Autonomous measurement of speech intelligibility utilizing automatic speech recognition

Bernd T. Meyer, Birger Kollmeier, Jasper Ooster

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Germany

{bernd.meyer, birger.kollmeier, jasper.ooster}@uni-oldenburg.de

10.21437/Interspeech.2015-617

Abstract

Measures of speech intelligibility are an essential tool for diagnosing hearing impairment and for tuning hearing aid parameters. This study explores the potential of automatic speech recognition (ASR) for conducting autonomous listening tests. In these tests (e.g., in the Oldenburg sentence matrix test employed here) the responses of participants are usually logged by a (human) supervisor. The target value is the speech reception threshold (SRT), i.e., the signal-to-noise ratio at which 50% speech intelligibility is achieved. We explore what ASR error rates can be obtained for such responses, and how ASR errors affect the measured SRT value. To this end, a speech database was recorded that contains utterances from 20 speakers and covers different levels of language complexity, ranging from simple five-word sentences to utterances as produced in typical human-human interactions during testing. While for the most complex speech material, the achievable SRT accuracy was not satisfactory, the ASR performance for sentences without out-of-vocabulary words was below 1.3% and hence sufficient to obtain a test-retest reliability of only 0.5 dB, which is identical to the reliability in human-supervised tests.

Index Terms: speech intelligibility, automatic speech recognition, speech reception threshold

1. Introduction

Tests to measure the average speech intelligibility of a listener gain more and more importance in an aging society: For instance, about 17% of the German population exhibit a hearing impairment that needs to be compensated [5]. A smaller, but still considerably large proportion of 13% of US citizens aged 12 years or older is suffering from binaural hearing loss [11].

Apart from pure-tone audiograms, the speech reception threshold (SRT), i.e., the signal-to-noise ratio (SNR) at which 50% speech intelligibility is achieved is the most important quantity in audiometry and hearing aid fitting. Unfortunately, SRT testing requires significant resources, since it is usually performed clinically with an expert logging the responses of the listener (who is asked to repeat the words he or she recognized from a stream of noisy words). The aim of this paper is to investigate the potential of SRT measurements based on automatic speech recognition (ASR) that do not require a supervisor and hence could facilitate access to SRT testing for the general public.

One successful approach to speech-based SRT testing is the use of matrix tests based on sentence lists that cannot be memorized. The first matrix test was designed by Hagerman [7] in Swedish and was based on random walks through a word matrix. A refined recipe for designing a matrix test that provides natural word transitions was proposed by Wagener et al. [19, 20, 21], which resulted in the Oldenburg Sentence test OLSA (where the name is derived from the German translation *Oldenburger Satztest*). During testing, the SNR is adapted based on the word score of a noisy sentence to obtain 50% word recognition rate, as shown in the example in Figure 1. The recipe was used to design matrix tests for a total of 14 different languages, e.g., English [6], Dutch [10], Italian [16], and Spanish [9]. A review of these tests is given in [12].

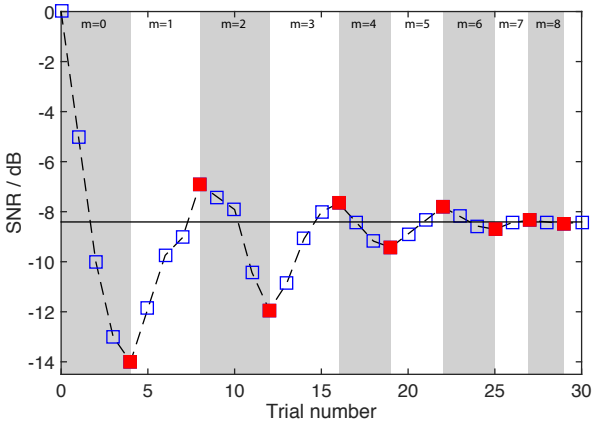

Figure 1: Example for SNR values in an SRT measurement. Filled markers denote extreme values after which the maximum step width is decreased (Eq. 1). The solid line marks the measured SRT of -8.4 dB. Note that lower SRT values correspond to better performance in this test.

By using a graphical user interface that shows all response alternatives, these tests are in principle suited for a closed-set format in unsupervised SRT measurements. For a purely acoustic measurement, this is not feasible. However, an acoustic-only ASR-based testing setup is desirable for testing children, visually-impaired people, and for phone-based testing. Further, since OLSA and its international counterparts have been compiled for a large number of languages, the ASR-based approached may also be used during regular (supervised) measurements when the audiologist does not speak the same language as the test subject, but could be assisted by ASR.

There are several services already established that provide a diagnosis via telephone or a smartphone app that can be used at home. In the context of ambient assisted living for instance, a system for classification of coughs was proposed in [18], which was later continued as a call-in service for remote diagnosis. A speech-in-noise test implemented as a smartphone app with graphical user interface called HearContrOl was presented in

[2]. However, to our knowledge a speech-based autonomous system to perform hearing diagnostics has not been developed yet, nor has the impact of ASR error sources on SRT measurements been investigated.

Therefore, our study analyzes if current ASR implementations are suitable for performing SRT tests autonomously, and to what extent errors produced by ASR are acceptable for an accurate measurement. The experiments are based on the OLSA matrix tests since it is widely used in hearing screenings and provides very accurate results with 0.5 dB of test-retest standard deviation (Section 2.1). We were especially interested in how the complexity of response formats affects ASR performance, and hence recorded a database with speech from 20 speakers containing complete and incomplete sentences, as well as utterances that resemble typical responses in a human-human interaction during a regular OLSA SRT measurement. This database is described in Section 2.2. After presenting the configuration of the ASR system (Section 2.3), we report the overall results for test sets of different complexity and also discuss consequences for using ASR in SRT measurements (Section 3).

2. Methods

2.1. Oldenburg Sentence Test OLSA

In this study we investigate the applicability of ASR to unsupervised matrix tests and picked the Oldenburg Sentence Test OLSA as a specific example, which is hence described in more detail. The sentences in the OLSA test follow the pattern (subject)(verb)(numeral)(adjective)(object). For each of these five groups, ten word alternatives exist which enables the generation of 10^5 different sentences, e.g., *Peter kauft acht nasse Steine* (German for *Peter buys eight wet stones*), which are syntactically correct but semantically unpredictable.

For the original OLSA test, 100 different sentences of those 10^5 alternatives were recorded, produced by a male speaker with normal articulation and moderate speaking rate [19]. Since the vocabulary size of OLSA is only 50, listeners can quickly adapt to it during a training session, but they cannot remember specific word sequences. The test procedure is adaptive: If only 0-2 words are identified, the SNR is increased, otherwise it is decreased. One important design parameter is the choice of the change in SNR, which needs to change quickly at the beginning (to cover a wide range of possible test outcomes) and converge towards the end of testing (to obtain a stable measure for the SRT). For our analysis, we employ the procedure by Brand and Kollmeier [1] which has been established as a standard rule for OLSA testing. The SNR change ΔS is given by

$$\Delta S(k) = -\frac{f(m) \cdot (P(k-1) - T)}{s}. \quad (1)$$

where k is the trial number, T is the target speech intelligibility (in our case 50%), and s is the average slope of SRT curves for matrix tests with a value of $0.15\,dB^{-1}$ (meaning a 15% increase of intelligibility for an SNR increase of 1 dB for the steepest part of the curve). $P(k-1)$ is the word recognition accuracy of the previous trial, and the parameter $f(m)$ controls the rate of convergence, where m denotes the index of the reversal (cf. filled markers in Figure 1). The optimal adaptation rate determined with Monte Carlo simulations is $f(m) = 1.5 \cdot 1.41^{-m}$ [1], which ensures potentially large SNR fluctuations for early trials and small changes after a couple of reversals. In conventional testing scenarios, a short training list is used to familiarize the listener with the vocabulary and the procedure. Thereby, a test-retest standard deviation of 0.5 dB is achieved [21].

The actual speech reception threshold is then calculated by

$$\text{SRT} = \text{SNR}_{init} + \sum_{k=1}^{K} \Delta S(k) \quad (2)$$

with an initial SNR of 0 dB for tests with mono signals in additive noise, and typically $K = 30$ sentences. Trained normal-hearing listeners usually achieve an SRT of -8.4 dB in this test [22].

2.2. Matrix test speech data

As prerequisite for the experiments in this study, we recorded speech data that resembled the original OLSA matrix test, yet was comprised of utterances with varying levels of complexity. The amount of data for each set is reported in Table 1. The data falls into one of three categories:

Complete, random sentences using the structure and vocabulary of the matrix test. Although this set is not sufficient to answer the questions posed in this study (since in every SRT measurement, there should be unrecognized/missing words), it is required to obtain representative acoustic models for a robust ASR system. The utterances exhibit a low complexity level and can be recognized with a simple left-to-right ASR model without skips.

Incomplete sentences were recorded to simulate a listener's response during an actual measurement session when the tested person is aware he or she is interacting with an ASR system. The utterances contain one to four words with the same order of grammatical functions as in the original OLSA.

Additionally, a test set that very much resembles responses occurring in a **human-human interaction** when an OLSA measurement is performed was created. To compile a list of utterances, we analyzed audio recordings of actual OLSA measurements provided by HörTech gGmbH Oldenburg and searched for typical response formats, frequency of out-of-vocabulary (OOV) words, and other factors that could affect ASR. Both hearing-impaired and normal-hearing subjects usually participate in clinical measurements, some of which are familiar with the OLSA vocabulary while others are not. This was considered when selecting four representative patients who consented to being recorded. Based on the analysis, a realistic test list containing 500 utterances (mix of complete, incomplete, and OOV utterances) was created. 200 of those were transcripts of responses given during the original OLSA test.

The amount of required speech data was estimated in experiments based on the small vocabulary task Aurora 2 [8], in which we systematically increased the number of speakers/training utterances per word class and kept track of the word error rate of the Aurora 2 baseline system using Mel-Frequency Cepstral Coefficients (MFCCs) [3]. A probable application scenario for SRT testing is a relatively noise-free environment, since noise sources would not only hurt ASR performance, but also interfere with the SRT measurement. Therefore, clean training was performed with data from 2 to 110 speakers with a balanced number of male and female speakers in each training set. Each speaker produced each word approximately 23 times, i.e., the full training set corresponds to 2.5k word observations per word class. Error rates are reported in Figure 2 for clean and noisy utterances, where the WER for the noisy condition is an average of SNRs from 0 to 20 dB, as suggested in [8]. Scores for the clean test set were further analyzed by plotting the relative increase of word error rate related to the full test set with 110 speakers (lower panel in Figure 2). Compared to the

fully-trained system, the WER is increased by ca. 50% when only 460 word representations for each word class are used. For more than 1840 classes, the relative increase stays below 20%. Hence, we aim at 2000 word representations per class for the recordings of our database for autonomous SRT testing.

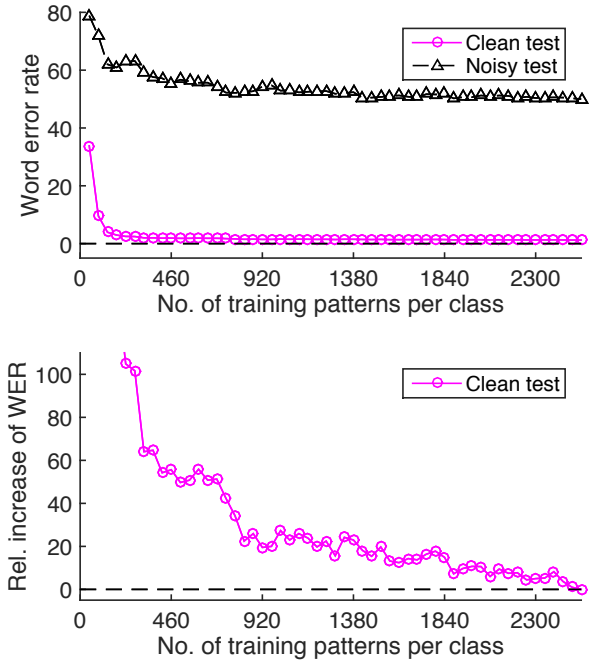

Figure 2: ASR error rates for an Aurora 2 system, trained with clean data from two to 110 speakers

20 native German speakers were recorded (10 male, 10 female, mean age 35.8 years, range: 21-60 years, standard deviation 15.3 years) in a sound-insulated booth designed for speech recordings using a Neumann KM184 microphone and an EDIROL UA-25 USB sound card recording at 44.1 kHz. For Set HH1, 17 of the 20 speakers recorded 100 sentences due to restricted recording time, while two speakers recorded the complete list of 500 sentences (Set HH2). A simple graphical user interface in Matlab was used to present sentences visually to speakers, and start and stop the actual recording that was performed with SoundMexPro, a toolbox for studio quality recordings in Matlab (HörTech GmbH, Oldenburg, Germany). Speakers were paid for their participation, and were instructed both verbally and in writing. On average, we collected 2200 recordings for each word in the OLSA test.

Set name	(label)	No. of utts.	No. of speakers	Duration (hrs)
Complete	(C)	15,730	20	15
Incomplete	(I)	5,998	20	4.6
Human-human interaction	(HH1)	1,675	17	1.2
	(HH2)	998	2	0.8
Total		27,170		23.2

Table 1: Properties of the recorded matrix test speech data.

2.3. Speech recognition system

ASR experiments were performed using a hidden Markov model (HMM) implemented in Kaldi [14]. The training procedure for the backend was as follows: A class initialization was performed with a monophone-based alignment using standard MFCC features with deltas and double-deltas. A refinement of temporal class labels was obtained from a triphone training step which used 2.5k HMMs corresponding to triphone classes and a total of 15k Gaussians. Finally, the refined classes are employed to train a triphone model with spliced MFCC features as input (using a total of 4+4 context frames added to the center frame, resulting in 117-dimensional features). Features were transformed using Maximum Likelihood Linear Transform and Linear Discriminant Analysis with a downprojection to 40-dimensional feature vectors. The backend used 3k Gaussians in 264 HMMs. Training of the acoustic model was performed following a *leave-one-out* procedure, i.e., speech data of 19 speakers was used for training, while the remaining data was used as test material. This was repeated for all 20 speakers for experiments with Sets C and I, respectively. When testing Set HH2, training was performed with Sets C, I, and HH1, since HH1 alone does not contain sufficient training data. A garbage model was used for OOV words. The division of data resulted in speaker-independent recognizers for all experiments.

3. Results

3.1. Overall results

ASR error rates obtained with complete (C) and incomplete (I) sentences without out-of-vocabulary words are presented in Figure 3. Two speakers produced error rates that were at least 8 to 14 times higher than average, and 5 times higher than the next best result. Since the audio material of these speakers sounded normally, we suspect a systematic error and excluded the data from further analysis. The average WERs for Set C (0.66% $\pm$ 0.39%) and Set I (1.26% $\pm$ 0.09%) are rather low, which was to be expected for a small vocabulary task when no additional sources of noise or reverberation exist, and also shows that the amount of training data estimated based on Aurora 2 experiments is sufficient to obtain this good performance. The error rates observed for recognition of Set HH2 was found to be far higher with 22.7%. The main difference to the other test sets is that many utterances from this set resembling human-human interactions contain OOV words and do not share the very simple grammatical structure of Sets C and I. The question if these errors affect an SRT result when ASR is used to control test parameters is addressed in the next section.

3.2. Effect of ASR errors on measured speech intelligibility

The effect of ASR errors on an SRT measurement depends on a specific sequence of trials. For instance, the SRT is influenced by potential changes of the SNR, which on their part depend on the extreme values of the measurement curve: When local extreme values occur late in the measurement, $f(m)$ in Eq. 1 is rather large, and the measurement curve cannot convergence quickly. Hence, we constructed a conservative example with very late local extrema that are rarely observed in measurements, but may result in continuous reduction of the SNR, thereby delaying the first local minimum (Figure 1). The first local extremum is assumed for the 5th trial, whereas in typical measurements it reached after the second or third trial.

Further, we assume that no transcription errors occur when

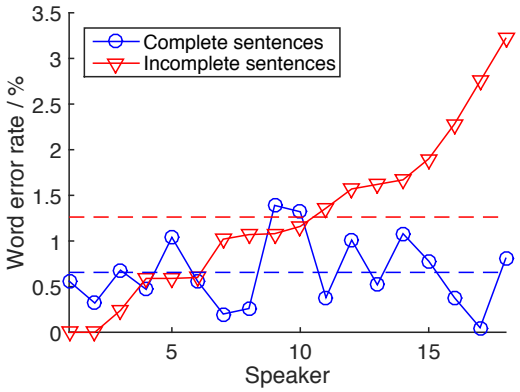

Figure 3: ASR error rates for complete and incomplete sentences. Dashed lines indicate average results for these test sets.

the test is conducted by a human listener, since the test subject gives his or her response in a acoustically clean condition rendering errors in human-human communication extremely unlikely: For a digit recognition task of similar complexity, human error rates of 0.5% were observed for 5 dB SNR and 0% error for clean condition [13].

ASR errors will affect the percentage of correct words to be reported during measurement, i.e., the term $P(k-1)$ in Eq. 1 that controls SNR changes. The deviation ΔP from the true percent correct value P depends on how often an error occurs (the WER), and what the effect of an error is (because a higher number of word errors results in larger SNR changes). This effect is approximated by the factor κ that is defined as the ratio of erroneous words in *misclassified* sentences. For Set C, the average number of word errors for incorrect sentences was $\kappa = 1.47$, for Set I $\kappa = 1.16$ was found, and for Set HH2 we measured $\kappa = 1.98$. The maximum value of ΔP should be 1, corresponding to the range or $P(k-1) - T$ in Eq. 1, which is achieved with an normalization factor of 0.2 for five-word sentences. Hence we obtain

$$\Delta P = 0.2 \cdot \text{WER} \cdot \kappa.$$

For our conservative example in Figure 1, the worst-case scenario is that the ASR-induced errors have the same sign and sum up to the maximum measurement error ΔSRT:

$$\Delta\text{SRT} = \sum_{k=1}^{K} \frac{f(m) \cdot \Delta P}{s},$$

where the index m is implicitly increased after each local extremum (cf. shades of grey in Figure 1). With a standard value $K = 30$ and the definition of $f(m)$ given above, we obtain $\Delta\text{SRT} = 18.5 \cdot \Delta P/s = 18.5 \cdot 0.2 \cdot \text{WER} \cdot \kappa/s$. If our goal is to achieve the same accuracy obtained with supervised measurements, it follows that $\Delta\text{SRT} \leq 0.5\,\text{dB}$ and consequently $\text{WER} \cdot \kappa \leq 2.03\%$.

We compare our experimental data to this target value and find that WER $\cdot\ \kappa$ is sufficiently small for Sets C and I (with values of 0.97% and 1.46%, respectively) to obtain a very reliable estimate of the SRT. For the data set that resembles human-human interaction, we obtain $\text{WER} \cdot \kappa = 44.9$, i.e., the required reliability of SRT measurement cannot be reached with a maximal potential error of 11 dB.

4. Summary and discussion

In this paper we investigated the potential use of ASR for autonomous measurement of the speech reception threshold (SRT), which is one of the most important measures for diagnosis of hearing impairment and for optimizing hearing aid settings. A very accurate method for SRT measurement is the Oldenburg sentence matrix test OLSA that is typically conducted by an audiologist logging the subject's responses. For an ASR-based test conduction, it is important to quantify the errors that occur for test responses. To this end, we recorded a speech database specifically for this study. Data was collected from 20 speakers and with different levels of complexity that either exactly follow the grammatical structure of matrix sentences (with possible skips over words, which is relevant for responding to partially masked sentences) or simulate responses given during a regular supervised measurement in a human-human interaction. The overall recognition rates for complete and incomplete sentences without out-of-vocabulary words were very low, with only 0.66% and 1.26%, respectively. For speech material that closely resembles test scenarios with a human supervisor however, the error rate was far higher with 22.7%.

There are several methods that could potentially lower ASR error rates, e.g., by adapting the system to the current speaker by maximum a posteriori (MAP) adaptation [4] or maximum likelihood linear regression (MLLR) [15]. Additionally, ASR training could be performed with a database suitable for large vocabulary continuous speech recognition; although one design criterion of the OLSA test was a representative phoneme distribution, it is likely that the identification of less frequent phones could profit from a larger amount of training data. This is especially important for recognizing out-of-vocabulary words (OOV) that contain phones with a low *a-priori* probability of occurence. At the same time, this would allow to move away from using a garbage model for OOV words, which should result in a strong decrease of WER for Set HH2.

Our post-analysis has shown that the current ASR system's performance is not sufficient for conducting SRT measurements with Set HH2. However, our analysis, which is based on a very conservative example for a complete measurement with 30 sentences, has shown that ASR is generally suitable for autonomous SRT measurements for test sets without OOV words. In fact, we have shown that a very low test-retest standard deviation below 0.5 % can be achieved, which is in the same range that is obtained with a human test supervisor [21].

In future work, our findings could be combined with a dialogue system that provides feedback to the user and hence guides him or her through the testing procedure. For telephone-based testing, additional experiments need to be performed for analyzing the effect of monaural signal presentation via telephone channels. A huge advantage for the matrix test under consideration combined with ASR-based autonomous testing via telephone is the availability of the original test in 14 languages, which has a high potential of improving widespread access to an important diagnostic tool in audiometry.

5. Acknowledgements

This work was funded by the DFG (Cluster of Excellence 1077/1 Hearing4All (http://hearing4all.eu), and the SFB/TRR 31 "The Active Auditory System" (http://www.sfb-trr31.uni-oldenburg.de/)). We thank Constantin Spille, Angel Mario Castro Martinez, Sabine Hochmuth, and Jörg-Hendrik Bach for valuable discussions and help with experimental details.

6. References

[1] Brand, T., Kollmeier, B. (2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," Journal of the Acoustical Society of America 111, pp. 2801-2810.

[2] Buschermöhle, M., Wagener, K., Berg, D., Meis, M., Kollmeier, B. (2014). "The German Digit Triplets Test (Part I): Implementations for Telephone, Internet and Mobile Devices," Zeitschrift für Audiologie 53, pp. 139-145.

[3] Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, 28 (4), pp. 357-366.

[4] Gauvain, J., Chin-Hui, Lee (1994). "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Transactions on Speech and Audio Processing 2 (2), pp. 291-298.

[5] Heger, D., and Holube, I. (2010). "Wie viele Menschen sind schwerhörig (How many people are hearing-impaired)?," Zeitschrift für Audiologie, 49, pp. 61-70.

[6] Hewitt, D.R. (2008). "Evaluation of an English speech-in-noise audiometry test," Faculty of Engineering, Science and Mathematics Institute of sound and vibration research, University of Southampton.

[7] Hagerman, B., Kinnefors, C. (1995). "Efficient adaptive methods for measuring speech reception threshold in quiet and noise," Scand Audiol 24(1), pp. 71-77.

[8] Hirsch, H. G., and Pearce, D. (2000). 'The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," Proc. Autom. Speech Recognit. Challenges for the new Millenium, pp. 29-32.

[9] Hochmuth, S., Brand, T., Zokoll, M.A., Zenker Castro, F., Wardenga, N., Kollmeier, B. (2012). "A Spanish matrix sentence test for assessing speech reception thresholds in noise," Int J Audiol, 51, pp. 536-544.

[10] Houben, R., Koopman, J., Luts, H., Wagener, K.C., van Wieringen, A., Verschuure, H., Dreschler, W.A. (2014). "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," Int J Audiol. 53 (10), pp. 760-763.

[11] Lin, F.R., Niparko J.K., Ferrucci L. (2011). "Hearing loss prevalence in the United States," Letter Arch Intern Med. 171(20), pp. 1851-1852.

[12] Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V. N., Brand, T., Wagener, K. C. (2015). "The multilingual matrix test: principles, applications and comparison across languages - a review," International Journal of Audiology, accepted.

[13] Meyer, B.T. (2013). "What's the difference? Comparing humans and machines on the Aurora2 speech database," in Proc. Interspeech 2013, pp. 2634-2638.

[14] 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 Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, pp. 1-4.

[15] Povey, D., Yao, K. (2011). "A Basis Method for Robust Estimation of Constrained MLLR," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4460-4463.

[16] Puglisi, G.E., Astolfi, A., Prodi N., Visentin, Ch., Warzybok, A. (2014). "Construction and first evaluation of the Italian Matrix Sentence Test for the assessment of speech intelligibility in noise," in Proc. of Forum Acusticum, Krakow, Poland.

[17] Psutka, J.V.; Müller, Ludek (2007). "Comparison of various feature decorrelation techniques in automatic speech recognition", Journal of Systemics, Cybernetics and Informatics 5, pp. 27-30.

[18] Schroeder, J., Wabnik, S., van Hengel, P. W., Goetze, S. (2011). "Detection and classification of acoustic events for in-home care", in Ambient Assisted Living, Springer Berlin Heidelberg, pp. 181-195.

[19] Wagener, K., Kühnel, V., Kollmeier, B. (1999). "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," Zeitschrift für Audiologie 38 (1), pp. 1-32.

[20] Wagener, K. C., Brand, T., and Kollmeier, B. (1999). "Development and evaluation of a German sentence test Part II: Optimization of the Oldenburg sentence test," Zeitschrift für Audiol., 38, 44-56.

[21] Wagener, K. C., Brand, T., and Kollmeier, B. (1999). "Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test," Zeitschrift für Audiologie (38), pp. 86-95.

[22] Wagener, D. and Brand, T. (2005). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters," International Journal of Audiology 44, pp. 144-156.