied when the target sentences are grammatically more complex (e.g. (Uslar et al. 2013)), since the actual grammar is not considered in the language model of the ASR system. The main parameter for the complexity of the language model for this method is just the number of words of the target sentence, which does not vary a lot between different sentence-based speech intelligibility tests. This hypothesis is supported by yet unpublished experiments performed during an external stay in the Hearing Systems group of the Technical University of Denmark, where work on a self-measurement procedure for the Danish hearing in noise test (HINT) (Nielsen and Dau 2011) was done. The HINT is a list-based sentence speech intelligibility test using 712 unique words overall. The number of unique words rises to 2380 when accounting for allowed deviations (e.g. wrong tense) in the subject's response. However, when accounting for only one sentence, the number of target words (five to ten words depending on the number of allowed deviations) differs only marginally from the five words per sentence in matrix tests. The acoustic model was trained with a 350 h publicly available Danish speech corpora (Danish Språkbanken compiled by Nordisk Språkteknologi (NST)) in the same way as the acoustic model in Chapter 2. The language model from the training data did not cover all of the target words. This was solved by manually adding phonetic transcriptions of the missing words to the lexicon of the ASR system. Using preliminary test data from 10 NH subjects, the proposed ASR structure with a sentence specific language model reached error rates similar to results shown in this thesis. These results indicate that the proposed system can be implemented for most of the available sentence-based speech

intelligibility tests.
This assumption has one limitation. Even though the proposed ASR system from Chapter 3 is transferable to other sentence-based speech intelligibility tests, it is not clear how different SNR adaption schemes and ways to calculate the SRT from the measurement interact with the errors from the ASR system, which should therefore be investigated in future research.

In Chapter 4 the unsupervised measurement was transferred to a screening application utilizing a smart speaker. This highly increases the accessibility of sentence-based speech intelligibility tests, since users can measure their hearing in noisy conditions at home themselves. Furthermore, once implemented for the smart speaker ecosystem, the application is also accessible via a virtual assistant application on every smartphone. Despite several uncertainties introduced due to the implementation on a smart speaker, it was possible to show that the measurement reliability in terms of intra-subject standard deviation is barely affected. The biggest drawback of the smart speaker-based self-conducted measurement is a varying bias between the different subject groups, which prohibits an accurate recreation of a clinically measured SRT. Even though it seems not to be indicated by ASR errors, the reason for the varying bias with the smart speaker-based measurement application is not clear and requires further investigation. Even though it seems not to be indicated by ASR errors, the reason for the varying bias with the smart speaker-based measurement application is not clear and requires further investigation. Furthermore, the screening application of the unsupervised measurement is only evaluated in the laboratory. Even though a broad range of reverberation times was tested (with rather extreme parameters, including a concert hall), the full variation of real living rooms, as well as all possible misuses from users, have not been evaluated yet. To evaluate the full classification performance, the application needs to be investigated with a representative cohort study. Despite these limitations, in the ROC analyses on a binary decision of a normal-hearing or elevated SRT a high decision accuracy was found with an AUC value of 0.95. This indicates that the smart speaker-based at-home measurement can be an accurate method for complementing clinical tests and diagnosis.

5.2 Relevance in the context of the digitalization of auditory healthcare

The increased accessibility of self-measurements is especially important in the context of providing healthcare to those who do not have the chance to visit professional facilities. It is estimated that 1.5 billion people worldwide are affected by hearing loss (World Health Organization 2021). 430 million of these people have a moderate or higher HL, of which the vast majority live in low- and middle-income countries (World Health Organization 2021). Due to the lack of trained professionals and high equipment cost, HI people in the low- and middle-income countries often do not have access to professional hearing care. However, already in the year 2011, 70% of the smartphones worldwide were used in low- and middle-income countries (World Health Organization 2011), so mobile health (mHealth) solutions have the potential to overcome the lack of professional healthcare (World Health Organization 2011; Källander et al. 2013). Since virtual speech-based assistants are available in many languages (Amazon's *Alexa*: 8 languages, Googles *Assistant*: 12 languages, Apples *Siri*: 21 languages), such a hearing test application for a virtual assistant can also provide hearing tests worldwide using a smartphone with headphones —removing the requirement for a smart speaker. Nevertheless, despite the high distribution of smartphones so far, most mHealth applications have not been very successful in low- and middle-income countries (Roess 2017). Furthermore, despite this potential in low- and middle-income countries, most virtual assistants are implemented in languages that are in demand in high-income countries. One reason for that is that for languages of small markets or markets with likely low volume of sales, there is often only very little annotated speech data available for training ASR systems. Generating large annotated training data sets is time-consuming and expensive, and therefore not very likely to be created for small and/or low-income countries. This issue is addressed by low- or zero-resource ASR approaches (T. Nguyen et al. 2021; NIST and IARPA 2020), which could be a remedy to this issue.

Besides the potential of providing hearing healthcare in low- and middle-

income countries, mHealth has gained increased importance in high-income countries. The goal is to provide a patient centered healthcare system (Alessia Paglialonga et al. 2019). Contactless and remote testing have furthermore become crucial during the current COVID-19 pandemic. In a survey of 120 British audiologists, 98% stated that, due to the restrictions enforced during the pandemic, they used remote testing methods (Saunders and Roughley 2020). Even though there are concerns regarding the quality of interpersonal interaction, 83.7% of the respondents are willing to continue the remote testing after the end of the pandemic restrictions and they expect only little impact on satisfaction and quality of auditory healthcare.
Another driving factor for self-measurement procedures is the increasing distribution of over-the-counter devices, e.g., self-fitting hearing aids without a medical license (Senate of the united states 2018; Sabin et al. 2020). A speech audiometric self-measurement as presented in this thesis could be used to verify a self-fitting process in a user's living room without the need to visit a specialist. This could also be driven by the ASR-based speech interfaces which have become ubiquitous today, and by the fast-growing distribution of smart speakers and virtual assistants on smartphones. Easily accessible hearing tests for these ASR-based speech interfaces could result in low thresholds for many listeners and provide screening methods, which hopefully will contribute to an increased awareness of hearing loss, as well as to an early supply of hearing support for those who need it.

5.3 Future research

The matrix sentence test is implemented with the same structure in over 20 languages. Even though it would be possible to train a new acoustic model for each language in the way of the ASR system proposed in this thesis, a multilingual implementation would be more versatile since it could be adapted to new, unseen languages (Hermann et al. 2021). To approach this goal, the output of the DNN acoustic model, i.e. phoneme posterior probabilities, could be used as features for a DTW recognizer (Hazen et al. 2009; Cetinkaya et al. 2016; Gundogdu and Saraçlar 2017). The referenced studies use this approach for either efficient query-by-

example tasks, i.e., searching keywords in an audio database, or to address the OOV problem in keyword detection. It could also be used with explicit knowledge about the target matrix words of the respective language. When directly using the acoustic model output, a language specific HMM is not required anymore. Therefore, a DTW-based approach could be combined with the approach in Vu et al. (2014) where an acoustic model is trained on multilingual phoneme set using the GlobalPhone database (Schultz 2002). Such a multilingual DTW recognizer has the potential to enable unsupervised measurements in all languages of the matrix sentence test without the need to train an acoustic model for each language separately. A drawback of DTW is that it usually uses one template per keyword and, hence, the variability of speech is not reflected by this template (to which the test item should be compared). Therefore, further research is required to quantify the advantageous and disadvantageous properties of the DTW approach.

Another question that should be addressed in the future is how the unsupervised system performs when speech audiometric tests are conducted with children (Wagener and Kollmeier 2005). Detecting a hearing loss in children is especially important, since an unaided hearing loss is related to communication and mental health disorders (Hogan et al. 2011), increases the probability to develop a speech disorder (Blamey et al. 2001) and even a small hearing loss can lead to more difficulties with educational and functional test measures (Bess et al. 1998). An early diagnosis is therefore crucial and and early fitting of a hearing aid can reduce the negative consequences (Tomblin et al. 2014). Frequently and easy accessible hearing screening can contribute to the early diagnosis and fitting. Even though there has been some progress in recognizing children speech, the higher variance in acoustic features and the lack of training data makes this a challenging task (Wu et al. 2019). It would be interesting to investigate the ASR performance with children's speech with the locally running ASR system for the laboratory application as well as with the smart speaker for the screening purpose.

Bibliography

Akeroyd, Michael A. et al. (May 2015). "International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests". In: *International Journal of Audiology* sup2, pp. 17–22. DOI: 10.3109/14992027.2015.1030513.

Amazon Inc. (2018). *Alexa Skills Kit SDK for Python*. https://github.com/alexa/alexa-skills-kit-sdk-for-python. accessed: 2019-06-25.

ANOVUM (2018). *EuroTrak Germany 2018 Survey*. Tech. rep. EHIMA. URL: https://www.ehima.com/wp-content/uploads/2018/06/EuroTrak_2018_GERMANY.pdf.

Apple Inc. (2018). *Researchkit - speech-in-noise test*. https://github.com/ResearchKit/ResearchKit. accessed: 2019-03-29.

Arlinger, Stig (2003). "Negative consequences of uncorrected hearing loss-a review". In: *International Journal of Audiology* 42, pp. 17–20. DOI: 10.3109/14992020309074639.

Baljić, I, A Winkler, T Schmidt, and I Holube (2016). "Untersuchungen zur perzeptiven Äquivalenz der Testlisten im Freiburger Einsilbertest". In: *HNO* 64.8, pp. 572–583. ISSN: 1433-0458. DOI: 10.1007/s00106-016-0192-0.

Barker, Jon, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe (2015). "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines". In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

Bess, Fred H, Jeanne Dodd-Murphy, and Robert A Parker (1998). "Children with Minimal Sensorineural Hearing Loss: Prevalence, Educational Performance, and Functional Status". In: *Ear and Hearing* 19.5. ISSN: 0196-0202. DOI: 10.1097/00003446-199810000-00001.

Blamey, Peter J., Julia Z. Sarant, Louise E. Paatsch, Johanna G. Barry, Catherine P. Bow, Roger J. Wales, Maree Wright, Colleen Psarros, Kylie Rattigan, and Rebecca Tooher (Apr. 2001). "Relationships Among Speech Perception, Production, Language, Hearing Loss, and Age in Children With

Impaired Hearing". In: *Journal of Speech, Language, and Hearing Research* 44.2, pp. 264–285. ISSN: 1092-4388. DOI: 10.1044/1092-4388(2001/022).

Boumpa, Eleni, Anargyros Gkogkidis, Ioanna Charalampou, Argyro Ntaliani, Athanasios Kakarountas, and Vasileios Kokkinos (2019). "An Acoustic-Based Smart Home System for People Suffering from Dementia". In: *Technologies* 7.1, p. 29. DOI: 10.3390/technologies7010029.

Brand, Thomas (2000). "Analysis and optimization of psychophysical procedures in audiology". PhD thesis. Carl von Ossietzky Universität Oldenburg.

Brand, Thomas and Birger Kollmeier (2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests". In: *The Journal of the Acoustical Society of America* 111.6, pp. 2801–2810. DOI: 10.1121/1.1479152.

Brand, Thomas and Kirsten Carola Wagener (2017). "Eigenschaften, Leistungen und Grenzen von Matrixtests (Characteristics, advantages, and limits of matrix tests)". In: *HNO* 65.3, pp. 182–188. DOI: 10.1007/s00106-016-0224-9.

Brand, Thomas, Thomas Wittkop, Kirsten Carola Wagener, and Birger Kollmeier (2004). "Vergleich von Oldenburger Satztest und Freiburger Wörtertest als geschlossene Versionen (Comparison of the Oldenburg sentence test and the Freiburg word test as closed versions)". In: *Proceedings of the 7th annual meeting of the Deutsche Gesellschaft für Audiologie (DGA), Leipzig.*

Bronkhorst, Adelbert W, Thomas Brand, and Kirsten Carola Wagener (2002). "Evaluation of context effects in sentence recognition". In: *The Journal of the Acoustical Society of America* 111.6, p. 2874. ISSN: 00014966. DOI: 10.1121/1.1458025.

Cetinkaya, Gozde, Batuhan Gundogdu, and Murat Saraçlar (2016). "Pre-filtered dynamic time warping for posteriorgram based keyword search". In: *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)* December 2016, pp. 376–382. DOI: 10.1109/SLT.2016.7846292.

Chen, Guoguo, Sanjeev Khudanpur, Daniel Povey, Jan Trmal, David Yarowsky, and Oguz Yilmaz (2013). "Quantifying the value of pronunciation lexicons for keyword search in low resource languages". In: pp. 8560–8564. ISBN: 9781479903566. DOI: 10.1109/ICASSP.2013.6639336.

Chen, Guoguo, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur (2015). "Pronunciation and Silence Probability Modeling for ASR". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 533–537.

Davis, Steven and Paul Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sen-

tences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4, pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

De Sousa, Karina C., De Wet Swanepoel, David R. Moore, Hermanus Carel Myburgh, and Cas Smits (2020). "Improving Sensitivity of the Digits-In-Noise Test Using Antiphasic Stimuli". In: *Ear and Hearing* 41.2, pp. 442–450. ISSN: 15384667. DOI: 10.1097/AUD.0000000000000775.

Deprez, Hanne, Emre Yilmaz, Stefan Lievens, and Hugo Van Hamme (2013). "Automating speech reception threshold measurements using automatic speech recognition". In: *Workshop of the special interest group on speech and language processing for assistive technologies*, pp. 1–6.

Farina, Angelo (2000). "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *Audio Engineering Society Convention 108*. Audio Engineering Society, pp. 1–24. DOI: 10.1109/ASPAA.1999.810884.

Feldmann, Harald (2004). "200 Jahre Hörprüfungen mit Sprache, 50 Jahre Deutsche Sprachaudiometrie - Ein Rückblick (200 years of hearing testing with speech, 50 years of German speech audiometry - a retrospective view)". In: *Laryngo- Rhino- Otologie* 83.11, pp. 735–742. ISSN: 09358943. DOI: 10.1055/s-2004-825717.

Folmer, Robert L., Jay Vachhani, Garnett P. McMillan, Charles Watson, Gary R. Kidd, and M. Patrick Feeney (2017). "Validation of a computer-administered version of the digits-in-noise test for hearing screening in the United States". In: *Journal of the American Academy of Audiology* 28.2. ISSN: 21573107. DOI: 10.3766/jaaa.16038.

Francart, Tom, Marc Moonen, and Jan Wouters (2009). "Automatic testing of speech recognition." In: *International Journal of Audiology* 48.2, pp. 80–90. DOI: 10.1080/14992020802400662.

Fraser, Norman M and G Nigel Gilbert (1991). "Simulating speech systems". In: *Computer Speech & Language* 5.1, pp. 81–99. DOI: 10.1016/0885-2308(91)90019-M.

Gablenz, Petra Von, Inga Holube, and Michael Buschermöhle (2014). "Was soll und kann ein Hörtest per Telefon erreichen?" In: *17. Jahrestagung der Deutschen Gesellschaft fÜr Audiologie*, pp. 1–4.

Gales, Mark JF (1998). "Maximum likelihood linear transformations for HMM-based speech recognition". In: *Computer speech & language* 12.2, pp. 75–98. DOI: 10.1006/csla.1998.0043.

Gemeinsamer Bundesausschuss der Ärzte und Krankenkassen (G-BA) (2012). *Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung (Guideline of the Fed-*

eral Joint Committee on the regulation of aids in public health care). URL: https://www.g-ba.de/richtlinien/13/.

Gopinath, Ramesh A (1998). "Maximum likelihood modeling with Gaussian distributions for classification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2. IEEE, pp. 661–664. DOI: 10.1109/ICASSP.1998.675351.

Grotlüschen, Anke, Klaus Buddeberg, Gregor Dutz, Lisanne Heilmann, and Christopher Stammer (2018). *Leben mit geringer Literalität LEO*. Tech. rep. URL: https://blogs.epb.uni-hamburg.de/leo.

Grotlüschen, Anke and Wibke Riekmann (2012). *Funktionaler Analphabetismus in Deutschland - Ergebnisse der ersten leo. – Level-One Studie*. 10th ed. Bundesverband Alphabetisierung und Grundbildung e.V, p. 298. ISBN: 978-3-8309-2775-4. URL: http://blogs.epb.uni-hamburg.de/leo/.

Gundogdu, Batuhan and Murat Saraçlar (2017). "Distance metric learning for posteriorgram based keyword search". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* March, pp. 5660–5664. DOI: 10.1109/ICASSP.2017.7953240. URL: http://ieeexplore.ieee.org/document/7953240/.

Hagerman, Björn (1982). "Sentences for testing speech intelligibility in noise". In: *Scandinavian Audiology* 11.2, pp. 79–87. DOI: 10.3109/01050398209076203.

Hahlbrock, Karl-Heinz (1953). "Über Sprachaudiometrie und neue Wörterteste". In: *Archiv für Ohren-, Nasen- und Kehlkopfheilkunde* 162.5, pp. 394–431. ISSN: 1434-4726. DOI: 10.1007/BF02105664.

Hazen, Timothy J., Wade Shen, and Christopher White (Dec. 2009). "Query-by-example spoken term detection using phonetic posteriorgram templates". In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 421–426. ISBN: 9781424454792. DOI: 10.1109/ASRU.2009.5372889.

Hermann, Enno, Herman Kamper, and Sharon Goldwater (2021). "Multilingual and unsupervised subword modeling for zero-resource languages". In: *Computer Speech & Language* 65, p. 101098. ISSN: 0885-2308. DOI: 10.1016/j.csl.2020.101098.

Hewitt, DR (2008). *Evaluation of an English speech-in-noise audiometry test*. Faculty of Engineering, Science and Mathematics Institute of Sound and Vibration Research, University of Southampton UK.

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE*

Signal Processing Magazine 29.6, pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

Hirsh, Ira J., Davis Hallowell, Silverman S Richard, Reynolds Elizabeth G., Eldert Elizabeth, and Benson Robert W. (Sept. 1952). "Development Of Materials For Speech Audiometry". In: *Journal of Speech and Hearing Disorders* 17.3, pp. 321–337. DOI: 10.1044/jshd.1703.321.

Hochmair-Desoyer, I, E Schulz, L Moser, and M Schmidt (Nov. 1997). "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users." In: *The American journal of otology* 18.6 Suppl, S83. ISSN: 0192-9763 (Print).

Hochmuth, Sabine, Thomas Brand, Melanie A Zokoll, Franz Zenker Castro, Nina Wardenga, and Birger Kollmeier (2012). "A Spanish matrix sentence test for assessing speech reception thresholds in noise". In: *International Journal of Audiology* 51.7, pp. 536–544. DOI: 10.3109/14992027.2012.670731.

Hogan, Anthony, Megan Shipley, Lyndall Strazdins, Alison Purcell, and Elise Baker (Aug. 2011). "Communication and behavioural disorders among children with hearing loss increases risk of mental health disorders". In: *Australian and New Zealand Journal of Public Health* 35.4, pp. 377–383. ISSN: 1326-0200. DOI: 10.1111/j.1753-6405.2011.00744.x.

HörTech (2019). *International matrix tests - Reliable speech audiometry in noise*. Oldenburg. URL: https://www.hoertech.de/images/hoertech/pdf/mp/produkte/intma/Broschre_Internationale_Tests_2019_WEB_klein.pdf.

Houben, Rolph, Jan Koopman, Heleen Luts, Kirsten Carola Wagener, Astrid Van Wieringen, Hans Verschuure, and Wouter A Dreschler (2014). "Development of a Dutch matrix sentence test to assess speech intelligibility in noise". In: *International Journal of Audiology* 53.10, pp. 760–763. DOI: 10.3109/14992027.2014.920111.

Hudgins, Clarence Virginius, JE Hawkins, JE Kaklin, and SS Stevens (1947). "The development of recorded auditory tests for measuring hearing loss for speech". In: *The Laryngoscope* 57.1, pp. 57–89.

Källander, Karin, James K Tibenderana, Onome J Akpogheneta, Daniel L Strachan, Zelee Hill, Augustinus H A ten Asbroek, Lesong Conteh, Betty R Kirkwood, and Sylvia R Meek (2013). "Mobile Health (mHealth) Approaches and Lessons for Increased Performance and Retention of Community Health Workers in Low- and Middle-Income Countries: A Review". In: *J Med Internet Res* 15.1, e17. ISSN: 1438-8871. DOI: 10.2196/jmir.2130.

Kinoshita, Keisuke, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al. (2016). "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research". In: *EURASIP Journal on Advances in Signal Processing* 2016.1, p. 7. DOI: 10.1186/s13634-016-0306-6.

Kliem, Kathrin and Birger Kollmeier (1994). "Entwicklung und Evaluation eines Zweisilber-Reimtestverfahrens für die deutsche Sprachaudiometrie". In: *Audiol Akust* 33.6, pp. 4–14.

Kollmeier, Birger, Anna Warzybok, Sabine Hochmuth, Melanie A Zokoll, Verena Uslar, Thomas Brand, and Kirsten Carola Wagener (2015). "The multilingual matrix test: Principles, applications, and comparison across languages: A review". In: *International Journal of Audiology* 54.sup2, pp. 3–16. DOI: 10.3109/14992027.2015.1020971.

Kollmeier, Birger and Matthias Wesselkamp (1997). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment". In: *Citation: The Journal of the Acoustical Society of America* 102, p. 2412. DOI: 10.1121/1.419624.

Leder, S B and J B Spitzer (June 1990). "A perceptual evaluation of the speech of adventitiously deaf adult males." eng. In: *Ear and Hearing* 11.3, pp. 169–175. ISSN: 0196-0202 (Print). DOI: 10.1097/00003446-199006000-00001.

Lincoln, Mike, Iain McCowan, Jithendra Vepa, and Hari Krishna Maganti (2005). "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments". In: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 357–362. DOI: 10.1109/ASRU.2005.1566470.

Littlejohn, Jenna, Annalena Venneri, Antonia Marsden, and Christopher J Plack (Mar. 2021). "Self-reported hearing difficulties are associated with loneliness, depression and cognitive dysfunction during the COVID-19 pandemic". In: *International Journal of Audiology*, pp. 1–5. ISSN: 1499-2027. DOI: 10.1080/14992027.2021.1894492.

Marxer, Ricard, Jon Barker, Najwa Alghamdi, and Steve Maddock (2018). "The impact of the Lombard effect on audio and visual speech recognition systems". In: *Speech Communication* 100.July 2017, pp. 58–68. ISSN: 01676393. DOI: 10.1016/j.specom.2018.04.006.

Mathers, Colin, Andrew Smith, and Marisol Concha (2000). "Global burden of hearing loss in the year 2000". In: *Global burden of Disease* 18.4, pp. 1–30.

Meyer, Bernd T., Birger Kollmeier, and Jasper Ooster (2015). "Autonomous measurement of speech intelligibility utilizing automatic speech recogni-

tion". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2982–2986. URL: https://www.isca-speech.org/archive/interspeech_2015/i15_2982.html.

Mohamed, Abdel-rahman, George E Dahl, and Geoffrey Hinton (2012). "Acoustic modeling using deep belief networks". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.1, pp. 14–22.

Mohri, Mehryar, Fernando Pereira, and Michael Riley (2008). "Speech Recognition with Weighted Finite-State Transducers". In: *Springer Handb. Speech Process. Speech Commun.* Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 559–584. ISBN: 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_28.

Moore, Meredith, Hemanth Venkateswara, and Sethuraman Panchanathan (2018). "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 466–470. DOI: 10.21437/Interspeech.2018-2391.

Mortensen, Linda, Antje S Meyer, and Glyn W Humphreys (Jan. 2006). "Age-related effects on speech production: A review". In: *Lang. Cogn. Process.* 21.1-3, pp. 238–290. ISSN: 0169-0965. DOI: 10.1080/01690960444000278.

Mühler, R, M Ziese, and D Rostalski (2009). "Development of a Speaker Discrimination Test for Cochlear Implant Users Based on the Oldenburg Logatome Corpus". In: *ORL* 71.1, pp. 14–20. ISSN: 0301-1569. DOI: 10.1159/000165170.

Nguyen, Tu Anh, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux (2021). "The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling". In: *Self-Supervised Learning for Speech and Audio Processing Workshop at NeurIPS*.

Nielsen, Jens Bo and Torsten Dau (2011). "The Danish hearing in noise test". In: *International Journal of Audiology* 50.3, pp. 202–208. ISSN: 1708-8186. DOI: 10.3109/14992027.2010.524254.

Nilsson, Michael, Sigfrid D Soli, and Jean A Sullivan (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." In: *The Journal of the Acoustical Society of America* 95.2, pp. 1085–99. ISSN: 0001-4966. DOI: 10.1121/1.408469.

NIST and IARPA (2020). *OpenASR Challenge*. URL: https://sat.nist.gov/openasr20.

Nogueira, Waldo, Filiep Vanpoucke, Philippe Dykmans, Leo De Raeve, Hugo Van Hamme, and Jan Roelens (2010). "Speech recognition technology in CI

rehabilitation." In: *Cochlear Implants Int.* 11 Suppl 1, pp. 449–453. ISSN: 14670100. DOI: 10.1179/146701010X12671177204507.

Nuesse, Theresa, Bianca Wiercinski, Thomas Brand, and Inga Holube (2019). "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech". In: *Trends in Hearing* 23, p. 233121651986298. ISSN: 2331-2165. DOI: 10.1177/2331216519862982.

Ooster, Jasper, Rainer Huber, Birger Kollmeier, and Bernd T. Meyer (Apr. 2018). "Evaluation of an automated speech-controlled listening test with spontaneous and read responses". In: *Speech Communication* 98, pp. 85–94. ISSN: 01676393. DOI: 10.1016/j.specom.2018.01.005.

Ooster, Jasper, Pia Nancy Porysek Moreta, Jörg-Hendrik Bach, Inga Holube, and Bernd T. Meyer (Sept. 2019). ""Computer, Test My Hearing": Accurate Speech Audiometry with Smart Speakers". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA: ISCA, pp. 4095–4099. DOI: 10.21437/Interspeech.2019-2118.

Ooster, Jasper, Kirsten Carola Wagener, Melanie Krueger, Jörg-Hendrik Bach, and Bernd T. Meyer (2020). "Potential of self-conducted speech audiometry with smart speakers". In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 7, pp. 373–380. URL: https://proceedings.isaar.eu/index.php/isaarproc/article/view/2019-43.

Paglialonga, A., G. Tognola, and F. Grandori (2011). "SUN-test (Speech Understanding in Noise): A Method for Hearing Disability Screening". In: *Audiology Research* 1.1, pp. 49–54. ISSN: 2039-4330. DOI: 10.4081/audiores.2011.e13.

Paglialonga, Alessia, Anisha A Patel, Erica Pinto, Dora Mugambi, and Karim Keshavjee (2019). "The Healthcare System Perspective in mHealth BT - m_Health Current and Future Applications". In: ed. by Giuseppe Andreoni, Paolo Perego, and Enrico Frumento. Cham: Springer International Publishing, pp. 127–142. ISBN: 978-3-030-02182-5. DOI: 10.1007/978-3-030-02182-5_9.

Parthasarathi, Sree Hari Krishnan and Nikko Strom (2019). "Lessons from building acoustic models with a million hours of speech". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6670–6674. DOI: 10.1109/ICASSP.2019.8683690.

Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur (2015). "A time delay neural network architecture for efficient modeling of long temporal contexts". In: *Proceedings of the Annual Conference of the Interna-*

tional Speech Communication Association, INTERSPEECH 2015-January, pp. 3214–3218. ISSN: 19909772.

Plomp, R and A M Mimpen (1979). "Improving the reliability of testing the speech reception threshold for sentences." In: *Audiology: official organ of the International Society of Audiology* 18.1, pp. 43–52. ISSN: 1499-2027. DOI: 10.3109/00206097909072618.

Potgieter, Jenni Marí, De Wet Swanepoel, Hermanus Carel Myburgh, Thomas Christopher Hopper, and Cas Smits (2016). "Development and validation of a smartphone-based digits-in-noise hearing test in South African English". In: *International Journal of Audiology* 55.7, pp. 405–411. ISSN: 17088186. DOI: 10.3109/14992027.2016.1172269.

Povey, Daniel, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohamadi, and Sanjeev Khudanpur (2018). "Semi-orthogonal low-rank matrix factorization for deep neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3743–3747. ISSN: 19909772. DOI: 10.21437/Interspeech.2018-1417.

Povey, Daniel, Arnab Ghoshal, et al. (2011). "The Kaldi speech recognition toolkit". In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, pp. 1–4.

Povey, Daniel, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur (2016). "Purely sequence-trained neural networks for ASR based on lattice-free MMI". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISSN: 19909772. DOI: 10.21437/Interspeech.2016-595.

Puglisi, Giuseppina Emma, Anna Warzybok, Sabine Hochmuth, Arianna Astolfi, N Prodi, C Visentin, and Birger Kollmeier (2014). "Construction and first evaluation of the Italian Matrix Sentence Test for the assessment of speech intelligibility in noise". In: *Proceedings of Forum Acusticum*. PAS-Polish Acoustical Society, pp. 1–5.

Rath, Shakti P, Daniel Povey, Karel Vesely, and Jan Cernocky (2013). "Improved feature processing for deep neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 109–113.

Roess, Amira (2017). "The promise, growth, and reality of mobile health-another data-free zone". In: *The NEW ENGLAND JOURNAL of MEDICINE* 377.21, pp. 2010–2011.

Ruff, S, T Bocklet, E Nöth, J Müller, E Hoster, and M Schuster (2017). "Speech Production Quality of Cochlear Implant Users with Respect to Duration and Onset of Hearing Loss". In: *ORL* 79.5, pp. 282–294. ISSN: 0301-1569. DOI: 10.1159/000479819.

Sabin, Andrew T, Dianne J Van Tasell, Bill Rabinowitz, and Sumitrajit Dhar (Jan. 2020). "Validation of a Self-Fitting Method for Over-the-Counter Hearing Aids". In: *Trends in Hearing* 24, p. 2331216519900589. ISSN: 2331-2165. DOI: 10.1177/2331216519900589.

Saon, George, Hagen Soltau, David Nahamoo, and Michael Picheny (2013). "Speaker adaptation of neural network acoustic models using i-vectors". In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 55–59. DOI: 10.1109/ASRU.2013.6707705.

Saunders, Gabrielle H. and Amber Roughley (2020). "Audiology in the time of COVID-19: practices and opinions of audiologists in the UK". In: *International Journal of Audiology* 60.4, pp. 255–262. ISSN: 17088186. DOI: 10.1080/14992027.2020.1814432.

Schmalz, Eduard (1846). *Erfahrungen über die Krankheiten des Gehöres und ihre Heilung*. Teubner.

Schröder, Marc and Jürgen Trouvain (2003). "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching". In: *International Journal of Speech Technology* 6, pp. 365–377.

Schuirmann, Donald J (1987). "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability". In: *Journal of Pharmacokinetics and Pharmacodynamics* 15.6, pp. 657–680.

Schultz, Tanja (2002). "Globalphone: a multilingual speech and text database developed at Karlsruhe university." In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. November. Denver, pp. 345–348.

Senate of the united states (2018). *A bill to provide for the regulation of over-the-counter hearing aids, 115th congress, session 1*. Accessed: 13/11/2018. URL: https://www.warren.senate.gov/files/documents/3_21_17_Hearing_Aids_Bill_Text.pdf.

Smits, Cas, Theo S Kapteyn, and Tammo Houtgast (2004). "Development and validation of an automatic speech-in-noise screening test by telephone". In: *International Journal of Audiology* 43.1, pp. 15–28. ISSN: 1499-2027. DOI: 10.1080/14992020400050004.

Smits, Cas, P. Merkus, Tammo Houtgast, Charles S. Watson, Gary R Kidd, James D Miller, Cas Smits, and Larry E Humes (2006). "How we do it:

The Dutch functional hearing screening tests by telephone and internet". In: *Clinical Otolaryngology* 31.5, pp. 436–440. ISSN: 0307-7772. DOI: 10.1111/j.1749-4486.2006.01195.x.

Smits, Cas, S. Theo Goverts, and Joost M. Festen (Mar. 2013). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise". In: *The Journal of the Acoustical Society of America* 133.3, pp. 1693–1706. ISSN: 0001-4966. DOI: 10.1121/1.4789933.

Snyder, David, Guoguo Chen, and Daniel Povey (2015). "MUSAN: A Music, Speech, and Noise Corpus". In: *CoRR* abs/1510.08484. arXiv: 1510.08484. URL: http://arxiv.org/abs/1510.08484.

Steffens, T (2016). "Verwendungshäufigkeit der Freiburger Einsilber in der Gegenwartssprache". In: *HNO* 64.8, pp. 549–556. ISSN: 1433-0458. DOI: 10.1007/s00106-016-0163-5.

— (Mar. 2017). "Die systematische Auswahl von sprachaudiometrischen Verfahren". In: *HNO* 65.3, pp. 219–227. ISSN: 0017-6192. DOI: 10.1007/s00106-016-0249-0. URL: http://link.springer.com/10.1007/s00106-016-0249-0.

Tomblin, J Bruce, Jacob J Oleson, Sophie E Ambrose, Elizabeth Walker, and Mary Pat Moeller (May 2014). "The Influence of Hearing Aids on the Speech and Language Development of Children With Hearing Loss". In: *JAMA Otolaryngology–Head & Neck Surgery* 140.5, pp. 403–409. ISSN: 2168-6181. DOI: 10.1001/jamaoto.2014.267.

Uma Maheswari, S., A. Shahina, and A. Nayeemulla Khan (2020). "Understanding Lombard speech: a review of compensation techniques towards improving speech based recognition systems". In: *Artificial Intelligence Review* 0123456789. ISSN: 15737462. DOI: 10.1007/s10462-020-09907-5.

Uslar, Verena N, Rebecca Carroll, Mirko Hanke, Cornelia Hamann, Esther Ruigendijk, Thomas Brand, and Birger Kollmeier (2013). "Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test". In: *The Journal of the Acoustical Society of America* 134.4, pp. 3039–3056.

Van Wieringen, Astrid and Jan Wouters (2008). "LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands". In: *International journal of audiology* 47.6, pp. 348–355.

Venail, F., E. Legris, B. Vaerenberg, J. L. Puel, P. J. Govaerts, and J. C. Ceccato (Apr. 2016). "Validation of the French-language version of the OTOSPEECH automated scoring software package for speech audiometry". In: *European Annals of Otorhinolaryngology, Head and Neck Diseases* 133.2,

pp. 101–106. ISSN: 18797296. DOI: 10.1016/j.anorl.2016.01.001. URL: http://dx.doi.org/10.1016/j.anorl.2016.01.001.

Vesely, Karel, Arnab Ghoshal, Lukás Burget, and Daniel Povey (2013). "Sequence-discriminative training of deep neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2345–2349.

Vipperla, Ravichander, Steve Renals, and Joe Frankel (2008). "Longitudinal study of ASR performance on ageing voices". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2550–2553. ISSN: 19909772.

Vlaming, Marcel S. M. G., Birger Kollmeier, Wouter A. Dreschler, Rainer Martin, Jan Wouters, Brian Grover, Yehya Mohammadh, and Tammo Mohammadh (Apr. 2011). *Hearcom: Hearing in the communication society.* DOI: 10.3813/AAA.918397.

von Helmholtz, Hermann (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Vieweg.

Vora, Jayneel, Sudeep Tanwar, Sudhanshu Tyagi, Neeraj Kumar, and Joel J.P.C. Rodrigues (Dec. 2017). "Home-based exercise system for patients using IoT enabled smart speaker". In: *IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom 2017.* Institute of Electrical and Electronics Engineers Inc., pp. 1–6. ISBN: 9781509067046. DOI: 10.1109/HealthCom.2017.8210826.

Vu, Ngoc Thang, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Herve Bourlard (May 2014). "Multilingual deep neural network based acoustic modeling for rapid language adaptation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Florence: IEEE, pp. 7639–7643. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6855086.

Wagener, Kirsten Carola and Thomas Brand (2005). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters". In: *International Journal of Audiology* 44.3, pp. 144–156. DOI: 10.1080/14992020500057517.

Wagener, Kirsten Carola, Thomas Brand, and Birger Kollmeier (1999). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (Development and evaluation of a German speech intelligibility test. Part III: Evaluation of the Oldenburg sentence test)". In: *Zeitschrift für Audiologie* 38.3.

Wagener, Kirsten Carola, Sabine Hochmuth, M. Ahrlich, Melanie A. Zokoll, and Birger Kollmeier (2014). "Der weibliche Oldenburger Satztest (The female Oldenburger sentence test)". In: *17. DGA Jahrestagung, Oldenburg*.

Wagener, Kirsten Carola and Birger Kollmeier (2005). "Evaluation des Oldenburger Satztests mit Kindern und Oldenburger Kinder-Satztest". In: *Z Audiol* 44.3, pp. 134–143.

Wagener, Kirsten Carola, Volker Kühnel, and Birger Kollmeier (1999a). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil I: Design des Oldenburger Satztests (Development and evaluation of a German speech intelligibility test. Part I: Design of the Oldenburg sentence test)". In: *Zeitschrift für Audiologie* 38.1.

— (1999b). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests (Development and evaluation of a German speech intelligibility test. Part II: Optimization of the Oldenburg sentence test)". In: *Zeitschrift für Audiologie* 38.2.

Waibel, Alexander, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang (1989). *Phoneme Recognition Using Time-Delay Neural Networks*. DOI: 10.1109/29.21701.

Wallenberg, E-L von and Birger Kollmeier (1989). "Sprachverständlichkeitsmessungen für die Audiologie mit einem Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten". In: *Audiol Akust* 28, pp. 50–65.

Wang, Dong, Xiaodong Wang, and Shaohe Lv (2019). *An Overview of End-to-End Automatic Speech Recognition*. DOI: 10.3390/sym11081018.

Wardenga, Nina, Cornelia Batsoulis, Kirsten Carola Wagener, Thomas Brand, Thomas Lenarz, and Hannes Maier (May 2015). "Do you hear the noise? The German matrix sentence test with a fixed noise level in subjects with normal hearing and hearing impairment". In: *International Journal of Audiology* 54.sup2, pp. 71–79. ISSN: 1499-2027. DOI: 10.3109/14992027.2015.1079929.

Warzybok, Anna, Thomas Brand, Kirsten Carola Wagener, and Birger Kollmeier (2015). "How much does language proficiency by non-native listeners influence speech audiometric tests in noise?" In: *International Journal of Audiology* 54, pp. 88–99. DOI: 10.3109/14992027.2015.1063715.

Wichmann, Felix A and N Jeremy Hill (2001). "The psychometric function: I. Fitting, sampling, and goodness of fit". In: *Percept. Psychophys.* 63.8, pp. 1293–1313. ISSN: 1532-5962. DOI: 10.3758/BF03194544.

Winkler, A and Inga Holube (2016). "Test-Retest-Reliabilität des Freiburger Einsilbertests". In: *HNO* 64.8, pp. 564–571. ISSN: 1433-0458. DOI: 10.1007/s00106-016-0166-2.

World Health Organization (2011). "mHealth: new horizons for health through mobile technologies." In: *mHealth: new horizons for health through mobile technologies.* ISSN: 9241564253.

— (2021). *World report on Hearing*, pp. 1–272. ISBN: 9789240020481.

Wu, Fei, Leibny Paola García-Perera, Daniel Povey, and Sanjeev Khudanpur (2019). "Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network." In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1–5. DOI: 10.21437/Interspeech.2019-2980.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig (2017). "Toward Human Parity in Conversational Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2410–2423. DOI: 10.1109/TASLP.2017.2756440.

Zokoll, Melanie A, Sabine Hochmuth, Anna Warzybok, Kirsten Carola Wagener, Michael Buschermöhle, and Birger Kollmeier (2013). "Speech-in-Noise Tests for Multilingual Hearing Screening and Diagnostics". In: *American Journal of Audiology* 22, pp. 175–178. DOI: 10.1044/1059-0889(2013/12-0061).

Zokoll, Melanie A., Kirsten Carola Wagener, Thomas Brand, Michael Buschermöhle, and Birger Kollmeier (2012). "Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype". In: *International Journal of Audiology.* ISSN: 1708-8186. DOI: 10.3109/14992027.2012.690078.

Danksagung

Bernd T. Meyer danke ich für die langjährige außergewöhnliche Betreuung, welche geprägt war von einem hohen Maß an Unterstützung, Vertrauen in meine Person und einer Vielzahl an Möglichkeiten, welche ich dank ihm erhalten habe. Zudem möchte ich mich für die unzähligen Anregungen seinerseits und sein Feedback bedanken, sowie die sehr positive und produktive Arbeitsatmosphäre, die zunächst in seiner Arbeitsgruppe „Automatic Speech and Audio Processing“ herrschte und seit zwei Jahren in der Arbeitsgruppe Kommunikationsakustik seiner eigenen Professur. Ich habe Bernd T. Meyer immer als meinen Mentor angesehen, welcher durch sein hohes Engagement in der Betreuung, diese Dissertation überhaupt erst ermöglicht hat.
Birger Kollmeier danke ich ebenfalls für die Unterstützung und Betreuung in den letzten Jahren und für das von ihm geprägte, anregende Arbeitsklima in der Medi, die von Ihm aufgebauten Strukturen die es mir ermöglicht haben mit dem Projekt überhaupt zu starten und einen großen Anteil an dem erfolgreichen Abschluss haben, als auch für außerordentliche Aktivitäten innerhalb der Medi, wie z.B. den Schreibworkshop auf Brač.
Odette Scharenborg danke ich, dass sie, als renommierte Wissenschaftlerin, das unabhängige Gutachten der Dissertation übernimmt.
Angel Castro Martínez und Constantin Spille danke ich für die Hilfe, die ich zu Beginn bekommen habe und für die Anstöße und Diskussionen.
Tom Brand danke ich für den fruchtbaren Input bei den Simulationen.
Dirk Eike Hoffner danke ich für die Hilfe bei zahlreichen Messungen mit Probanden.
Den Bürokollegen und -kolleginnen und Nachbarn und vor allem Tobias de Taillez, Feifei Xiong, Nils Westhausen, Jana Roßbach und Nadine El-Dajani danke ich für die angenehme Zusammenarbeit. Jana, Nadine und Connor Fitzgerald danke ich zudem für das Korrektur Lesen zum

Schluss.
Ein besonderer Dank geht auch an die TAs der Medi, an Anita Gorges, unter anderem für ihre Hilfe bei den Messungen mit Probanden, sowie an Frank Grunau für den technischen Support.
Nicht unerwähnt lassen möchte ich den viermonatigen Aufenthalt in Dänemark in der Hearing Systems Gruppe an der DTU in Kopenhagen im Rahmen meiner Promotion: Dies war eine inspirierende und fruchtbare Zeit für mich, welche aufgrund der COVID-19-Pandemie leider nicht mehr ausgewertet und somit in die Dissertation einfließen konnte, aber trotz allem für mich eine wichtige Erfahrung innerhalb der Promotion darstellte. Danke an Torsten Dau, Tobias May und Johannes Zaar für den herzlichen Empfang und die tolle Betreuung dort.
Abschließend möchte ich meiner Familie und Freunden danken. Meinen Eltern und Schwiegereltern, welche uns als junge Familie während des Studiums und während der Promotion immer als Babysitter oder ganz allgemein unterstützt haben. Insbesondere aber meiner Frau Clara für ihre Unterstützung und Beistand, sowie meinen Kindern Felix und Frieda für die Lebensfreude und den Ausgleich den sie mir Tag für Tag geben.

Eidesstattliche Erklärung

Hiermit erkläre ich, Jasper Ooster, dass ich die vorliegende Dissertation mit dem Titel „Automatic speech recognition interfaces for accurate self-conducted speech audiometry“ selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Ich bestätige, dass diese Dissertation weder in Teilen noch in ihrer Gesamtheit einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegen hat. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe.
Teile dieser Dissertation wurden bereits veröffentlicht bzw. sind zur Veröffentlichung eingereicht. Beiträge der Koautoren sind zu Beginn der jeweiligen Kapitel angegeben. Abgesehen davon bestand der Anteil der Koautoren an den Veröffentlichungen in der Betreuung und Korrektur der Manuskripte. Die Entwicklung der Methoden, die Konzeption, Durchführung und Auswertung der Experimente sowie das Schreiben der Manuskripte lagen in meiner Hand.

Oldenburg, den 24.06.21

(Jasper Ooster)