Computer Speech and Language 78 (2023) 101447

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Self-conducted speech audiometry using automatic speech recognition: Simulation results for listeners with hearing loss

Jasper Ooster [a,c,*], Laura Tuschen [b,c], Bernd T. Meyer [a,c]

[a] *Communication Acoustics, Carl von Ossietzky University, Oldenburg, 26129, Germany*
[b] *Fraunhofer-Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Oldenburg, 26129, Germany*
[c] *Cluster of Excellence Hearing4all, Germany*

ARTICLE INFO

Keywords:
Speech audiometry
Automatic speech recognition
Matrix sentence test
Unsupervised measurement

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 et al., 2018), as well as a state-of-the-art system utilizing a time-delay neural network (TDNN). The dynamic 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 73 listeners, ranging from normal-hearing to severely hearing-impaired as well as read speech from 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 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.

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 one 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 hearing loss or the performance of a hearing aid. However, an important drawback of such tests

[*] Corresponding author at: Communication Acoustics, Carl von Ossietzky University, Oldenburg, 26129, Germany.
E-mail address: jasper.ooster@uni-oldenburg.de (J. Ooster).

https://doi.org/10.1016/j.csl.2022.101447
Received 21 June 2021; Received in revised form 15 July 2022; Accepted 21 August 2022
Available online 27 August 2022

is the effort to conduct these in a clinical or laboratory test environment, since a human supervisor needs to be present during the measurement to log the responses of test subjects. While for percent correct test with fixed stimuli it is possible to record the responses and have them rated later, this is in particular important for adaptive speech tests that vary the difficulty depending on the listener responses. 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 et al., 2004, 2006; Vlaming et al., 2011; 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 strongly 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.

Borrie et al. (2019) analyzed a system for scoring responses of non-adaptive tests, i.e., the responses can be analyzed after the test conduction. That system requires a written transcript which is compared to the stimulus or reference text in order to calculate the score. It is stated that the automated system makes fever errors (<1%) than a human judge (>2%) in the scoring process.

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 $-0.2\,\mathrm{dB}$ (overestimating the hearing performance) 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 (GUI) for capturing the listeners' performance in a clinical context by providing all possible responses (Kollmeier et al., 2015). In contrast to the Digit Triplet Tests the matrix sentence tests are using complete, grammatically correct, sentences, with a careful selection of stimulus words to resemble the phoneme distribution of the German language, which increases the ecological validity. However, in clinical practice the GUI approach often remains unused since elderly subjects require too much time to find the correct responses in the 5 × 10 word matrix. Furthermore, it excludes visually or motorically impaired as well as illiterate listeners. For example, more than 12% of the adult German population is functionally illiterate (Grotlüschen et al., 2018).

The limited vocabulary of the matrix test 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 et al., 2018). This prototype features an ASR system 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, no influence on the SRT measurement accuracy was observed (with 0.9% deletion and 2.9% insertion errors on the scoring, respectively).

To evaluate the influence of ASR errors on the SRT measurement accuracy with the matrix sentence test, two simulations using Monte-Carlo methods were used. These simulations use a psychometric function to simulate the recognition probability of a stimulus word by the subject by a Bernoulli trial. The (simulated) recognized words are used to run the adaptive measurement procedure as in the original text conduction. The first method addressed the general influence of ASR errors on the simulated measurement procedure by taking the insertion and deletion errors as free parameters and falsely increasing or decreasing the number of recognized words based on these error rates. This simulation suggested a high robustness of the adaptive procedure against ASR errors with a stronger sensitivity to deletion errors. The second method utilized actual transcripts from an ASR system that uses real speech responses as input to estimate the performance for a specific target group. This second simulation scheme is also used in this study to evaluate SRT measurement accuracy with the here presented ASR systems and explained in more detail in the methods section.

While it seemed promising that no difference was found in measurement accuracy between measurements conducted with a human supervisor and unsupervised measurements utilizing an ASR system, all measurements in Ooster et al. (2018) were carried out using headphones for stimulus presentation. 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 hearing loss 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 in this study (ranging from NH over unaided mild & moderate hearing loss, aided severe hearing loss, to listeners with a CI). The ASR system from Ooster 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. However, since the baseline model in only trained on clean audio data, it was not expected to perform well under the above mentioned more difficult conditions. Therefore, the baseline system's performance is compared to an 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 et al., 2018), trained using data augmentation methods, which included training on noisy recordings. This system represents one of the most advanced hybrid systems available in the open source speech recognition toolkit *Kaldi* (Povey 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 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, i.e., training and test data need to have a similar syntax, 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. This garbage model is trained on typically phoneme transitions of the German language, i.e., it can handle filler words during the ASR systems decoding process without explicitly displaying them. Training data containing both filler words and matrix words is not required for this method.

The aim of this study is to quantify the reliability for ASR technology for different types of patients and noisy scenarios for automated speech audiometry.

To this end, we use a two-stage evaluation scheme; in a first stage the errors of the ASR systems are evaluated by comparing the ASR systems' output with labels transcribed by a human judge utilizing speech responses that were collected with the above-mentioned 73 subjects. In a second stage, we use a simulation scheme from our previous work (Ooster et al., 2018) to consider the dynamic nature of matrix tests (i.e., the SNR adaptation based on the number of correct/incorrect response words). This simulation directly uses the transcriptions from the ASR systems (including their errors) to evaluate the resulting SRT measurement accuracy. The results obtained with the two investigated ASR systems are compared against each other to address the question if a reliable measurement, i.e., with the same accuracy as with a human supervisor, would be possible when the adaptive measurement procedure is controlled by that ASR system.

2. Methods

2.1. Matrix sentence test

Speech intelligibility tests try to capture the capability of individuals to understand speech and are an important measure in hearing research as well as in clinical practice. The matrix sentence tests (Hagerman, 1982; Wagener et al., 1999b; Kollmeier 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. The speech audiometric test considered in this study is the German matrix sentence test. The stimulus material in this test is constructed from a five-by-ten matrix so that the sentences are syntactically fixed but semantically unpredictable (Wagener et al., 1999b). 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 et al., 1999a). 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.

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 deep neural network-hidden Markov model (DNN-HMM) hybrid systems (Hinton et al., 2012), which have a separate acoustic and language models. The first system is the exact ASR system which was used in Ooster 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 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 et al., 2011) and are based on publicly available training recipes.

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. These posterior probabilities 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) consisting of 16,000 utterances (18 h) from 40 different speakers. The self-recorded in-house database consists of 27,000 utterances (23 h) 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) were generated with a Gaussian mixture model (GMM)-HMM system trained on the clean audio on 25 ms frames with a 10 ms shift 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, DNN 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 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 artificially increases the number of training samples and 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. This is done to increase the general robustness of the acoustic model, rather than training it with the test noise condition. The resulting training set is six times larger compared to the original set and contains clean and noisy signals at different speaking rates.

The input to the DNN is for each frame (25 ms length, 10 ms shift) 40-dimensional MFCCs (without any speaker adaptation), which are grouped with 100-dimensional speaker identity vectors (i-vectors) (Saon et al., 2013). The acoustic classification is done with a factorized time-delay deep neural network (TDNN-F) and follows the implementation of Povey 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 et al., 2016). To prevent overfitting, four different regularization methods are used: l_2-norm regularization, 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 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.

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 during training to translate the word labels into phonetic labels. During testing, only the phonetic transcriptions of the German matrix sentence words are required. To handle 'out-of-vocabulary (OOV)' words observed during test time, i.e., words that are not part of the matrix test, 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 of the language model, which defines word transition probabilities. 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 five words of the stimulus sentence. These five words are more likely to be in the response of the subject than the other 45 words of the matrix test. This way, word substitution with the 45 other matrix test words cannot occur anymore, which would lead to score insertion errors. But substitutions of OOVs with the target words still can happen.

Furthermore, in the grammar of *TDNN-LM5*, the 20 most frequent OOV words from the NH subjects data in Ooster 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 are included such as *Rosen* (roses) instead of *Dosen* (cans) and *Boris* instead of *Doris*. This NH data was also used as a development set to fine-tune hyper parameters such as the word 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 et al., 2015). This also includes phoneme-specific probabilities for the silence class.

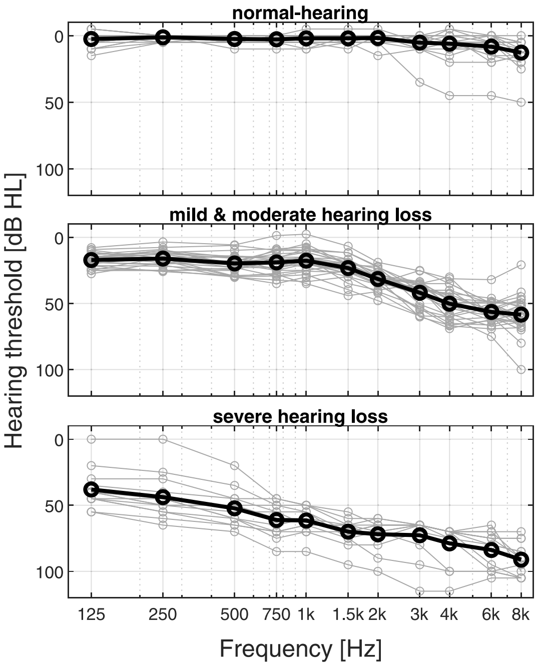

Fig. 1. Individual audiograms of the better-hearing ear of our subjects (gray lines) together with the average audiogram for the respective subject group (black lines).

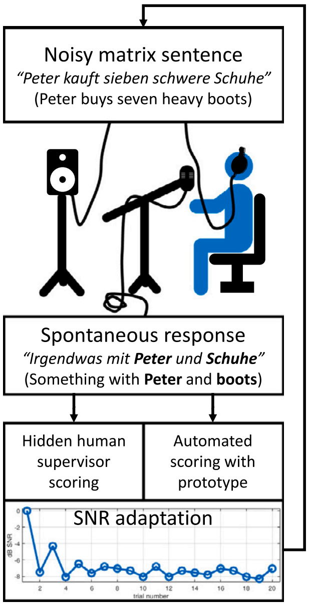
Noisy matrix sentence
"Peter kauft sieben schwere Schuhe"
(Peter buys seven heavy boots)
Spontaneous response
*"Irgendwas mit **Peter** und **Schuhe**"*
(Something with **Peter** and **boots**)
Hidden human supervisor scoring
Automated scoring with prototype
SNR adaptation
dB SNR
trial number

Fig. 2. To record spontaneous responses, matrix sentence test measurements are either conducted with a loudspeaker setup and a hidden human supervisor or in a headphone setup controlled by an ASR system, i.e., a prototype of the unsupervised measurement system.

2.3. Evaluation data

The recorded responses which are used for evaluation are collected with four different subject groups as described in Table 1, Fig. 1, and the following subsections. Overall 446 measurement lists 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 noisy versions of the audio, a total of 15k utterances are analyzed.

Fig. 2 shows an overview of the measurement and the recording of the spontaneous evaluation data. Two different approaches are used for audio recordings:

Table 1
Mean and standard deviation of the four subject groups who participated in the evaluation.

	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

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 *FC-DNN-LM50* 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 audio recordings of the subjects' responses were manually transcribed by a single judge after the measurements to generate human reference labels for the ASR system's evaluation. After the first round of labeling substantial effort has been put into cleaning the labels. Utterances were systematically rechecked on typos, random files have been chosen for relistening, and utterances that are extremely long or short have also been double checked. 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 and the same label cleaning methods are applied as for the NH listeners.

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. For all measurements the noise level was calibrated to a presentation level of 65 dB(A) for the noise stimulus. The noise level was kept fixed during the measurements and the stimulus speech levels were adopted based on the listeners response starting at 0 dB SNR.

The subjects with a CI did not record spontaneous responses, but read a list of 30 matrix sentences recorded in a quiet environment. All subjects were compensated for their participation in this study.

2.3.1. Normal-hearing listeners

This data set contains 2.420 spontaneous responses collected from 20 young NH listeners. Their responses were recorded using the *FC-DNN-LM50* ASR system and originally published in Ooster 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 on average six measurement lists with twenty sentences each (including two training lists) with the standard speech-shaped noise of the matrix test (so-called *olnoise*) 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.

2.3.2. Unaided hearing-impaired listeners

The unaided HI subjects had mild to moderate hearing losses ($25\,\mathrm{dB\,HL} < \mathrm{PTA}_{0.5,1,2,4} < 45\,\mathrm{dB\,HL}$). Eight of the 26 subjects used a hearing aid on a regular basis but did not wear it during the measurements. The data consist of overall 4660 spontaneous responses which were collected with the *FC-DNN-LM50* ASR 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 (better ear) 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, since the microphone does not record the stimulus signal. Therefore, 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.

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 hearing loss ($\mathrm{PTA}_{0.5,1,2,4} > 55\,\mathrm{dB\,HL}$) in both ears. They were all experienced users of their hearing aids (>10 years) and used their own hearing aids during the measurements. They were not tested for speech disorders related to hearing loss 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.

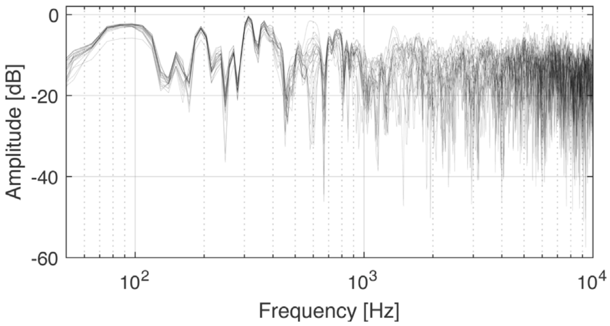

Fig. 3. 24 different frequency responses from the stimulus loudspeaker to the speech recording microphone. The frequency responses are measured with an exponential frequency sweep with 13 different subjects.

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 loudspeaker (Genelec 8030B) was placed at 1.5 m distance in front (0° azimuth) of the subjects. 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.

2.3.4. Noisy versions NH and unaided HI

A (simulated) noisy version of the clean data was generated to disentangle the differences between the acoustic condition and potentially 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. To this end the transfer function between the stimulus loudspeaker and the recording microphone was measured, the speech shaped noise of the matrix test convolved with this transfer function and than mixed to the clean data of the NH and unaided HI subjects. This procedure resembles the noise part of the severely hearing impaired listeners recordings, but does not cover changes in speaking style introduced by talking in a noisy environment (e.g. Lombard effect).

Transfer function:. 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 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). Fig. 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).

Signal-to-noise-ratio:. 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. Fig. 4 shows a box plot over the six measurements for each of the 13 severely HI subjects.

Creation of noisy audio files:. 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, the speech shaped noise of the matrix gets convolved with this transfer function and mixed at the SNR to the clean utterance to create a noisy version.

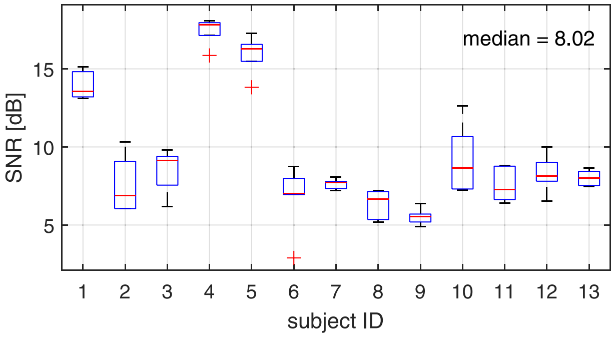

Fig. 4. SNR of the speaker's responses to the recorded background noise during loudspeaker measurements with hearing aids for each of the 13 subjects of the severely HI group. One SNR is calculated per measurement list, so there are six data points per subject in the box plot.

Table 2
Hypothetical example response and ASR transcript for showing the error metric. The (human) supervisor's task is to identify if the bold words from the stimulus sentence were in the listeners response. The gray words do not affect this (human supervisor's) scoring process and are also completely neglected by the error metric. <Filler> are non-matrix test words which are handled by a garbage model from the ASR system. The underlined words in the listener response and in the ASR transcript are a score deletion and a score insertion error, respectively. In this example the numbers are: $N_{subject\ score} = 4$, $N_{score\ insertions} = 1$, $N_{score\ deletions} = 1$, $SIR = SDR = 1/4$. Note that, due to the dynamic procedure of the test, $N_{subject\ score}$ cannot become zero for the complete list of sentences for which the metrics are calculated. Hence, a division by zero does not occur. Second, the order and repetition of words are neglected by these metrics as in human scoring.

Stimulus sentence	**Peter buys five wet stones.**
Listeners response (ground truth)	I got something like **Peter buys stones.** Ohh no there was also **wet** in it.
ASR transcript	<Filler> Nina **buys** <Filler> **Peter five** <Filler> five **stones**

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 system to train hearing and articulation of highly hearing-impaired persons. These are listeners who either have two CIs or have bimodal hearing devices, i.e., they have one hearing aid and one CI. As part of the data collection, a phoniatric and audiological examination was conducted by physicians for all subjects. Hearing disorders as well as speech and voice disorders were tested. 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.

2.4. System evaluation

The performance and accuracy of the whole unsupervised measurement system is evaluated in two stages: In the first stage the errors in the transcription of the ASR system are analyzed. In the second stage a Monte Carlo simulation method is used to evaluate how the ASR errors affect measurement accuracy with respect to the SRT. This simulation directly uses the output of the ASR system — including potential errors in the transcription (see Table 2).

2.4.1. ASR evaluation metrics

The ASR systems presented in this study differs from regular ASR use, since all words that do not occur in the matrix tests are ignored and the metrics introduced in this study are based on errors that could have an effect on the SRT, i.e., the final measurement value of the test. The ASR systems in this study are optimized for this particular task. This is in contrast to the usual task of an ASR system — to produce a full transcript, for which the performance would be quantified in terms of the word error rate (WER). Therefore, in this study two metrics are used which directly resembles the accuracy in the scoring process and the ASR systems' errors are quantified by the *score deletion rate* (SDR) and the *score insertion rate* (SIR):

$$SIR = \frac{N_{score\ insertions}}{N_{subject\ score}} \tag{1}$$

$$SDR = \frac{N_{score\ deletions}}{N_{subject\ score}} \tag{2}$$

$N_{subject\ score}$ is the ground truth number of correctly recognized and uttered words of the subject in response to the stimuli and $N_{score\ insertions}$ or $N_{score\ deletions}$ the number of falsely increased or decreased scores, where these referred to the errors made by the ASR system. These metrics are 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 $N_{subject\ score} \approx 50$.

2.4.2. *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 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 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 method a subject is simulated by defining a psychometric function, which gives the probability of understanding a word at a given SNR. This probability is used to perform a Bernoulli trial of recognizing a word or not of the stimulus sentence. Five of these trials (for the five words of a stimulus sentence) give the (simulated) number of correctly recognized words (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 transcribed 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. 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.

The parameters of the psychometric function, which defines the (simulated) subject, are randomly selected from the SRT distribution of the respective subject group to account for the higher variability in the (severe) HI subjects. 200 measurement lists are simulated with different psychometric functions. The number of independent elements per stimulus sentence (between 3.18 and 4.29) is also considered when performing the Bernoulli trials (Bronkhorst et al., 2002; Brand and Kollmeier, 2002).

3. Results

3.1. *Errors of the ASR system*

Fig. 5 shows the statistics of the ASR system's errors for each of the four subject groups. Table 3 summarizes the mean error rates. 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 (with significance level of $p < 0.05$).

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 listener groups, 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*. The *FC-DNN-LM50* results are used partially as a reference.

Table 3
Summary of the mean ASR system's error rates on the utterances from the four different subject groups.

	FC-DNN-LM50		TDNN-LM5	
	SDR [%]	SIR [%]	SDR [%]	SIR [%]
NH	0.76	2.55	0.27	2.34
NH noisy	3.19	2.59	0.52	2.27
Unaided HI	1.10	2.59	0.33	3.30
Unaided HI noisy	8.17	3.08	0.56	3.57
Severely HI aided	10.60	4.41	1.77	3.66
CI read	19.32	–	3.96	–

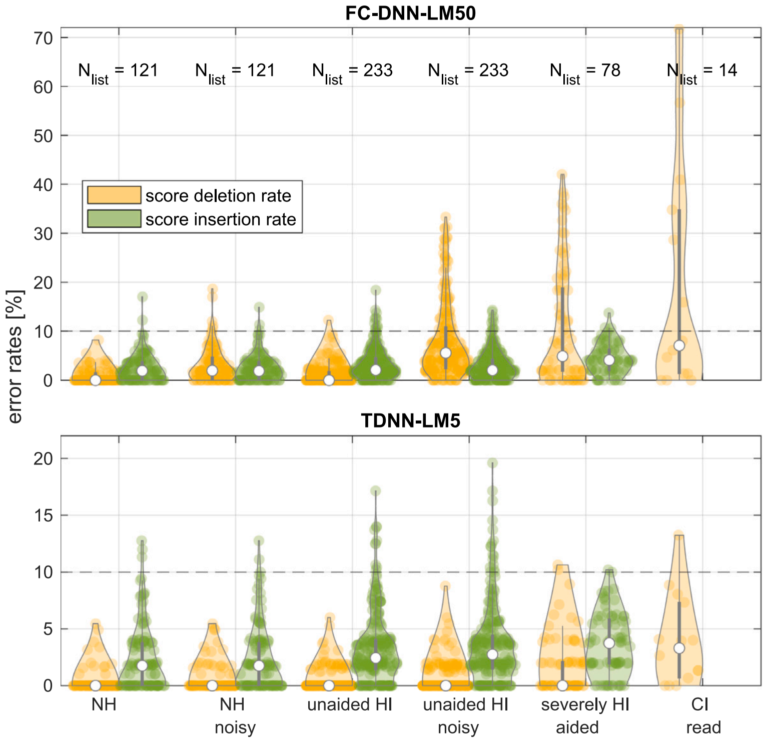

Fig. 5. Violin plot of the ASR system's error rates on the utterances from the four different subject groups. The individual data points denote error rates of N_{list} single measurement lists; the width of the violin denotes the estimated probability density of the error rates. The median value and the interquartile range are denoted by white dots and the bold gray lines, respectively. The whiskers (thin gray lines) have a maximum length of 1.5 times the interquartile range. All values above the whiskers are treated as outliers. The horizontal-dashed lines serves as orientation for better comparing the results of the upper and lower panel with the two different ASR systems.

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), $p < 0.002$).

3.1.2. 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

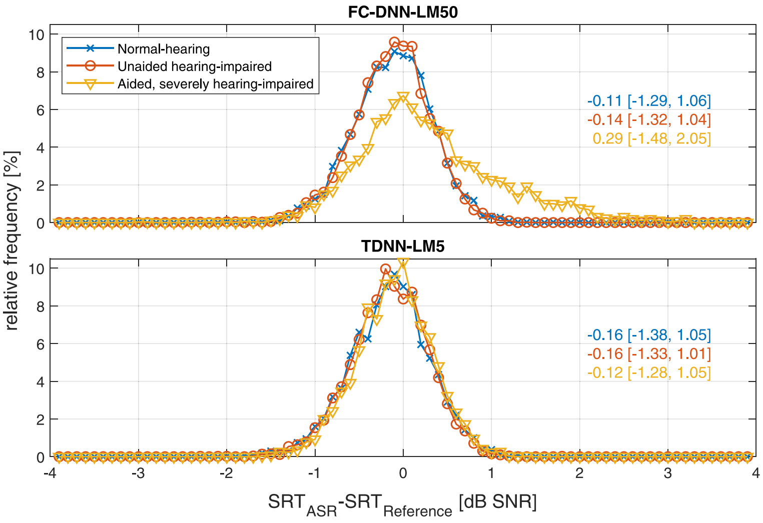

Fig. 6. Normalized histogram of the difference between simulated SRT outcomes with (SRT_{ASR}) and without ($SRT_{Reference}$) errors from the ASR system. For each point, the subjects' response behavior is simulated based on a psychometric function; and an audio file is selected based on the simulated 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, i.e., the thresholds where 2.5% or 97.5% of the data points are below this value.

($p < 0.01$ 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 ($p = 0.74$).

3.1.3. 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 Fig. 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 ($p = 0.10$).

3.2. Simulated influence on the SRT measurement accuracy

The results of the simulations are shown in Fig. 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 Fig. 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 Fig. 6, the error rates of the ASR system were lower with the *TDNN-LM5* system, hence all three subject groups show an overlap.

4. Discussion

This study explored the unsupervised conduction of speech intelligibility tests. To this end, the ASR performance of two different ASR systems was 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 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 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.

This system produced relatively accurate measurement results compared to the supervised system under good acoustic conditions 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), which is inline with the results from the previous study (Ooster et al., 2018). 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. This 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 persons 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 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. 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. Including Lombard speech as well as pathological speech to the training data of the ASR system could diminish the remaining difference in SDR.

The SIR only showed small differences across different data sets. The previous study (Ooster 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 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 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 this specific ASR approach to open-set (list-based) speech-in-noise tests (e.g. Kollmeier and Wesselkamp (1997), Nilsson et al. (1994) and 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.

With overall 0.6% and 3.0% score deletion and insertion errors, the *TDNN-LM5* ASR system could be even more accurate that a human supervisor during the test conduction. In Xiong et al. (2017) the human error rates in the labeling process is between 6% and 11% for large vocabulary continuous speech. In Borrie et al. (2019) more than 2% errors were found in the scoring process (after transcribing) for open-set hearing tests where different rules for allowed errors need to be applied. However, the labeling task for the matrix sentence test is considerably easier. It is only required to mark 50 words and no further rules need to be applied, since all deviation count as an error.

To verify this hypothesis, we performed an additional experiment where an additional person familiar with the matrix sentence test created labels for the unaided severely HI data which were later compared with the original transcripts. All 1560 sentences relabeled, which resulted in 3952 transcript words that were compared. 0.6% (22/3952) score deletion and 0.3% (11/3952) score insertion errors words were identified in comparison to the original labels. This compares to an average of 1.8% score deletion and 3.7% score insertion errors of the *TDNN-LM5* ASR system with the unaided, severely HI data.

The results shown in this study are conducted under (noisy) controlled acoustic conditions with an optimal microphone placement and the SRT results are based on a simulation scheme. Further 'real-world' measurements need to confirm the result of the simulation that results are as good as with human supervisors even when different kinds of microphones are used which might not be optimally placed.

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 hearing loss. 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 hearing loss of subjects and the stochastic elements in responses. While a baseline ASR system from a previous study (Ooster et al., 2018) was not able to produce accurate results for altered speech and noisy condition, a new proposed system that represents the current state-of-the-art, using a time delay neural network (TDNN) and combining it with a current cost function (lattice-free maximum mutual information) and a language model tailored to the target sentence, was able to accurately score recordings from all four subject groups. With overall 0.6% and 3.0% deletion and insertion errors in the automated scoring process, we estimate the SRT results to be within 1.38 dB for 95% of the users, with an average bias of -0.16 dB or lower. Therefore, we conclude that the proposed automated measurement procedure, using a state-of-the-art ASR system, can be used to accurately conduct speech audiometric tests, i.e., with the same accuracy as with a human supervisor.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

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, Germany under Germany's Excellence Strategy – EXC 2177/1 (Hearing4all)- Project ID 390895286 and the CRC TRR 31, Transfer Project T01.

References

Borrie, S.A., Barrett, T.S., Yoho, S.E., 2019. Autoscore: An open-source automated tool for scoring listener perception of speech. J. Acoust. Soc. Am. 145 (1), 392–399. http://dx.doi.org/10.1121/1.5087276.

Brand, T., Kollmeier, B., 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J. Acoust. Soc. Am. 111 (6), 2801–2810. http://dx.doi.org/10.1121/1.1479152.

Bronkhorst, A.W., Brand, T., Wagener, K.C., 2002. Evaluation of context effects in sentence recognition. J. Acoust. Soc. Am. 111 (6), 2874. http://dx.doi.org/10.1121/1.1458025.

Chen, G., Xu, H., Wu, M., Povey, D., Khudanpur, S., 2015. Pronunciation and silence probability modeling for ASR. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. pp. 533–537.

Deprez, H., Yilmaz, E., Lievens, S., Van Hamme, H., 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, A., 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. http://dx.doi.org/10.1109/ASPAA.1999.810884.

Francart, T., Moonen, M., Wouters, J., 2009. Automatic testing of speech recognition. Int. J. Audiol. 48 (2), 80–90. http://dx.doi.org/10.1080/14992020802400662.

Grotlüschen, A., Buddeberg, K., Dutz, G., Heilmann, L., Stammer, C., 2018. Leben mit Geringer Literalität LEO (Living with Low Literacy LEO). Technical Report, Universität Hamburg, URL: https://blogs.epb.uni-hamburg.de/leo.

Hagerman, B., 1982. Sentences for testing speech intelligibility in noise. Scand. Audiol. 11 (2), 79–87. http://dx.doi.org/10.3109/01050398209076203.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97. http://dx.doi.org/10.1109/MSP.2012.2205597.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M.A., Uslar, V., Brand, T., Wagener, K.C., 2015. The multilingual matrix test: Principles, applications, and comparison across languages: A review. Int. J. Audiol. 54 (sup2), 3–16. http://dx.doi.org/10.3109/14992027.2015.1020971.

Kollmeier, B., Wesselkamp, M., 1997. Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. Cit. J. Acoust. Soc. Am. 102, 2412. http://dx.doi.org/10.1121/1.419624.

Leder, S.B., Spitzer, J.B., 1990. A perceptual evaluation of the speech of adventitiously deaf adult males. Ear Hear. 11 (3), 169–175. http://dx.doi.org/10.1097/00003446-199006000-00001.

Marxer, R., Barker, J., Alghamdi, N., Maddock, S., 2018. The impact of the lombard effect on audio and visual speech recognition systems. Speech Commun. 100 (July 2017), 58–68. http://dx.doi.org/10.1016/j.specom.2018.04.006.

Mathers, C., Smith, A., Concha, M., 2000. Global burden of hearing loss in the year 2000. Glob. Burd. Dis. 18 (4), 1–30.

Meyer, B.T., Kollmeier, B., Ooster, J., 2015. Autonomous measurement of speech intelligibility utilizing automatic speech recognition. 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.

Mohri, M., Pereira, F., Riley, M., 2008. Speech recognition with weighted finite-state transducers. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (Eds.), Springer Handb. Speech Process. Speech Commun.. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 559—-584. http://dx.doi.org/10.1007/978-3-540-49127-9_28.

Moore, M., Venkateswara, H., Panchanathan, S., 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. http://dx.doi.org/10.21437/Interspeech.2018-2391.

Mortensen, L., Meyer, A.S., Humphreys, G.W., 2006. Age-related effects on speech production: A review. Lang. Cogn. Process. 21 (1–3), 238–290. http://dx.doi.org/10.1080/01690960444000278.

Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95 (2), 1085–1099. http://dx.doi.org/10.1121/1.408469.

Ooster, J., Huber, R., Kollmeier, B., Meyer, B.T., 2018. Evaluation of an automated speech-controlled listening test with spontaneous and read responses. Speech Commun. 98, 85–94. http://dx.doi.org/10.1016/j.specom.2018.01.005.

Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of the Annual Conference of the International Speech Communication Association. In: INTERSPEECH, Vol. 2015-January, pp. 3214–3218.

Potgieter, J.M., Swanepoel, D.W., Myburgh, H.C., Hopper, T.C., Smits, C., 2016. Development and validation of a smartphone-based digits-in-noise hearing test in South African english. Int. J. Audiol. 55 (7), 405–411. http://dx.doi.org/10.3109/14992027.2016.1172269.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., Khudanpur, S., 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. http://dx.doi.org/10.21437/Interspeech.2018-1417.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 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, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 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. http://dx.doi.org/10.21437/Interspeech.2016-595.

Ruff, S., Bocklet, T., Nöth, E., Müller, J., Hoster, E., Schuster, M., 2017. Speech production quality of cochlear implant users with respect to duration and onset of hearing loss. ORL 79 (5), 282–294. http://dx.doi.org/10.1159/000479819.

Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 55–59. http://dx.doi.org/10.1109/ASRU.2013.6707705.

Schröder, M., Trouvain, J., 2003. The german text-to-speech synthesis system MARY: a tool for research, development and teaching. Int. J. Speech Technol. 6, 365–377.

Smits, C., Kapteyn, T.S., Houtgast, T., 2004. Development and validation of an automatic speech-in-noise screening test by telephone. Int. J. Audiol. 43 (1), 15–28. http://dx.doi.org/10.1080/14992020400050004.

Smits, C., Merkus, P., Houtgast, T., Watson, C.S., Kidd, G.R., Miller, J.D., Smits, C., Humes, L.E., 2006. How we do it: The dutch functional hearing screening tests by telephone and internet. Clin. Otolaryngol. 31 (5), 436–440. http://dx.doi.org/10.1111/j.1749-4486.2006.01195.x.

Snyder, D., Chen, G., Povey, D., 2015. MUSAN: a music, speech, and noise corpus. CoRR abs/1510.08484. URL: http://arxiv.org/abs/1510.08484. arXiv:1510.08484.

Uma Maheswari, S., Shahina, A., Nayeemulla Khan, A., 2020. Understanding lombard speech: a review of compensation techniques towards improving speech based recognition systems. Artif. Intell. Rev. (0123456789), http://dx.doi.org/10.1007/s10462-020-09907-5.

Van Wieringen, A., Wouters, J., 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for flanders and the netherlands. Int. J. Audiol. 47 (6), 348–355.

Vesely, K., Ghoshal, A., Burget, L., Povey, D., 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, R., Renals, S., Frankel, J., 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.

Vlaming, M.S.M.G., Kollmeier, B., Dreschler, W.A., Martin, R., Wouters, J., Grover, B., Mohammadh, Y., Mohammadh, T., 2011. Hearcom: Hearing in the communication society. http://dx.doi.org/10.3813/AAA.918397.

Wagener, K.C., Brand, T., Kollmeier, B., 1999a. 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). Z. Audiol. 38 (3).

Wagener, K.C., Kühnel, V., Kollmeier, B., 1999b. 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). Z. Audiol. 38 (1).

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J., 1989. Phoneme recognition using time-delay neural networks. http://dx.doi.org/10.1109/29.21701.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M.L., Stolcke, A., Yu, D., Zweig, G., 2017. Toward human parity in conversational speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (12), 2410–2423. http://dx.doi.org/10.1109/TASLP.2017.2756440.

Zokoll, M.A., Hochmuth, S., Warzybok, A., Wagener, K.C., Buschermöhle, M., Kollmeier, B., 2013. Speech-in-noise tests for multilingual hearing screening and diagnostics. Am. J. Audiol. 22, 175–178. http://dx.doi.org/10.1044/1059-0889(2013/12-0061).