INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

"Computer, test my hearing": Accurate speech audiometry with smart speakers

Jasper Ooster[1,4], *Pia Nancy Porysek Moreta*[1,4], *Jörg-Hendrik Bach*[2,4], *Inga Holube*[3,4], *Bernd T. Meyer*[1,4]

[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]HörTech gGmbH, Oldenburg, Germany
[3]Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany
[4]Cluster of Excellence Hearing4all, Germany

jasper.ooster@uol.de, pia.nancy.porysek.moreta@uol.de, j.bach@hoertech.de, inga.holube@jade-hs.de, bernd.meyer@uol.de

Abstract

Speech audiometry based on matrix sentence tests is an important diagnostic tool for hearing impairment and fitting of hearing aids. This paper introduces a self-conducted measurement for estimating the speech reception threshold (SRT) of a subject, i.e., the signal-to-noise ratio corresponding to 50% intelligibility, based on a smart speaker. While the original measurement procedure is well-evaluated and provides a very high measurement accuracy (< 1 dB test-retest standard deviation), the measurement using a smart speaker differs in several aspects from the commercially available implementation, such as missing control over the absolute presentation level, mode of presentation (headphones vs. loudspeaker), potential errors from the automated response logging, and influence from room acoustics. The SRT measurement accuracy is evaluated with six normal-hearing subjects conducted with an Amazon Alexa application on an Echo Plus loudspeaker in a quiet office environment. We found a significant difference of 0.6 dB in SRT between the proposed and the commercially available testing procedure. However, this bias is smaller than the inter-subject standard deviation, and the measurement accuracy is similar to the original test for normal-hearing listeners, which indicates that smart speakers may become a helpful addition for the screening of hearing deficits.

Index Terms: auditory healthcare, speech audiometry, speech reception threshold, speech recognition, smart speakers

1. Introduction

The ability of normal-hearing listeners to recognize and understand spoken language is crucial for communication and generally for social interactions. This ability is often limited in hearing-impaired (HI) listeners, which has negative consequences for communication and their quality of life. The early diagnosis of hearing deficits can contribute to a compensation of these negative consequences, e.g., by providing hearing aids shortly after a diagnosis of hearing impairment [1]. A well-established measurement to assess hearing deficits is the matrix sentence test, which is available in 21 languages [2]. The target value of this test is the speech reception threshold (SRT), i.e., the signal-to-noise ratio at which 50% of words are correctly identified by the listener. In this test, noisy nonsense sentences with a fixed grammar are used as stimulus, and the SNR is dynamically adapted until it converges to the SRT after 20 to 30 presentations of different matrix sentences. In its commercially available application, the test is performed in the presence of a professional supervisor to record the responses. This requires a lot of resources and therefore could prevent people from actually being tested. A potential remedy is to perform the closed test with a graphical user interface (GUI) to select responses from a displayed matrix. A GUI-based testing however skews the results since the full vocabulary is shown to the participant and, maybe more importantly, it excludes people who cannot read, such as pre-school children, the visually impaired, or the functionally illiterate (14.5% of the German population between 18 and 64 [3]).

To address the limitations of GUI-based tests, we developed the Speech-controlled Automated Matrix Test (SAMT) where an automatic speech recognition (ASR) system was interfaced with the clinical measurement software of the commercial test to conduct the matrix sentence test measurements with a purely speech-based interface [4]. While it was shown that it is possible to conduct automated measurements without a loss in measurement accuracy, this system is not available to end users at home.

Recently, smart speakers such as Amazon's *Echo*, Apple's *HomePod* or *Google Home* have gained a lot of attention. These speakers are connected to a speech assistant for hands-free control of the device. Functions such as music control, organization of daily schedules or voice-controlled home automation are well-established as features for smart speakers, and there is also interest in using them in the context of healthcare: For instance, a smart home system was introduced that provides acoustic cues to support dementia patients' memory [5] and related research proposed the use of smart speakers to support elderly people with their physical therapy [6]. The Apple ResearchKit [7] contains a speech-in-noise test that is similar to our approach, but has not been compared to a commercial test in a clinical setting to the best of our knowledge.

In this paper, we explore for the first time the accuracy of self-conducted SRT measurements using a smart speaker while comparing the outcome to the commercially available German matrix sentence test. Four main differences exist between these two procedures, all of which could introduce a substantial error for the SRT measure: (1) We use a high-quality speech synthesis instead of the natural speech files that are protected by copyright, (2) the sound is presented via the speaker in a reverberant environment and not via audiological headphones, (3) com-

http://dx.doi.org/10.21437/Interspeech.2019-2118

pressed audio files are presented, and (4) the listener's response is transcribed via ASR and not logged by an audiometrist. Our analysis could contribute to applications of smart speakers in auditory healthcare, and to lowering the threshold for the diagnosis of hearing deficits.

2. Methods

2.1. Matrix Sentence Test

Experiments in this paper are performed with the German matrix sentence test [8], which is described in the following. The speech material used as stimulus is based on a random path through a five by ten word matrix, resulting in sentences with identical structure *Name Verb Numeral Adjective Object*. Due to this random structure, the sentences are syntactically fixed but semantically unpredictable. Therefore, it is not possible to predict the next word from the previous words of the sentence, which results in a very high test efficiency of less than 1 dB test-to-retest standard deviation for hearing-impaired subjects [9] and 0.5 dB for normal-hearing subjects [10]. These matrix sentences were presented in speech-shaped stationary noise to the subjects who respond with the recognized words. A supervisor compares the response to the stimulus sentence in order to capture the response score, i.e., the number of correct recognized words. Based on this response score, the SNR is adapted for the next presentation in order to converge the SNR to the SRT. After the presentation of 20 matrix sentences, the SRT is calculated by a likelihood fit of a psychometric function to the 20 data points. Due to the limited vocabulary, listeners can quickly learn the vocabulary which results in a strong training effect during the first six measurements of $\approx$ 2 dB [11]. The strongest decrease in SRT of 1 dB is observed between the first and the second measurement.

2.2. Listening tests with a smart speaker

The structure of the smart speaker application follows the diagram in Figure 1 and is implemented with the Alexa skill developer kit in Python.

Synthetic speech: In contrast to the original matrix sentence test that uses recordings of a male [8] or a female speaker, the sentences of our system were generated with a speech synthesis program. This was done since the original speech material is protected by copyright. We used synthetic speech signals that were found to exhibit the highest perceived naturalness as well as the highest speech intelligibility for a group of normal-hearing listeners. The corresponding study [12] compared three different synthesis algorithms for conducting matrix sentence tests. The commercial synthesis provided by the Acapela Group was found to produce the best results and was therefore used to synthesize the 150 sentences of the German matrix test with a female voice. The synthetic female speech shows similar characteristics and SRTs as the natural female speech [12].

Presentation level: The noisy sentences need to be presented at different SNRs and were therefore pre-mixed with a speech-shaped stationary noise in steps of 0.1 dB. When performing the clinical test, the noise is continuously presented at a fixed level while the speech level is dynamically adapted. The smart speaker does not allow a continuous playback when the ASR is active, and the system normalizes each presented sentence, i.e., there was no direct control over the presentation level of the speech or the noise component.

Configuration of the ASR component: To trigger the next event (changing the SNR) from recognized matrix words, the Alexa development software requires the definition of word patterns that need to be matched, and which are referred to as *intents*. To provide sufficient variability for listeners' responses, we defined several of these word patterns, e.g., regular matrix test sentences (either complete or with one word missing in an arbitrary position) as well as patterns that match only few matrix words (or only one word) for the case of trials with a low intelligibility. In some of these patterns, typical fill words that were observed in listeners' responses [4] were added to account for the variability in real-life responses. One example is "Ich habe {*Matrix Word*} verstanden" (I have understood {*Matrix Word*}). Due to the adaptive measurement procedure, the subjects also heard presentations without understanding any word. To address these responses, we used the pre-built skip patterns from Amazon extended with common phrases from our previous evaluation measurements.

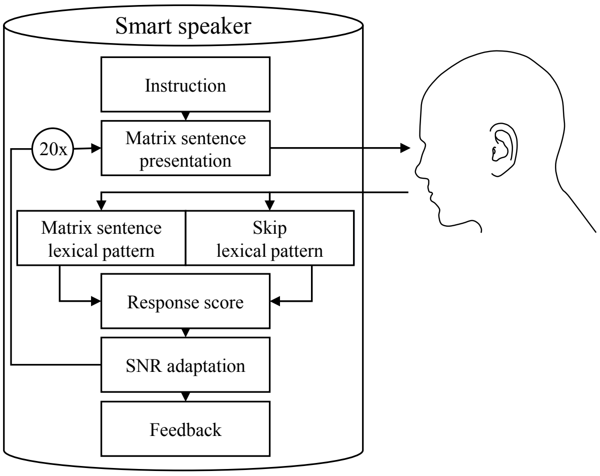

Figure 1: *Overview of the measurement loop of the matrix sentence test application.*

2.3. Error metric

To quantify the ASR performance, we analyzed score insertion errors (i.e., additional words in the ASR transcript in comparison to the true label, which incorrectly increases the response score) and score deletion errors (missing words, which incorrectly decreases the response score). Substitutions were not considered since all relevant errors that could affect the SRT measurement can be categorized as either score insertion or score deletion error. To calculate the error rate, the number of errors $N_{score\ insertion}/N_{score\ deletion}$ are normalized by the number of correctly repeated matrix sentence test words in the subject's response, i.e., the response score N_{score}:

$$SIR = \frac{N_{score\ insertions}}{N_{score}}; \quad SDR = \frac{N_{score\ deletions}}{N_{score}} \qquad (1)$$

All other words that do not belong to the 50-word matrix cannot effect the scoring and are therefore treated as non-matrix-vocabulary (NMV) words. Further details about the error metric can be found in [4].

2.4. Evaluation Measurements

To evaluate the self-measurement accuracy, the application was tested with six self-reported normal-hearing subjects in a small office (18 m^2, 48 m^3) with a reverberation time of T$_{60}$ =

550 ms. The subjects were between 20 and 28 years old (avg. 24 ± 2.8 years) and were seated at a distance of approximately 1.5 m from the smart speaker. Three of the subjects had prior experience with the matrix sentence test. The full instruction and training of the subjects was carried out with the smart speaker application. Each of the six subjects conducted six measurements with the smart speaker, and each measurement contained 20 sentences. The first two full presentation cycles (each with 20 sentences) are used to train the subjects, which resembles the procedure of the commercial matrix test. To make the training more effective, the first ten sentences were presented at a relatively high SNR (0 dB), so that the subjects understood all of the 50 words once before the adaptive measurement procedure began. The only additional instruction given to the subjects was that the device is only listening when its optical indicator is active since the subjects in our experiment had no experience with smart speakers. We assume that regular users of smart speakers do not need this additional instruction. The subjects' responses were additionally recorded with an independent microphone and subsequently manually transcribed. These manual transcripts are later compared to the ASR transcript to determine the ASR error rate. After this smart speaker measurement, the SRT of each subject was measured with a standard setup: Noisy sentences produced by a female speaker were presented in an isolated sound booth over Sennheiser HDA200 headphones to the subjects, and their response was logged by a human supervisor. Signals were presented diotically, i.e., the same signal was used for the left and right channel, since this should represent the listening task during the measurement with the smart-speaker application.

3. Results

In the following, we first describe the listeners' behaviour, since we found their interaction patterns to be important, and continue with the general ASR performance before the main result in terms of measurement accuracy is described.

3.1. Subject response behaviour

The subjects' response behaviour is similar to observed behaviour in our previous study with the SAMT system since we determined a similar average NMV word rate of $(10.8 \pm 5.7)\%$ (see Figure 2) in comparison to $(10.6 \pm 6.6)\%$ in [4]. The smart speaker stops the recording as soon as the subject interrupts the response and one of the predefined lexical patterns is already matched. A single matrix sentence word has to match the lexical pattern since it is a valid response during the measurement at low SNRs. When the subjects interrupt their response to think about the next word (something we observed frequently during the evaluation of the SAMT system, where we had up do 4 s waiting time at the end of a subject's response) the recording ends directly and the subject cannot enter more words. Nevertheless, this did not pose a problem in our measurements, since the subjects quickly learned to adapt their response behaviour based on the visual feedback through the light ring that indicates when the device is listening.

3.2. ASR performance

The ASR error rates obtained by comparing ASR-based and manual transcript as described above are shown in Figure 2. The highest SIR reaches 10% whereas the average score insertion rate is $(6.0 \pm 2.3)\%$. The SDR is always below 5% with an average of $(3.7 \pm 1.0)\%$. During the development of the application, we observed that ASR transcripts contained some systematic errors (e.g., the word *'Stein'* (Stone) was transcribed as *'Steine'* (Stones)), which was automatically corrected. Without compensating for six of such systematic errors, the score deletion and insertion rates are 10.1% and 5.7%, respectively.

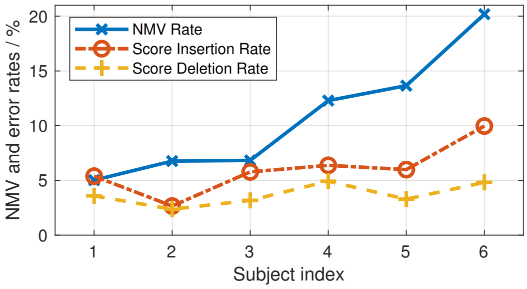

Figure 2: *Performance of the ASR system for each of the six subjects. The results are mean values over all six measurements with the smartspeaker application.*

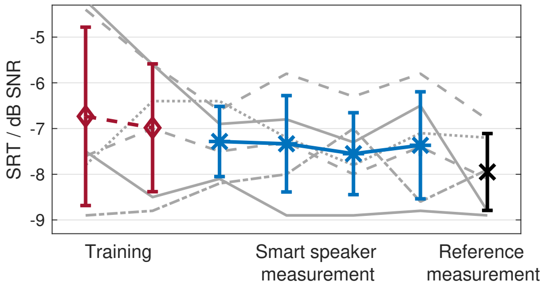

Figure 3: *Average SRTs and the inter-subject standard deviation. The reference value is measured with the commercially available setup. The gray lines indicate the individual SRTs.*

3.3. Measurement accuracy

Figure 3 shows the average SRTs with the inter-subject standard deviation of the six measurements with the smart speaker application and of the reference measurement with the commercially available setup. After the training, we measured an overall average SRT of (-7.4 ± 0.9) dB SNR with the smart speaker application and (-8.0 ± 0.9) dB SNR with the commercially available setup. This bias of 0.6 dB is significant (paired-sample t-test, $p = 0.002$). Regardless of the measurement setup, the average SRT measured in this study was higher than the reference value for normal-hearing subjects with the natural female speaker (-9.4 ± 1.0) dB SNR [13]) and with the synthetic female voice (-8.6 ± 0.6) dB SNR [12]) despite the fact that we measured with binaural stimuli (which gives a benefit for understanding speech in noise) and not with monaural stimuli as in the reference studies (note that lower SRTs represent *better* performance). Furthermore, the measured training effect is smaller than in the reference studies. The difference between the first and the third measurement (the first measurement after the training phase) is 0.55 dB. After the training phase, no additional significant training effect is observable.
To estimate the test-to-retest standard deviation, we subtracted each measured SRT value from the following value. This results in a test-to-retest standard deviation of 0.71 dB.

4. Discussion

Our experiments have shown that it is possible to conduct speech audiometric tests with a smart speaker. Although we introduced several differences in comparison to the commercially available measurement, the SRT measurement accuracy with the self-administered speech-controlled test is very high in terms of intra-subject standard deviation, and the observed bias is very small, which is encouraging. Comparable studies analyzed single differences to the commercial procedure used here, and in several cases stronger effects were observed. Each of the differences introduced in the present study should result in a higher test difficulty, and should therefore result in worse (higher) SRTs. Given the nature of these differences (as described in the final part of the introduction), it seems unlikely that their effects could cancel each other out.

For instance, in a study in which the synthesized speech signals used by us were first introduced [12], an SRT difference of 0.8 dB was measured when comparing synthetic and human voices. Presumably, this difference is therefore the main reason for the measured bias in our study, since we are using the same stimuli and the same reference.

For a related German speech test that used combinations of digits rather than sentences (the German digit triplet test), it was shown that a difference of 0.5 dB in SRT is introduced by the presentation in a room with a volume of 220 m^3 and a reverberation time of 0.86 s [14]. The authors therefore performed measurements in a larger room with a higher reverberation time, but also used smaller and simpler loudspeakers.

The same study showed that an MP3-based speech coding with a bitrate of 128 kBit/s increases the SRT by 0.5 dB for the digit triplet test. Given that our experiments used a bitrate of only 48 kBit/s, we could have expected a decrease in SRT accuracy by this compression alone. We used the FFmpeg encoder, while the BladeEnc was used in [14] which has not been in development since 2001, which could result in an inferior compression performance compared to the state-of-the-art.

Our previous study [4] reported no significant SRT difference introduced from the ASR errors obtained from a Kaldi-based system [15]. This is consistent with the results obtained in the current paper, since errors from the Alexa-based ASR were found to have a negligible effect with an estimated increase of 0.1 dB in intra-subject standard deviation and a bias of approximately 0.1 dB.

In the automated test introduced here, the presentation level is not calibrated to a specific level, and the individual noise and speech levels could not be controlled for at all (as explained in the methods section), which seems not to be crucial: Wagener and colleagues [9] did not measure a significant influence of the presentation level on the SRT result (for levels that are clearly above the hearing threshold).

Independently of the smart speaker measurements, the measured SRTs were higher (worse) on average compared to the reference values for normal-hearing subjects, which is presumably due to the specific group of listeners combined with the relative small number of listeners: The increased average SRT is dominated by two subjects, which had SRTs of -6.8 dB SNR and -7.1 dB SNR during the reference measurement.

A limitation of our study is that we explored only one specific condition, i.e., one office room, a fixed position for the listeners using the same device. Measurements in uncontrolled home measurements with loudspeakers could introduce a difference of 1.1 dB as reported for the the digit triplet test [16]. This effect could partially be circumvented by measuring the noise background level or estimate the reverberation time before a test can be conducted. Interestingly, reverberation could also have a positive effect for distinguishing between normal-hearing and HI users, since SRTs from HI users degrade more quickly, thereby increasing the difference between both groups [17]. This should however require a thorough analysis that should be linked to acoustic room parameters. Using different smart speakers could lead to very different results, since both the acoustical presentation as well as the ASR performance are crucial components of the proposed approach.

In future work the influence of different rooms and the interaction with different devices should be investigated. While the measurement with the matrix sentences has the potential to deliver accurate test results, the listeners should get a comprehensible summary after the measurement that is more informative than the SRT (which is a value that is usually interpretable by experts only). For the triplet digit test, the feedback 'good', 'insufficient', or 'poor' was provided [18, 19]. A similar feedback could be used by the smart speaker (possibly with a more friendly wording). This would require an evaluation with a broader subject base, including HI subjects.

In view of upcoming commercial consumer-grade hearing-assistant systems (in contrast to medically indicated hearing-aids), such an automated self-measurement could provide an objective and precise indicator for the hearing-assistant devices fitting success. Applications such as the one presented here may provide precise and useful tools for self-assessment of hearing in devices that rely entirely on self-fitting without intervention by an expert, e.g., over-the-counter devices [20] or personal sound amplifiers.

5. Summary

In this paper, we have explored the applicability of a smart speaker to conduct an automated speech audiometry test using a speech interface. Several changes were made to the procedure that is used for the commercially available German matrix test, including a synthesized instead of a natural voice, as well as ASR-based transcription instead of relying on a human supervisor. Despite these differences, the system achieves a relatively high accuracy when compared to the clinical test that was performed with the same group of six normal-hearing listeners: The test-to-retest standard deviation was found to be 0.7 dB for the smart speaker system, which is close to 0.5 dB achieved with the clinical application. This shows that smart speakers may provide reliable speech-controlled tools for self-screening of hearing deficits, which in turn could result in a lower-threshold for the diagnosis of hearing deficits. In future work, the reliability of the approach should be tested with hearing-impaired listeners and take into account the variability that could be introduced by different room acoustics, acoustic properties of the speaker, as well as the accuracy of the ASR component required to transcribe listeners' responses.

6. 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.

7. References

[1] S. Arlinger, “Negative consequences of uncorrected hearing loss - a review,” *International Journal of Audiology*, vol. 42, no. 2, pp. S17–S20, 2003.

[2] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, “The multilingual matrix test: Principles, applications, and comparison across languages: A review,” *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16, 2015.

[3] A. Grotlüschen and W. Riekmann, *Functional illiteracy in Germany - Results of the first leo. - Level-One Study*, 10th ed. Bundesverband Alphabetisierung und Grundbildung e.V, 2012. [Online]. Available: http://blogs.epb.uni-hamburg.de/leo/

[4] J. Ooster, R. Huber, B. Kollmeier, and B. T. Meyer, “Evaluation of an automated speech-controlled listening test with spontaneous and read responses,” *Speech Communication*, vol. 98, pp. 85 – 94, 2018. [Online]. Available: www.sciencedirect.com/science/article/pii/S0167639317302698

[5] E. Boumpa, A. Gkogkidis, I. Charalampou, A. Ntaliani, A. Kakarountas, and V. Kokkinos, “An Acoustic-Based Smart Home System for People Suffering from Dementia,” *Technologies*, vol. 7, no. 1, p. 29, 2019.

[6] J. Vora, S. Tanwar, S. Tyagi, N. Kumar, and J. J. P. C. Rodrigues, “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)*, Oct 2017, pp. 1–6.

[7] Apple Inc., “Researchkit - speech-in-noise test,” accessed: 2019-03-29. [Online]. Available: http://researchkit.org/docs/docs/ActiveTasks/ActiveTasks.html#speech_in_noise

[8] K. Wagener, V. Kühnel, and B. Kollmeier, “Development and evaluation of a German speech intelligibility test. Part I: Design of the Oldenburg sentence test,” *Zeitschrift für Audiologie*, vol. 38, no. 1, 1999.

[9] K. C. Wagener and T. Brand, “Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters,” *International Journal of Audiology*, vol. 44, no. 3, pp. 144–156, 2005.

[10] T. Brand and B. Kollmeier, “Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests,” *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2801–2810, 2002.

[11] K. Wagener, V. Kühnel, and B. Kollmeier, “Development and evaluation of a German speech intelligibility test. Part II: Optimization of the Oldenburg sentence test,” *Zeitschrift für Audiologie*, vol. 38, no. 2, 1999.

[12] T. Nuesse, B. Wiercinski, T. Brand, and I. Holube, “Measuring speech recognition with a matrix test using synthetic speech,” *Trends in Hearing*, 2019.

[13] M. Ahrlich, “Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit (Optimization and evaluation of the Oldenburger sentence test with female voice and the examination of the speakers effect on speech intelligibility),” Bachelor thesis, Carl von Ossietzky Universität Oldenburg, 2013.

[14] M. Buschermöhle, K. C. Wagener, D. Berg, M. Meis, and B. Kollmeier, “The German Digit Triplets Test (Part I): Implementations for Telephone , Internet and Mobile Devices,” *Zeitschrift für Audiologie*, vol. 53, no. 4, pp. 139–145, 2014.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[16] C. Smits, P. Merkus, and T. Houtgast, “How we do it: The Dutch functional hearing?screening tests by telephone and internet,” *Clinical Otolaryngology*, vol. 31, no. 5, pp. 436–440, 2006.

[17] H. Schepker, K. Haeder, J. Rennies, and I. Holube, “Perceived listening effort and speech intelligibility in reverberation and noise for hearing-impaired listeners,” *International Journal of Audiology*, vol. 55, no. 12, pp. 738–747, 2016.

[18] C. Smits, T. S. Kapteyn, and T. Houtgast, “Development and validation of an automatic speech-in-noise screening test by telephone,” *International Journal of Audiology*, vol. 43, no. 1, pp. 15–28, 2004.

[19] M. Buschermöhle, K. C. Wagener, D. Berg, M. Meis, and B. Kollmeier, “The German Digit Triplets Test (Part II): Validation and Pass/Fail Criteria,” *Zeitschrift für Audiologie*, vol. 54, no. 1, pp. 6–13, 2015.

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