Potential of self-conducted speech audiometry with smart speakers

JASPER OOSTER[1,4,*], KIRSTEN C. WAGENER[2,4], MELANIE KRUEGER[3,4], JÖRG-HENDRIK BACH[2,3,4] AND BERND T. MEYER[1,3,4]

[1]*Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany*
[2]*Hörzentrum GmbH, Oldenburg, Germany*
[3]*HörTech gGmbH, Oldenburg, Germany*
[4]*Cluster of Excellence Hearing4all, Germany*

Speech audiometry in noise based on matrix sentence tests is an important diagnostic tool to assess the speech reception threshold (SRT) of a subject, i.e., the signal-to-noise ratio corresponding to 50% intelligibility. Although the matrix test format allows for self-conducted measurements by applying a visual, closed response format, these tests are mostly performed in open response format with an experimenter entering the correct/incorrect responses (expert-conducted). Using automatic speech recognition (ASR) enables self-conducted measurements without the need of visual presentation of the response alternatives. A combination of these self-conducted measurement procedures with signal presentation via smart speakers could be used to assess individual speech intelligibility in an individual listening environment. Therefore, this paper compares self-conducted SRT measurements using smart speakers with expert-conducted lab measurements. With smart speakers, the experimenter has no control over the absolute presentation level, mode of presentation (headphones vs. loudspeaker), potential errors from the automated response logging, and room acoustics. We present the differences between measurements in the lab and with a smart speaker for normal-hearing, mildly hearing-impaired and moderate hearing-impaired subjects in low, medium, and high reverberation.

INTRODUCTION

Being able to understand speech, especially in noisy conditions, is a crucial factor of social interaction and is often limited for hearing impaired listeners, which can reduce their quality of life. An early diagnosis of the hearing loss can ease this limitation by an early supply of a hearing aid (Arlinger, 2003). A reliable measurement tool with a high accuracy for quantifying the ability of speech understanding in noise is available through matrix sentence tests (Kollmeier *et al.*, 2015). Due to the closed-vocabulary construction of this test, it allows for an unsupervised measurement with a graphical user interface. Nevertheless, such an interface excludes subjects who cannot read (children, illiterate, visually impaired). Hence, we focus on a system that uses

*Corresponding author: jasper.ooster@uni-oldenburg.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

only acoustic communication cues, i.e., speech. This we propose to do with automatic speech recognition (ASR) for the response logging (Ooster *et al.*, 2018). While that system created for clinical (and relatively controlled) environments, an ASR-based conduction has also the potential of increasing the accessibility by performing self-measurements at home.

Smart speakers such as Amazon's *Echo*, Apple's *HomePod* or *Google Home* have the potential of bringing such a test to a broader subject base, since they provide a good audio quality and have a built-in dialogue manager including an ASR component. There have already been 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).

In this work, we present a smart speaker application for measuring and validating the speech reception threshold (SRT), i.e. the signal-to-noise ratio (SNR) corresponding to 50% intelligibility with the matrix sentence test. Smart speaker-based measurements have several differences compared to established clinical setups: (i) We use a high-quality speech synthesis instead of the natural speech files that are protected by copyright, (ii) the sound is presented via the speaker in a reverberant environment, (iii) compressed audio files are presented, and (iv) the listener's response is transcribed via ASR and not logged by an audiometrist. In a first proof-of-concept study the smart speaker-based measurement was already evaluated in a single office room with six normal-hearing (NH) subjects (Ooster *et al.*, 2019). However, the accuracy for hearing-impaired (HI) subjects, which is crucial for speech audiometry, has not been determined. Furthermore, in a real use case, the acoustic conditions in which the test is conducted can exhibit large variability. The user can be asked to avoid any background noise (in order to get an accurate test result), but it is often not possible to easily change the acoustics of the room where the smart speaker is placed. Therefore, in this study, we evaluate the measurement procedure when testing mildly and moderately HI subjects and secondly quantify the influence of acoustic conditions by conducting the experiments with three different kinds of reverberation.

METHODS

Matrix sentence test

The speech audiometric test used in this study is the German matrix sentence test (Wagener *et al.*, 1999). During testing, the subject hears sentences in stationary speech-shaped noise, and the SNR is dynamically adapted to reach the SRT after presenting 20 noisy sentences. The final measurement outcome is estimated by a likelihood fit of a psychometric function to the 20 data points of the whole measurement. 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*. Through this procedure, the individual words of the sentence are independent. This results in a low test-to-retest standard deviation of 1 dB for HI subjects (Wagener and Brand, 2005) and 0.5 dB for NH subjects (Brand and Kollmeier, 2002).

The smart speaker application

The elements of the smart speaker application for the automated SRT measurement are shown in the overview Figure 1 (Ooster *et al.*, 2019). The application is implemented with the Alexa Skill Developer Kit in Python (`github.com/alexa/alexa-skills-kit-sdk-for-python`). When the measurement application is

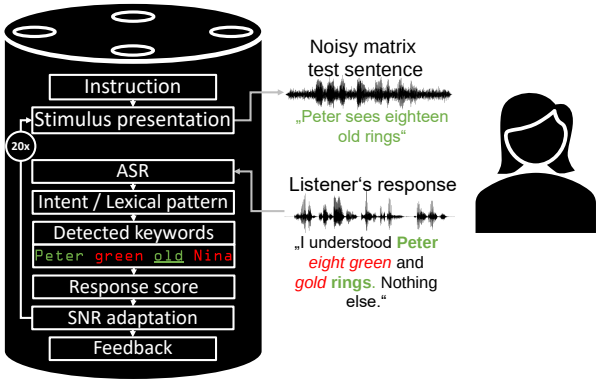

Fig. 1: Overview of the smart speaker measurement application.

started, the subject first hears an instruction about the general measurement procedure and the structure of this hearing test. These instructions are based on the guidelines for the clinical application of the matrix sentence test. During the measurement itself the dialogue manager of the smart speaker uses *intents*, which are derived from lexical patterns on the ASR output to take the next action within the measurement application. The core intent of the measurement application is the response to a matrix stimulus sentence. The lexical pattern to invoke this intent are based on realistic responses obtained in previous work (Ooster *et al.*, 2018). Based on this intent, the keywords in the subjects response are collected; the SNR for the next presentation is adapted based on the resulting score. Since the original speech material of the matrix sentence test is protected by copyright, 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). All stimulus matrix sentence audio files 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) in order to be properly played on the smart speaker.

Evaluation measurements

The prototype application was evaluated using an *Amazon Echo Plus 2nd Generation* loudspeaker. The measurements were conducted with subjects with three different hearing-profiles; normal hearing, mildly hearing impaired, and moderately hearing impaired. The subjects were categorized with pure tone average (PTA) criteria from .5 to 4 kHz (Mathers *et al.*, 2001). All subjects were paid for the participation in this study. The smart speaker measurements were conducted in a room that uses

distributed microphones and loudspeakers to simulate different room acoustics. The subjects were sitting in the center of the room with the smart speaker in front on a table at a distance of 2 m. At the beginning of the measurement, the subjects were asked to adjust the volume of the speaker to an easy intelligibility of the speech assistance's voice. To account for different acoustic conditions, rooms with different reverberation times T_{30} were simulated: *Living Room* ($T_{30} = 0.51$s), *Poor Classroom* ($T_{30} = 1.12$s) and *Concert Hall* ($T_{30} = 1.52$s). Every time the room settings were changed, the subjects heard four random sentences at different SNR so they could adapt to the new room and also could adjust the volume of the speaker. The subjects were always allowed to change the volume of the speaker again during the measurement. The subjects were invited for two measurement sessions each with nine SRT measurements in total, as described in Table 1. Overall, each subject conducted 16 measurements with the smart speaker application as well as two clinical reference measurements. The clinical reference measurements were conducted in an isolated sound booth,

Room A				Room B		Room C		Booth
Training1	Training2	Test1	Test2	Test3	Test4	Test5	Test6	Reference

Table 1: Measurement procedure during one of the two sessions for each subject. While the reference measurement with the clinical setup was always in the end, the order of the room settings during the smart speaker measurement was randomly chosen for each subject.

with a calibrated and equalized loudspeaker, the original, female, natural voice for the stimulus sentences (Wagener *et al.*, 2014) and a human supervisor for response scoring. At the end of each measurement session, all recorded audio files in the cloud of the smart speaker were deleted so that the ASR system is not adapted to that speaker for the measurement with the next subject. Parallel to the measurements, a human supervisor scored the subjects responses to have the true value for the scoring for each sentence (assuming that the experienced human supervisor produces no errors when logging the reported words). These true transcripts were later used to quantify the errors of the smart speaker ASR in terms of the score insertion rate (SIR) and the score deletion rate (SDR), i.e., the errors that could actually have an influence on the SRT by inserting or deleting a word. They are defined by

$$SIR = \frac{N_{score\ insertions}}{N_{score}}; SDR = \frac{N_{score\ deletions}}{N_{score}}, \qquad \text{(Eq. 1)}$$

where the number of errors $N_{score\ insertion}$ and $N_{score\ deletion}$ are normalized by the number of correctly repeated matrix sentence test words in the subject's response N_{score}. The order of the words is neglected in this error metric. The full error rates in the classical sense of an ASR system cannot be calculated since the full transcript (including non score relevant words) was not created. Details on the evaluation metrics can be found in Ooster *et al.* (2018).

RESULTS

The evaluation measurements were conducted with 5 subjects from each subject group, resulting in a total of 15 subjects as described in Table 2. Figure 2 describes the

	Normal-hearing	Mild hearing Loss	Moderate hearing loss
N (f/m)	5 (2/3)	5 (1/4)	5 (2/3)
Age	62 +/- 6 years	68 +/- 1 years	60 +/- 11 years
PTA	13 +/- 7 dB	29 +/- 5 dB	47 +/- 8 dB

Table 2: Description of the subjects who participated in the evaluation.

SRT measurement accuracy of the measurement with the smart speaker application compared to clinically acquired estimates. While the black line indicates a potential perfect match between the clinically measured value and the value estimated with the smart speaker application, most of the measured points are above this line. This highly

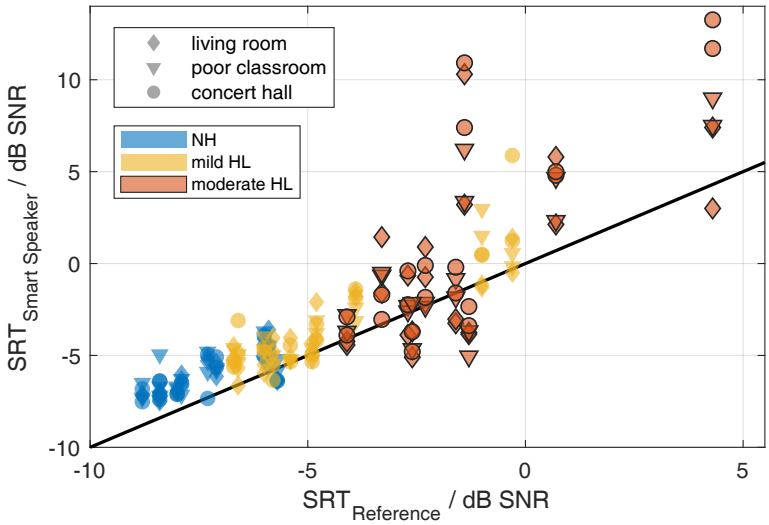

Fig. 2: The measured SRTs with the smart speaker application plotted vs. the SRTs measured with the clinical reference setup in the same session. Color depicts the hearing loss categorization based on the PTA criterion; the shape of data points denotes the room setting during the smart speaker measurement.

significant bias (paired-sample t-test, $p = 3.1 \cdot 10^{-15}$) that amounts to 1.38 dB on average is constant over the acoustic conditions and subjects groups. We did not find any significant difference with the t-test between the different room settings and the subject groups. The intra- and inter-subject standard deviation (SD; 1.37dB/3.79dB) are higher than with the clinical setup (0.76dB/3.11dB) over all subject groups and acoustic conditions. The inter- and the intra-subject SD varied slightly in the three

different acoustic conditions, with the highest increase of the inter-subject SD in the *Concert Hall* condition of about 1 dB. The intra-subject SD is increased by 0.39 dB in this condition. For the mildly HI subjects the intra- and inter-subject SD is slightly increased in comparrsion to the NH subjects by 0.32 dB and 0.30 dB, respectively. For the moderately HI subjects both the intra- and the inter-subject SD is increased by 1.45 dB and 1.76 dB, respectively, in comparison to the mildly HI subjects. This is also due to the fact that two measurement sessions failed completely: In one of these measurement sessions, the ASR performance was with 19.8% SDR (6.9% SIR) very low in comparison to the other subjects, which resulted in an bias of 8.3 dB and a SD of 3.3 dB. In the next measurement session of this subject, the ASR performance was better (at 8.5% SDR and 4.1% SIR), which is consistent with a much higher measurement accuracy with a bias of -0.2 dB and a SD of 1.2 dB. The second inaccurate measurement session is not explainable by the error rates of the ASR system (SDR = 7.4%, SIR = 5.8%). However, we noticed that the subject was speaking very quietly which resulted in several terminations of the measurement application. Although the subject spoke in normal volume towards the end of the session, the terminations could have had a large effects on the ability of the subjects to focus on the listening task. When excluding these two subjects the increase of the inter- and intra-subject SD for the moderately HI subjects goes down to 0.33 dB and 0.70 dB, respectively.

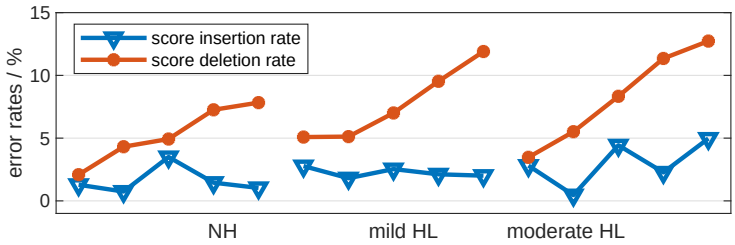

Fig. 3: The ASR performance of the smart speaker for each of the 15 subjects.

The ASR performance of the smart speaker for all subjects is shown in Figure 3. With an average of SDR = $(8.0 \pm 3.2)\%$ are the score deletion errors significantly higher for the HI subjects than the score deletion error of the NH subjects, which had an average of SDR = $(5.3 \pm 2.3)\%$ (two sample t-test, equal variances not assumed, $p = 3.0\%$). Three of the HI subjects showed SDRs above 10%; through post analysis (as discussed above) ASR errors for one of those subjects could be attributed to a strong decrease of the SRT measurement accuracy. The score insertion errors are below 5% for all of the subjects and no significant difference was found between NH and HI subjects.

DISCUSSION

In this study, we investigated the SRT measurement accuracy with a smart speaker-based application in three different acoustic conditions and with three different subject groups. In our previous study regarding the speech-controlled automated matrix

test (SAMT; Ooster *et al.*, 2018), we didn't find any significant decrease of the measurement accuracy conditioned by the errors from the ASR system. In this study, an ASR system, that was not that fine-tuned to the words of the matrix test was used in a more challenging acoustic condition (far-field recognition with reverberation) and therefore the obtained error rates are higher. For one subject this resulted in a very inaccurate measurement, but overall the observed intra-subject SD is very similar to the one of the clinical application. The only subject group with an increase intra-subject SD are the moderately HI subject. This subject group has very similar ASR error rates as the mildly HI subjects, so the decreased replicability of the measurement outcome seems not to be indicated by the smart speaker based measurement itself, but presumably due to an insecurity of subjects in terms of speech-based interaction with a speaker. The obtained ASR error rates during this study represent a lower boundary of the ASR performance, since in a real use case the ASR system should be adapted to the specific user and secondly owners of smart speakers are probably used to speech-based inputs and normal patterns of interaction. The moderately HI subjects showed a decreased measurement accuracy with the smart speaker application, but most of the variance during the measurement with the moderate HI subjects is towards higher (worse) SRTs and therefore would not change the screening result.

CONCLUSIONS

In this paper, we have shown that speech audiometry conducted with a smart speaker for at-home screening of hearing deficits is possible with an intra-subject SD of 1.37 dB. The bias between the clinical and the smart-speaker measurement is significant, but consistent across subject groups and room settings, and no significant difference was found between the groups and conditions, respectively. While normal-hearing and mildly hearing-impaired subjects showed a very similar measurement accuracy as the clinical reference measurement, the inter- and intra-subject SD is increased for moderately hearing-impaired subjects by 1.39 dB and 1.89 dB, respectively. This was attributed to the results for single subjects, whose speech produced high ASR error rates or was too low to properly conduct the measurement. When excluding these subjects from the analyses the increase of inter- and intra-subject SD goes down to 0.33 dB and 0.70 dB, respectively and the overall intra-subjects SD goes down to 0.91 dB which is comparable to the intra-subject SD with the clinical measurement setup of 0.67 dB.

In future work, we will develop an SRT-based criterion to produce a recommendation for the test user (e.g., to seek advice from an audiometrist), based on ratings of his or her performance for the test. This will require a larger number of listeners to be tested to establish a reliable statistical foundation for such a recommendation.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project

ID 390895286 and the CRC TRR 31, Transfer Project T01.

REFERENCES

Arlinger, S. (**2003**) "Negative consequences of uncorrected hearing loss - a review." Int. J. Audiol., **42**(2), S17–S20. doi: 10.3109/14992020309074639

Boumpa, E., Gkogkidis, A., Charalampou, I., Ntaliani, A., Kakarountas, A., and Kokkinos, V. (**2019**). "An Acoustic-Based Smart Home System for People Suffering from Dementia". Technol., **7**(1), 29. doi: 10.3390/technologies7010029

Brand, T., and 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. doi: 10.1121/1.1479152

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

Mathers, C., Smith, A., and Concha, M. (**2001**). "Global burden of hearing loss in the year 2000." Global Burden of Disease 2000, **18**(4), 1–30.

Nuesse, T., Wiercinski, B., Brand, T. and Holube, I. (**2019**). "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech." Trends Hear. doi: 10.1177/2331216519862982

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

Ooster, J., Porysek Moreta, P. N., Bach, J.-H., Holube, I. and Meyer, B.T. (**2019**). "Computer, test my hearing: Accurate speech audiometry with smart speakers" Proceedings of the Interspeech 2019, Graz, Austria. doi 10.21437/Interspeech.2019-2118

Vora, J., Tanwar, S., Tyagi, S., Kumar, N., and Rodrigues, J. J. P. C. (**2017**). "Home-based exercise system for patients using IoT enabled smart speaker". In 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom. doi: 10.1109/HealthCom.2017.8210826

Wagener, K. C., Kühne, V., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztests" (Development and evaluation of a German speech intelligibility test. Part I-III: Design, optimization and evaluation of the Oldenburg sentence test). Z Audiol., **38**(1-3), 4-15, 44-56, 86-95.

Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters". Int. J. Audiol., **44**(3), 144–156. doi: 10.1080/14992020500057517

Wagener K., Hochmuth S., Ahrlich M., Zokoll M. and Kollmeier B. (**2014**), "Der weibliche Oldenburger Satztest". 17. DGA Jahrestagung, Oldenburg, Germany.