Interspeech 2018
2-6 September 2018, Hyderabad

Prediction of Subjective Listening Effort from Acoustic Data with Non-Intrusive Deep Models

Paul Kranzusch[1,3], *Rainer Huber*[2,3], *Melanie Krüger*[4], *Birger Kollmeier*[1,3], *Bernd T. Meyer*[1,3]

[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Fraunhofer IDMT - Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Germany
[4]Hörzentrum Oldenburg GmbH, Oldenburg, Germany

paul.kranzusch@uol.de, rainer.huber@idmt.fraunhofer.de, bernd.meyer@uol.de

Abstract

The effort of listening to spoken language is a highly important perceptive measure for the design of speech enhancement algorithms and hearing-aid processing. In previous research, we proposed a model that quantifies the phoneme output probabilities obtained from a deep neural net (DNN), which resulted in accurate predictions for unseen speech samples. However, high correlations between subjective ratings and model output were observed in known noise types, which is an unrealistic assumption in real-life scenarios. This paper explores non-intrusive listening effort prediction in unseen noisy environments. A set of different noise types are used for training a standard automatic speech recognition (ASR) system. Model predictions are produced by measuring the mean temporal distances of phoneme vectors from the DNN. These are compared to subjective ratings of hearing-impaired and normal-hearing listener responses from three databases that cover a variety of noise types and signal enhancement algorithms. We obtain an average correlation of 0.88 and outperform three baseline measures in most conditions.

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

1. Introduction

Listening effort (LE) is a perceptual measure that is highly relevant for speech perception in the context of speech enhancement and transmission, both for normal-hearing (NH) and hearing-impaired (HI) listeners [24]. It can be assessed with subjective listening tests, surveys and with non-acoustic measurements since it is related to several physiological factors such as heart rate variability [18], memory performance [25], or pupil diameter [16, 15]. Since these measurements are extensive, computational models for estimating the perceived LE from acoustic data (and especially speech) are very desirable. However, relatively few studies have so far explored this topic. A rough correlation between LE and the speech transmission index (STI) [8] was reported in [26, 24]. A high correlation was observed when analyzing the effect of single-channel noise reduction schemes on HI listeners between LE and the audio quality model PEMO-Q [10]. However, both the STI and PEMO-Q are intrusive (or double-ended models) and therefore require a clean speech reference signal that is usually not available in real-life scenarios. Such models are principally not suitable for predicting perception as model-in-the-loop that could for instance be used for continuous parameter optimization in hearing aids with the aim of decreasing LE.

Recently, a model for LE prediction was therefore proposed on the basis of an automatic speech recognition (ASR) system [12], which does not require the clean speech signal. The system is trained as regular ASR system; when applied as LE model, the degradation of phoneme posterior probabilities of a deep neural net (DNN) are quantified using the mean temporal distance (MTD) as introduced earlier for of multi-stream ASR [6, 19]. The model — which is referred to as LEAP (LE prediction from Acoustic Parameters) — was shown to produce accurate predictions for different signal enhancement strategies in commercial hearing aids. When analyzing model prediction for three different databases, it was shown that good model predictions were obtained for HI and NH listeners in different types of noises and speech enhancement strategies [11]. In contrast to many classic models of speech perception, this model has the advantage of being blind with respect to speech signals (since it is based on speaker-independent ASR). However, accurate LE predictions were only obtained when training and test noise were matched. This is a severe limitation of the model when considering its use as model-in-the-loop, since it requires separate models for each noise type and acoustic scene, which is not feasible for real-life systems.

This paper addresses this limitation by exploring non-intrusive models that do not require a speech or noise reference for LE prediction. Similar to previous work [12, 11] we test the hypothesis that phoneme representations obtained from a DNN are informative about signal degradation (and therefore listening effort) as quantified by the mean temporal distance. In contrast to previous work, a multi-condition training set is designed that is based on the Wall Street Journal corpus [4] and is completely independent from the stimuli used in subjective LE experiments. The number of training noise types is systematically increased while the amount of training data is kept constant to isolate the effect of different maskers. The resulting ASR systems are used to produce LE predictions that are compared to subjective listening data from three datasets that cover NH and HI listeners, stationary and modulated noise types, and various speech enhancement strategies. Earlier studies reported high correlation between LE and perceived speech quality (SQ) [10]. Since non-intrusive models of LE prediction are not available, we use current non-intrusive SQ models, i.e., the ITU-T standard P.563 [13] and the American National Standard ANIQUE+ [14] as baseline models as proposed in [12].

In the following, we describe the structure of the LEAP model, the training data sets and the subjective listening tests. Section 4 compares the correlation values of different model implementations to speech quality models as well as a signal-to-noise (SNR) estimator [1] that serves as additional baseline before results are discussed and the paper is summarized.

10.21437/Interspeech.2018-1375

2. Methods

2.1. Model Description

The model explored in this paper is based on the LEAP model first proposed in [12] that was shown to predict listening effort for several speech enhancement algorithms [11]. At its core, the model as illustrated in Figure 1 is based on an ASR hybrid system that uses regular mel spectrogram features as input to a DNN that serves as acoustic model.

The training procedure for the DNN is performed with the open source training toolkit Kaldi [23] according to the standard approach as described in [29]. The DNN contains six hidden layers with each 2048 hidden units. For initialization the net is trained as a deep belief network [20] as described in [7]. In the second step, a backpropagation with a learning rate of 0.008 is used initially and halved whenever the improvement is below 0.01. The training stops after 20 epochs or when relative improvements below 0.0001 are reached. The cross-entropy cost function is computed for an independent development set. The output of the DNN consists of 2026 units which represent phoneme posterior probabilities. The context-dependent triphone activations obtained from the net are decoded by a hidden Markov model (HMM) during training. To predict subjective listening effort, the quality of the DNN output is analyzed, which is based on the assumption that high-quality phoneme representations correspond to low LE and vice versa. As quality (or performance) measure, the mean temporal distance (MTD) [6] is applied to the phoneme posterior probabilities over time (the phoneme posteriorgram). It takes into account the average distance $\mathcal{D}(p(t), p(t-\Delta t))$ between phoneme vectors p with a temporal distance Δt, with the symmetric Kullback-Leibler divergence $\mathcal{D}$:

$$M(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=\Delta t}^{T} \mathcal{D}(p(t-\Delta t), p(t)) \qquad (1)$$

For degraded acoustic stimuli, the posteriorgram activations are often smeared over time, hence distant frames are relatively similar. On the other hand, clean utterances result in clear activations which are rather different on average when they are more than 200 ms apart (a time scale that was linked to coarticulation in [6]). To obtain a scalar value that can be compared to LE, we first calculate the mean distance of vectors with a Δt range in which coarticulation effects can be neglected, i.e., 350 to 800 ms in steps of 50 ms, and average the resulting values, which we refer to as MTD in the following.

2.2. ASR Training Data

The basis training data consists of 7,138 utterances produced by 83 speakers from the SI-84 Aurora4 data set with a vocabulary size of 5000 words [21]. The size of the set is increased by a factor of three by mixing each utterance with three different types of noise from a set as shown in Table 1 which are separated from test noises. The noise signals have a duration from 22 seconds to 4 minutes. One exception is the stationary speech-shaped noise (SSN) that was created for this study and exhibits the same long-term spectrum as the Aurora4 subset 020_16k and a duration of 60 minutes. The number of noise types as well as the following experimental parameters were chosen since they resemble the design decisions for Dataset 3 in our earlier study [11], which should provide comparable results. For each utterance, the starting sample of the noise type as well as the SNR between -3 and 22 dB are randomly chosen. The noisy speech files are further processed with three single-microphone noise reduction algorithms [3, 5, 22] that are used in one of the datasets (see below), which introduces noise reduction artifacts in the training data and should provide a higher robustness of the system towards such artifacts. Hence, we obtain 7138 × 3 noises × 4 conditions (3 noise reduction algorithms + unprocessed) = 85,656 utterances as training data. The development set for training the DNN is generated the same way as the training data containing 330 × 3 noises × 4 conditions = 3960 utterances.

To investigate the influence of additional maskers for training data, we systematically increase the number of maskers from three to all eleven maskers shown in Table 1. We also investigate if the specific selection of noises has a major impact on model prediction quality. Some noise types in the listening datasets (cf. next section) either are similar to speech (ICRA5-250 has a modulation frequency of 4 Hz) or contain speech elements (*Party noise*), and the robustness of the DNN against background voices could be an important factor. This is explored by compiling two sets that contain three noise types with background voices such as children laughing on a playground (*Gym, Mall*, and *Playground*) or without any background voice (*ICRA1, Factory 1*, and *Propeller Plane*). The number of training utterances is kept constant for all experiments presented in this paper to factor out the influence of the amount of data. When increasing the number of noise types, the number of utterances mixed with each noise type is adjusted accordingly. All generated files have a sampling rate of 16 kHz.

Idx.	**Noise**	**Source**
1	Gym	BBC
2	Mall	BBC
3	Playground	BBC
4	ICRA1	DRE01
5	Factory 1	Noisex
6	Propeller Plane	BBC
7	Rain	BBC
8	Factory 2	Noisex
9	Shower	BBC
10	Vacuum Cleaner	BBC
11	SSN	see description

Table 1: *List of noise types and their corresponding source used for ASR training. BBC Sound Effects Lib. and Noisex database are described in [28], and the label DRE01 refers to [2].*

3. Subjective LE Datasets

Below the datasets used for subjective ratings of LE are presented, which are identical to the databases used for the LEAP model based on matched noise for training and testing [11]. All LE experiments were performed with the Oldenburg Sentence test (OLSA) [30], which contains German matrix sentences produced by a male speaker in fixed grammatical structure which are syntactically correct but semantically not predictable. The listeners who participated in the experiments were native German speakers. Subjects rated their perceived LE on a 13-step graphical rating scale that range from *extremly effortful* to *effortless*. By averaging these ratings across subjects, Listening Effort Mean Opinion Scores (LE-MOS) [27] were obtained.

Dataset 1: Hearing aids with and without noise reduction OLSA sentences were mixed with two noises: A stationary

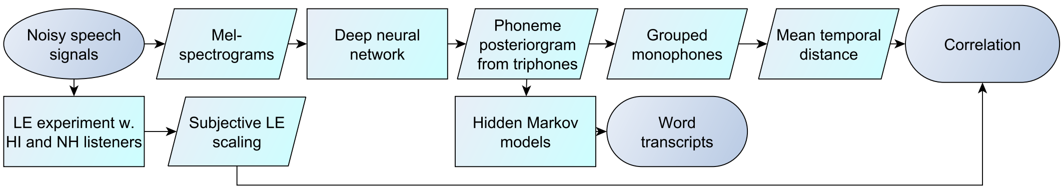

Figure 1: *Illustration of the proposed listening effort model. Decoding with the HMM is performed only during training the model, but not required for obtaining model predictions.*

	Noise	Subjects	LEAP Model						ANIQUE+	ITU-T p.563	SNR est.
			3A	3B	5	7	9	11			
DS1	Airplane	HI	0.665	0.601	0.610	0.603	0.634	0.551	0.642	0.703	0.866
	OLNOISE	HI	0.925	0.898	0.926	0.914	0.901	0.905	0.568	0.717	0.734
DS2	OLNOISE	NH	0.993	0.993	0.993	0.993	0.992	0.991	0.902	0.946	0.921
	ICRA5-250	NH	0.859	0.828	0.951	0.964	0.961	0.939	0.948	0.934	0.576
	OLNOISE	HI	0.995	0.994	0.995	0.995	0.994	0.995	0.966	0.982	0.986
	ICRA5-250	HI	0.967	0.961	0.983	0.985	0.986	0.982	0.980	0.978	0.110
DS3	Traffic2	NH	0.836	0.850	0.878	0.872	0.890	0.884	0.227	0.511	0.626
	Party	NH	0.741	0.791	0.820	0.822	0.822	0.797	0.418	0.304	0.681
	OLNOISE	NH	0.695	0.726	0.761	0.720	0.738	0.745	0.307	0.749	0.630
AVG			0.853	0.849	**0.880**	0.874	**0.880**	0.865	0.662	0.758	0.681

Table 2: *Correlation of LE model predictions and the average subjective listening effort from moderately hearing-impaired (HI) and normal-hearing (NH) subjects. Correlation is reported for three baseline measures and for the proposed non-intrusive LEAP models (gray shading) that have been exposed to a different number of noise types (column header, 3A/B: without/with background voices).*

speech-shaped noise that matches the long-term spectrum of OLSA (OLNOISE) with SNRs from -1 to 14 dB in 3 dB steps, and secondly, airplane cabin noise with SNRs from -10 to 5 dB in steps of 3 dB. The noisy speech files were processed by four commercial hearing aids. These hearing aids were fitted to the average hearing loss of 20 moderately HI listeners. The mean pure tone average is 58 dB, the subjects were between 26 and 83 years old with a median of 71. Each test is conducted with an activated and de-activated noise reduction program. The output signals were recorded at 44 kHz sampling frequency with an ear simulator placed in an artificial head (KEMAR). To obtain a similar long term spectrum and frequency responses from the different hearing aids, the recordings were filtered individually to only regard to the different noise reduction programs and presented via Sennheiser HDA200 headphones. The unaided subjects were tested with 96 generated test signals consisting of 6 SNRs × 4 hearing aids × 4 noise reduction settings [9].

Dataset 2: LE for this dataset was measured with a procedure that adaptively adjusts the SNR to cover the whole range of the LE scale [17]. Two maskers were employed, i.e., OLNOISE and ICRA5-250. The latter is a speech-simulating modulated noise with pauses up to 250 ms [2]. The stimuli with 16 kHz sampling frequency were presented over a loudspeaker in front of the subjects to 15 normal-hearing subjects (age: 21-31 years, avg. 24.6) and 15 hearing-impaired subjects (age: 50-78 years, mean 67.9, average hearing loss measured at frequencies 0.5, 1, 2, and 4 kHz: 41.6 dB). Due to the adaptive adjustments, various SNR values were used.

Dataset 3: Three different noises were mixed with OLSA sentences: OLNOISE, cocktail party and traffic noise with SNRs ranging from -3 to 12 dB in steps of 3 dB. All signals had a sampling rate of 16 kHz. 18 normal-hearing subjects were tested. The subjects were between 19 and 49 years old with a median age of 24 years. Each speech file was presented in four conditions: processed by three single-microphone noise reduction algorithms [3, 5, 22] as well as the unprocessed condition.

4. Results

An overview of correlation coefficients between model output and subjective LE is given in Table 2. Results with an even number of noise types are fully consistent with the data presented in Table 2 and are not shown due to space limitations. The Pearson correlation (r) and the Spearman rank correlation (rs) are computed from the LE-MOS from the LE experiments and the corresponding prediction curve. The LE prediction curve is a third order polynomial regression to take into account curvilinear relations between the MTD and the LE-MOS results as proposed in [11].

On average, all LEAP models outperform the baseline systems with correlation coefficients from 0.849 to 0.880. A comparison of the two models exposed to three maskers (3A and 3B in Table 2) suggests that the specific selection of noise types does not have a major influence on the predictive power of the model on average (with correlation values of 0.853 and 0.849). The following experiments with a larger number of maskers were therefore performed by sequentially adding noise types in the order shown in Table 1 and without considering different selections of noise types. Stable results are obtained across all conditions, with relatively consistent correlations even for individual listening situations. The lowest predictive power is achieved for HI listeners and airplane noise ($r = 0.61$ on average), while the correlations for OLNOISE and the modulated ICRA5-250 in Dataset 2 are consistently high ($r > 0.915$ on av-

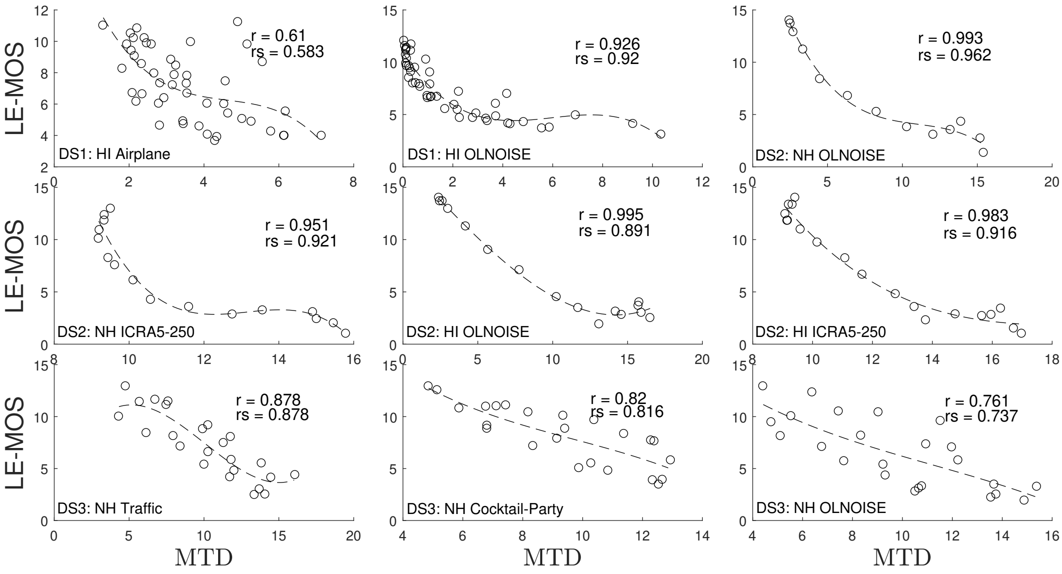

Figure 2: *Relation between the DNN-based performance measure MTD from the LEAP 5 model and the average subjective listening effort for three databases. The dashed line depicts the third order polynomial fitted curve derived from the data.*

erage). The best average prediction is achieved when including 5 or 9 maskers ($r = 0.880$) in the training; detailed results for the system with 5 noises are presented in Figure 2. This model outperforms the baseline measures in all conditions with the exception of airplane noise. The figure shows that a high variance is obtained in airplane noise, as well as with OLNOISE in experiments with signal enhancement (Dataset 1 for low MTD values and Dataset 3), although all OLNOISE conditions also achieve good correlations. On the other hand, variability of LE predictions for NH and HI listeners for in other conditions is small, especially for Dataset 2.

5. Discussion

With the proposed non-intrusive LEAP model that is blind with respect to speech and noise, we achieved good correlations with subjective listening effort. Our results indicate that prediction power does not strongly depend on the number of noise types seen by the DNN since similar results were observed across all trained models (which however peaked for 5 and 9 training masking signals). By using two sets of three very different noises (leftmost results in Table 2), it was also shown that the specific noise selection seems not to be critical for obtaining correlations over 0.8. However, when using just one masker for mismatched training, [11] reported an average correlation of only 0.63, which is lower than the average value of all three baseline measures. Therefore, it is important to expose the model to at least some variability captured by a few maskers. When using matched noises for training and model predictions, very high correlations (0.96 on average) were obtained in earlier research [11]. However, the requirement of knowing the noise type in advance and training a specific model for each new condition appears as a strong model limitation which hinders its application in real-life scenarios, which is alleviated for the current model at the cost of degrading the correlation by 0.08 on average. While the model is trained with English speech (Wall Street Journal Corpus), the evaluation is performed with *German* matrix sentences. The model does not require a decoding with an HMM (and therefore no dictionary), i.e., it can easily be applied to arbitrary acoustic signals. English phonemes that do not occur in German (such as [θ]) should remain unactivated in the posteriorgram, at least for clean utterances, and therefore not affect the MTD measure. German phonemes not present in US English (e.g., [x] as in *acht*) would presumably activate a small number of phonemes, which should reduce the MTD value for clean speech and therefore its potential to differentiate between clean and degraded conditions. The good results obtained with the current models indicate that the German and English phoneme inventory are similar enough so that this effect might not play a major role for model predictions. Nevertheless, it would be interesting to investigate other language pairs and language-matched models in future research.

6. Summary

This paper explored the non-intrusive prediction of subjective listening effort based on the output of a DNN-based phoneme classifier. It is based on the assumption that degraded acoustic signals result in degraded phoneme posterior probabilities which can be quantified with the mean temporal distance [6]. This LEAP model was first proposed in [12], but produced accurate predictions only when the noise type used in LE experiments was used during training of the model, i.e., a new model training was required for each condition. We therefore created multi-condition training sets with independent noise types, and gradually increased the number of maskers that our model was exposed to. The best average correlations of $r = 0.88$ were obtained when using 5 and 9 maskers for training, respectively. However, the specific number of noise types was found not to be crucial, since all systems produced correlations above $r = 0.849$, which is higher than the output of related speech quality models that served as baseline.

7. Acknowledgements

This research was supported by the DFG (Cluster of Excellence 1077/1 Hearing4all; URL: http://hearing4all.eu). The authors would like to thank Sivantos for kindly providing Dataset 1.

8. References

[1] Denk, F., da Costa, J. P. C. L., and Silveira, M. A. (2014). "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.

[2] Dreschler, W. A., Verschuure, H., Ludvigsen, C., & Westermann, S. (2001). ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment: Ruidos ICRA: Seates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos. Audiology, 40(3), 148-157.

[3] Ephraim, Yariv, and David Malah (1984). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator." IEEE Transactions on acoustics, speech, and signal processing 32.6, pp. 1109-1121.

[4] Garofalo, J., Graff, D., Paul, D., Pallett, D. (2007). CSR-I (WSJ0) Complete. Linguistic Data Consortium. Philadelphia.

[5] Hendriks, Richard C., Timo Gerkmann, and Jesper Jensen (2013). "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art." Synthesis Lectures on Speech and Audio Processing 9.1, pp. 1-80.

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

[7] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh (2006). "A fast learning algorithm for deep belief nets." Neural computation 18.7, pp. 1527-1554.

[8] Houtgast, T. , Steeneken, H.J.M (1971), "Evaluation of Speech Transmission Channels by Using Artificial Signals", Acustica 25, 355367.

[9] Huber, R., Schulte, M., Vormann, M., Chalupper, J. (2010). "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*.

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

[11] Huber, R., Krüger, M., Meyer, B.T. (2018). "Single-ended prediction of listening effort using deep neural networks," Hearing Research, https://doi.org/10.1016/j.heares.2017.12.014.

[12] Huber, R., Spille, C., Meyer, B.T. (2017). "Single-ended prediction of listening effort based on automatic speech recognition," in Proc. Interspeech, doi:10.21437/Interspeech.2017-1360.

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

[14] Kim, D.-S., Tarraf, A. (2007). "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.

[15] Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. Ear and Hearing, 33(2), pp. 291-300.

[16] Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. Audiology, 36(3), pp. 155-164.

[17] Krüger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. The Journal of the Acoustical Society of America, 141(6), pp. 4680-4693.

[18] Mackersie, Carol L., Imola X. MacPhee, and Emily W. Heldt (2015). "Effects of hearing loss on heart-rate variability and skin conductance measured during sentence recognition in noise." Ear and hearing 36.1, 145.

[19] Mallidi, S. H., Ogawa, T., and Hermansky, H. (2016). Uncertainty estimation of DNN classifiers, 2015 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2015 - Proc., pp. 283288. doi:10.1109/ASRU.2015.7404806

[20] Mohamed, Abdel-rahman, Geoffrey Hinton, and Gerald Penn (2012). "Understanding how deep belief networks perform acoustic modelling." Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[21] Parihar, N., Picone, J., Pearce, D., and Hirsch, H. (2003). Performance analysis of the Aurora large vocabulary baseline system, Proc. of Eurospeech03, 2004, pp. 1013.

[22] Plapous, C., Marro, C., Scalart, P. (2006). "Improved signal-to-noise ratio estimation for speech enhancement." IEEE Transactions on Audio, Speech, and Language Processing 14.6: pp. 2098-2108.

[23] 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, ASRU*, pp. 1-4.

[24] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier (2014). "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," *J. Acoust. Soc. Am.* vol. 136, p. 2642-2654, doi: 10.1121/1.4897398.

[25] Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research, 52(5), pp. 1230-1240.

[26] Schepker, H., Haeder, K., Rennies, R., Holube, I. (2016). "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, doi: 10.1080/14992027.2016.1219774

[27] Schulte, M., Meis, M., and Wagener, K. (2007). "Listening Effort and Speech Intelligibility." 8th EFAS Congress/10th Congress of the German Society of Audiology.

[28] Varga, A., and Steeneken, H.J.M. (1993). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems." Speech communication 12.3: 247-251.

[29] Vesel, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. Proc. Interspeech, pp. 2345-2349.

[30] Wagener, K. C., Kühnel, V., Kollmeier, B. (1999). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a sentence test for the German language)," *Zeitschrift für Audiologie*, vol. 38, pp. 4-15.