INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Single-ended prediction of listening effort based on automatic speech recognition

Rainer Huber, Constantin Spille, Bernd T. Meyer

Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

rainer.huber@uni-oldenburg.de, constantin.spille1@uni-oldenburg.de
bernd.meyer@uni-oldenburg.de

Abstract

A new, single-ended, i.e. reference-free measure for the prediction of perceived listening effort of noisy speech is presented. It is based on phoneme posterior probabilities (or posteriorgrams) obtained from a deep neural network of an automatic speech recognition system. Additive noisy or other distortions of speech tend to smear the posteriorgrams. The smearing is quantified by a performance measure, which is used as a predictor for the perceived listening effort required to understand the noisy speech. The proposed measure was evaluated using a database obtained from the subjective evaluation of noise reduction algorithms of commercial hearing aids. Listening effort ratings of processed noisy speech samples were gathered from 20 hearing-impaired subjects. Averaged subjective ratings were compared with corresponding predictions computed by the proposed new method, the ITU-T standard P.563 for single-ended speech quality assessment, the American National Standard ANIQUE+ for single-ended speech quality assessment, and a single-ended SNR estimator. The proposed method achieved a good correlation with mean subjective ratings and clearly outperformed the standard speech quality measures and the SNR estimator.

Index Terms: automatic speech recognition, deep neural networks, listening effort prediction

1. Introduction

The effort required to listen to and understand noisy speech is an important factor in the evaluation of noise reduction (NR) schemes. While speech intelligibility is often not improved by (especially single-channel) NR schemes or already near 100% before NR, the listening effort can still be affected by reduced noise levels, even at (more realistic) positive SNRs, where speech intelligibility is at or near 100% (e.g., [1], [2]). Consequently, Rennies et al. [2] conclude that “intelligibility is an insensitive measure to evaluate many everyday listening conditions” and “in some conditions, listening effort is a more sensitive measure of speech perception than intelligibility.”

Listening effort can be measured subjectively, e.g. by using rating scales, or objectively by using indirect measures such as memory performance, reaction times, pupil dilatation or other physiological measures (see [2] for an overview). However, all of these measurement methods are time consuming and costly. Hence, computational models to predict listening effort are of high interest. However, up to now, no models are known to the authors that have specifically been designed to predict listening effort. Schepker et al. [3] and Rennies et al. [2] found a rough correlation between subjectively measured (i.e. perceived) listening effort and the Speech Transmission Index (STI). Huber et al. [4] found high correlations between some output measures of the double-ended (i.e. reference-based) audio quality model PEMO-Q [5] and listening effort ratings of hearing-impaired subjects. Unfortunately, double-ended quality models need a clean or nearly-clean reference signal, which is often not available. Hence, the present study investigates the possible qualification of single-ended speech quality models for the prediction of perceived listening effort. It has been found that speech quality and listening effort are highly correlated [6], hence speech quality measures could also be qualified to predict listening effort. Consequently, two established standard methods for single-ended speech quality assessment, namely, the ITU-T standard P.563 [7] and the American National Standard ANIQUE+ [8] have been evaluated with regard to listening effort prediction. Moreover, a new single-ended method is proposed that is based on automatic speech recognition (ASR).

The proposed new method is motivated by recent advances in the field of ASR, which were enabled by the advent of deep learning: A few years ago, the gap between human speech recognition (HSR) and ASR was reported to be 15dB, i.e., ASR reached human accuracy only when the signal-to-noise ratio was increased by 15dB. For conversational telephone speech, this gap has been recently closed by using deep neural networks (DNNs) in elaborate ASR systems, which achieve the same word error rate as human transcribers [9]. The increased performance indicates that robustness to noise or artifacts should be in the same range for human and machine listeners, and hence representations derived from DNNs could be useful for predicting speech quality.

ASR technology has been applied before for prediction of speech *intelligibility* by combining Gaussian mixture models with Hidden-Markov-Model (HMMs) [10], [11], [12]. However, these approaches are reference-based, either because glimpses above the noise floor need to be identified [10] or because the SNR of training data is a free variable, which can only be chosen if separate noise and speech are available [11], [12].

Our novel method uses DNN-based classifiers for single-ended listening effort prediction and therefore differs from all existing methods known to the authors. As in state-of-the-art ASR, a DNN is trained to classify the current phoneme given the input feature. The resulting output over time is referred to as phoneme posterior probabilities (or “posteriorgrams”), which are expected to degrade in suboptimal conditions, e.g., in the presence of noise, reverberation, or other signal distortions. The amount of degradation is estimated with performance measures, which have been applied before to predict the word error rate of ASR systems [13] or to select features in a multi-stream ASR system [14].

In the following, a measure to quantify the posteriorgram degradation as a predictor of listening effort will be introduced.

http://dx.doi.org/10.21437/Interspeech.2017-1360

Correlations with subjective quality ratings obtained from the evaluation of noise reduction schemes in hearing aids will be analyzed. The performance of the proposed method will be compared with current standard methods.

2. Methods

2.1. ASR system and training data

The ASR system was trained using the standard kaldi DNN recipe for Aurora 4 [15]: The DNN used six hidden layers, 2048 units per layer, and an additional softmax output layer. It was pre-trained as a restricted Boltzmann machine using contrastive divergence (CD-1) and supervised fine-tuning with the triphone targets via cross entropy. Every phone was modeled with three Hidden Markov Model (HMM) states except for the silence phone which was modeled with five states. 40-dimensional Mel-filterbank features were extracted from the 16 kHz audio data and fed to the DNN using an additional temporal context of 5+5 frames, resulting in 440-dimensional input to the neural net. Posteriorgrams were derived from the activations of the softmax output layer. Monophone posteriorgrams were obtained by grouping all triphones belonging to the same phone and subsequent summation of the corresponding activations. This was done since we found monophone processing to give similar results than its triphone equivalent at a lower computational cost. The training data uses the noise types also used for listening experiments (speech-simulating and airplane cabin noise, see Section 3), from which a multi-condition training set was compiled. We added noise to the clean Aurora 4 training data (7138 utterances, read speech, vocabulary size: 5,000) with SNRs drawn randomly from an interval that was found to be relevant for listening effort in earlier research (speech-simulating noise: 0 to 23 dB, airplane cabin noise: -9 to 14 dB). With two noise types, we obtain 14,276 noisy utterances used to train the acoustic model.

2.2. Performance measure

From the posteriorgrams, the mean temporal distance (or M-Measure) as proposed by Hermansky et al. [16] is computed. The M-Measure computes the average difference between two vectors of phoneme posteriors $p_{t-\Delta t}$ and p_t (i.e., two columns of the posteriorgram) with a temporal distance Δt:

$$M(\Delta t) = \frac{1}{T - \Delta t} \sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t)$$

with T being the temporal length of the analyzed posteriorgram, and D being the Kullback-Leibler divergence. In the present study, M is computed for Δt =35 to 80 ms (in 5 ms steps) and averaged.

Figure 1 shows posteriorgrams of clean speech (top panel) and of speech in noise at 5 dB SNR (bottom panel) of the same utterance. Additive noise and other distortions tend to smear the posteriorgram and make it more homogeneous along the time axis. Hence, the difference between different vectors of phoneme posteriors and consequently the M-Measure decrease. A more systematic illustration of the effect of distortions by additive noise on the M-Measure is shown in Figure 2, where $M(\Delta t)$ curves of speech in noise at different SNRs are plotted. “Washed out” phoneme activations by additive noise results in a higher similarity between distant posterior frames. This decreases the Kullback-Leibler divergence between frame vectors, and consequently noisy curves lie below curves with a higher SNR.

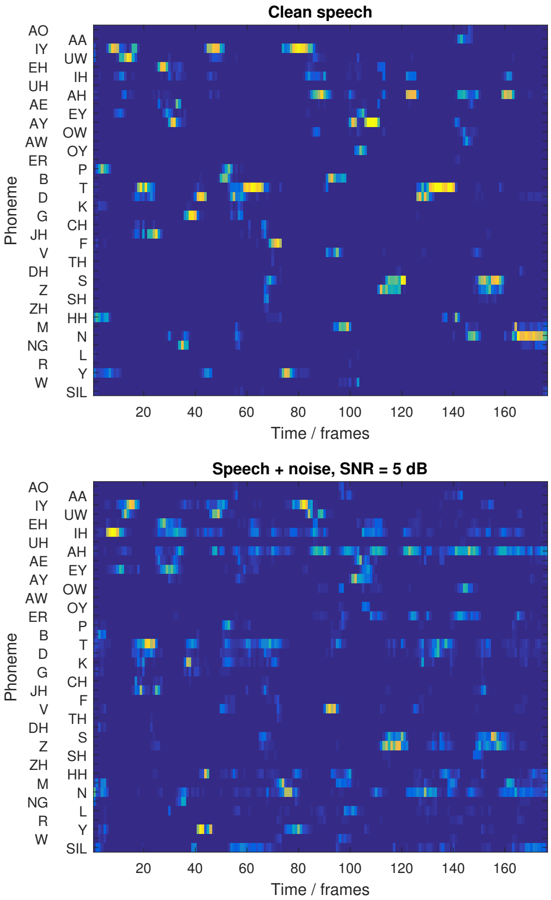

Figure 1: *Phoneme posterior probabilities (“posteriograms”) of clean speech (top) and speech in noise at 5 dB SNR (bottom) of the same utterance. The corresponding M-Measure is 15.4 for the clean speech sample and 7.6 for the noisy speech sample.*

3. Listening effort database

Speech signals and corresponding subjective listening effort ratings were obtained from a study carried out by the Hörzentrum Oldenburg [4]. In this study, the performances of the noise reduction programs of four commercial hearing aids were benchmarked. To this end, male speech taken from the Oldenburg sentence test [17] was mixed with (A) stationary speech-simulating noise taken from the same sentence test (OLNOISE) and (B) stationary airplane cabin noise at six different SNRs (speech-simulating noise: SNR = -1, 2, 5, 8, 11, 14 dB; airplane cabin noise: SNR = -10, -7, -4, -1, 2, 5 dB[1]). These source signals were processed by four commercial hearing aids, which were fitted to the average hearing loss of 20 moderately hearing-impaired test subjects (mean pure tone average = 58 dB; age 26-83, median=71), all native German speakers. Each hearing aid processed the noisy speech samples two times: with activated and with de-activated noise reduction program. The

[1] The sets of SNRs were derived from pre-tests, covering a listening effort range from “extremely effortful” to “effortless”.

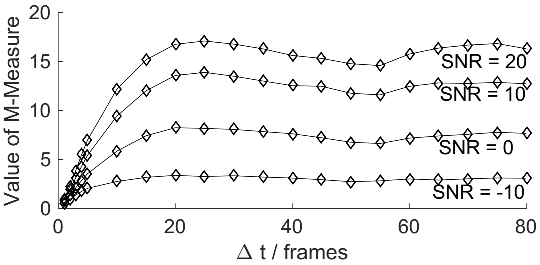

Figure 2: *M-Measure values for the same utterance at different SNRs. The temporal smearing of phoneme activations results in more similar posteriorgram frames for high Δt, i.e., we obtain smaller M-Measure values for noisy data.*

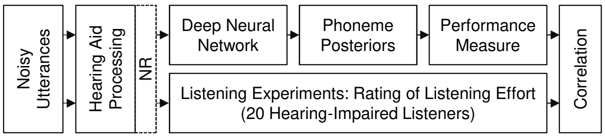

Figure 3: *Illustration of our analysis of listening effort prediction. The noise reduction (NR) program in the hearing aid is either switched on or off and hence illustrated with a dashed box.*

hearing aid output signals were recorded with an artificial head (KEMAR) with ear simulator. The recordings were filtered individually to have the same long term spectrum, so that the different hearing aids had similar frequency responses and only differed with regard to the effects of the individual noise reduction programs. In total, 96 test signal were generated (6 SNRs x 4 hearing aids x 2 noise reduction settings (on/off) x 2 noise types). The filtered recordings were presented to the unaided subjects via headphones (Sennheiser HDA200) with the original hearing aid output level. Amongst other measurands, the test subjects rated the perceived listening effort using a 13-step graphical rating scale with seven verbal categories: “extrem anstrengend” (extremely effortful), “sehr anstrengend” (very effortful) “deutlich anstrengend” (considerably effortful), “mittelgradig anstrengend” (middle effortful), “wenig anstrengend” (little effortful), “sehr wenig anstrengend” (very little effortful), “mühelos” (effortless). The subjects entered their responses on a touch screen. The rating results were averaged across subjects to build the Listening Effort Mean Opinion Score (LE-MOS).

4. Results

Posteriorgrams of the hearing aid recordings were computed by the ASR system described earlier and the M-Measure was calculated for all 96 posteriorgrams (cf. Figure 3). Moreover, the standard speech quality measures P.563 and ANIQUE+ were computed for the same signals, as well as a “blind” SNR estimator [18]. The outcome of these measures are compared with the averaged subjective listening effort ratings in scatterplots shown in Figure 4. The results are summarized by correlations between all objective measures (i.e. M-Measure, P.563, ANIQUE+ and SNR estimator) and subjective scores in Table1.

Figure 4, left panels, shows mean subjective listening effort ratings (LE-MOS) vs. corresponding values of the M-Measure for speech mixed with OLNOISE (upper left panel). A high negative correlation of $r = -0.93$ is obtained. (The correlation is negative, because listening effort decreases with higher SNR, whereas the M-Measure increases with higher SNR; cf. example shown in Figure 1.) The relation between LE-MOS and M-Measure is somewhat curvilinear. Consequently, if a second-order polynomial fit is applied, the correlation between LE-MOS and the fitted M-Measure values increase to 0.96. The rank correlation is very high, too ($rs = -0.95$), and the standard deviation between LE-MOS and a linear (non-linear) fit of the M-Measure data amounts to just 0.94 (0.76). The correlation achieved with the same measure for the airplane cabin noise (lower left panel) is not quite as high as for the OLNOISE, but still very good ($r = -0.92$). The relation between LE-MOS and M-Measure data is approximately linear in this case, so no non-linear fit was applied.

The results obtained with the standard speech quality measures P.563 and ANIQUE+ are very poor (middle panels) because a large portion of the predicted MOS values are floored at the lower end of the MOS scale, i.e., the qualities of the noisy signals are outside (below) the normal operation range of these measures. In case of the ITU-T standard P.563, a number of signals could not be processed successfully (indicated by the software returning a MOS value of -1; those results were excluded from the data inspected here).

The single-ended SNR estimation method achieves a decent correlation with LE-MOS data if a non-linear fit is applied to the data and as far as OLNOISE is concerned ($r = 0.85$, see upper right panel). However, for airplane cabin noise, the linear correlation is just 0.7 (lower right panel).

The comparison with the standard speech quality measures and the SNR estimator in Table 1 shows that the prediction performance of the proposed new measure is clearly superior to the performances of the other measures.

Table 1: *Pearson correlations between objective measures and mean subjective listening effort ratings. (The numbers in brackets apply if a non-linear fit (2nd-order polynomial) is applied.)*

	M	P.563	ANIQUE+	SNR est.
OLNOISE	0.93 (0.96)	0.63	0.60	0.77 (0.85)
airplane noise	0.92	0.69	0.02	0.70

5. Discussion

The prediction of listening effort based on DNN-based ASR systems achieves very good correlations when combined with a suitable performance measure. It appears that methods borrowed from ASR become more and more useful in HSR research now that the overall performance gap between humans and machines get smaller (or vanishes for single, well-studied databases [9]). In our experiments, the best correlations are obtained with the M-Measure, which was shown earlier to be clearly related to parameters that influence speech intelligibility in hearing aids, e.g., the optimal direction of a beamformer when spatial filtering is performed in multi-channel hearing aids [19]. This is one example of strategies developed for ASR (specifically stream-weighting in multi-stream ASR) which has a meaningful application in human speech perception (specifically hearing research), as advertised in [20].

The proposed approach may be applied for accelerated research and development of speech enhancement (SE) in hear-

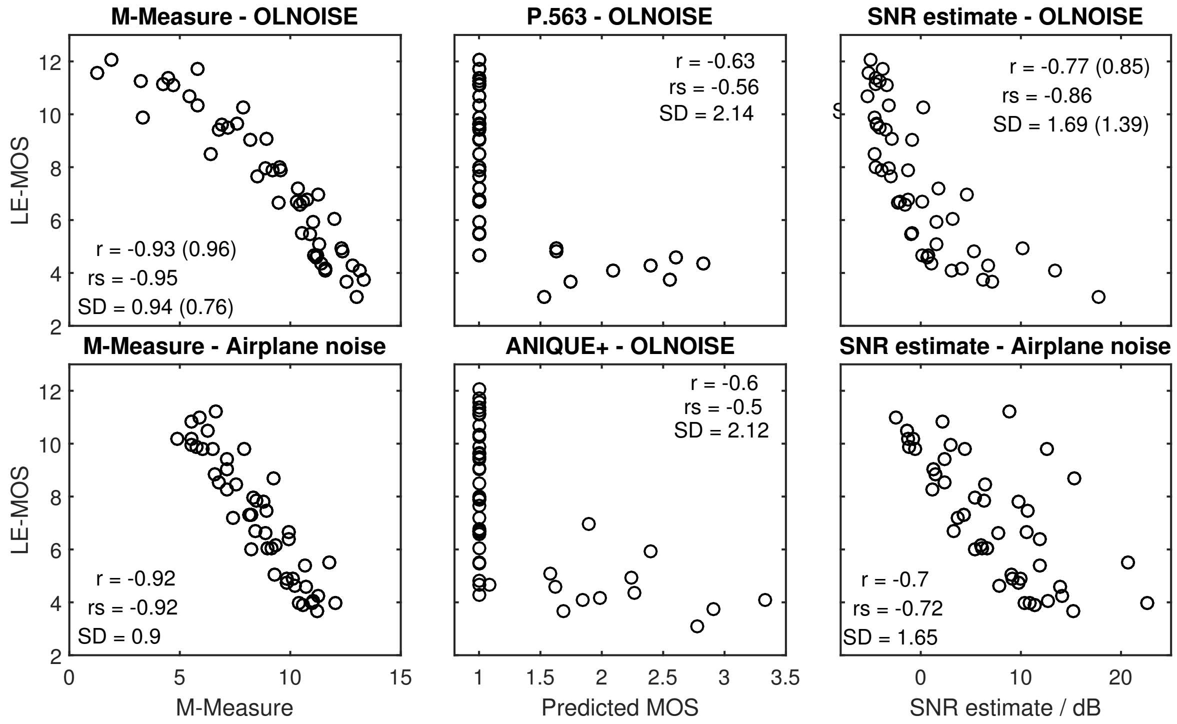

Figure 4: *Scatter plot of mean subjective listening effort ratings (Listening Effort Mean Opinion Scores - LE-MOS) vs. objective performance measures (left: M-Measure; middle: standard speech quality measures; right: SNR estimator) for the speech signals mixed with speech simulating noise (OLNOISE) or airplane cabin noise. r: linear correlation coefficient; rs: rank correlation coefficient after Spearman; SD: standard deviation from a fit. The bracketed values apply if a 2nd order polynomial fit is applied.*

ing aids. For instance, the evaluation of SE for a large parameter space should be feasible with the model, since a forward run with the DNN acoustic model is computationally relatively cheap. A limited number of optimized parameter sets could then be investigated in listening experiments.

A related application scenario of our algorithm is the use of a *model in the loop*, which constantly monitors hearing aid signals and selects the SE algorithm/parameter set optimal for the current acoustic scene. This would require running a DNN-classifier on hearing aid hardware in real-time. As estimated in [19], a forward run of a standard DNN as used in our experiments is not possible on current hearing aid hardware due to limitations in power consumption. However, when the model complexity is reduced by a factor of 10, such real-time processing becomes feasible. Considering the good results obtained with considerably smaller nets [21] and the fact that our measure is based on monophone activations (in contrast to high-dimensional triphone activations), this reduction seems within reach, which will be subject of future research.

A further limitation that needs to be overcome for this application is the fact that the multi-condition training set used noise types that are later encountered in testing, i.e., although the acoustic model has not seen the test *speech* signals during training, it has been exposed to the masker. Two potential remedies for this are (A) to create multi-condition training sets with a large number of very different maskers (which could be expected to generalize to other maskers) or (B) to estimate noise properties from speech pauses that could be used for online adaptation of the model, which also will be investigated in the future.

6. Conclusions

From the results presented, the following conclusions can be drawn:

- Single-ended listening effort prediction based on a DNN-based ASR system is possible with high accuracy.
- Standard single-ended quality measures are not qualified to predict listening effort of noisy speech.
- Listening effort depends not only on the SNR, but also on other factors such as type of noise and speech distortions by, e.g., processing artifacts.
- Single-ended listening effort prediction methods can be used as *models in the loop* in hearing devices if the computational complexity can be reduced.

7. Acknowledgements

This work was funded by the Cluster of Excellence 1077/1‘ “Hearing4all”. The authors would like to thank the Hörzentrum Oldenburg and Sivantos for kindly providing their data, and four anonymous reviewers for helpful comments on an earlier version of this paper.

8. References

[1] M. Krüger, M. Schulte, and I. Holube, "Entwicklung einer adaptiven Skalierungsmethode zur Ermittlung der subjektiven Höranstrengung," in *Proceedings of the 18th Annual Conference of the German Society for Audiology*, Bochum, Germany, 2015.

[2] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," *J. Acoust. Soc. Am.* vol. 136, p. 2642, 2014, doi: 10.1121/1.4897398

[3] 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, doi: 10.1080/14992027.2016.1219774

[4] R. Huber, M. Schulte, M. Vormann, and J. Chalupper, "Objective measures of speech quality in hearing aids: Prediction of listening effort reduction by noise reduction algorithms," in *2nd Workshop on Speech in Noise: Intelligibility and Quality, Amsterdam, The Netherlands*, 2010. Available: http://www.phon.ucl.ac.uk/events/quality2010/talks/RainerHuber.pdf

[5] R. Huber and B. Kollmeier, "PEMO-Q - A new Method for Objective Audio Quality Assessment using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech and Language processing*, vol. 14, no. 6, pp. 1902-1911, 2006.

[6] R. Huber, T. Bisitz, T. Gerkmann, J. Kiessling, H. Meister, and B. Kollmeier, "Comparison of single-microphone noise reduction schemes: can hearing impaired listeners tell the difference?" *Int J Audiol.* 2017 Jan 23:1-7. doi: 10.1080/14992027.2017.1279758. [Epub ahead of print], 2017

[7] ITU-T, "Single-ended method for objective speech quality assessment in narrow-band telephony applications", Recommendation P.563, International Telecommunication Union, Geneva, Switzerland, 2004.

[8] D.-S. K im and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221-236, 2007.

[9] W. Xiong, J. Droppo, X.Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," arXiv:1610.05256v1, 2016.

[10] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, pp. 1562-1573. doi:10.1121/1.2166600, 2006

[11] B. T. Meyer and B. Kollmeier, "Learning from human errors: Prediction of phoneme confusions based on modified ASR training", in *INTERSPEECH 2010 - 11th Annual Conference of the International Speech Communication Association, September 26-30, Makuhari, Japan, Proceedings*, 2010.

[12] M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier, "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception", *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2708-2723, 2016.

[13] B. T. Meyer, S. H. Mallidi, H. Kayser, and H. Hermansky, "Predicting error rates for unknown data in automatic speech recognition," in *ICASSP 2017 - 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, New Orleans, USA, Proceedings*, 2017.

[14] S. H. Mallidi, T. Ogawa, and H. Hermansky, "Uncertainty estimation of DNN classifiers," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2015.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, ASRU*, pp. 1-4, 2011.

[16] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013

[17] K. C. Wagener, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 4-15, 1999.

[18] F. Denk, J. P. C. L. da Costa, and M. A. Silveira, "Enhanced forensic multiple speaker recognition in the presence of coloured noise," in *Proceedings of the 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*, IEEE, 2014.

[19] B. T. Meyer, S. H. Mallidi, A. M. Castro Martnez, G. Paya-Vaya, H. Kayser, and H. Hermansky. "Performance monitoring for automatic speech recognition in noisy multi-channel environments," *IEEE Workshop on Spoken Language Technology*, 2016

[20] O. Scharenborg. "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, 49, pp. 336-347, 2007

[21] T. Nagamine, M. L. Seltzer, and N. Mesgarani. "On the Role of Nonlinear Transformations in Deep Neural Network Acoustic Models," in *INTERSPEECH 2016 - 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016