Objective Assessment of a Speech Enhancement Scheme with an Automatic Speech Recognition-Based System

Rainer Huber[1], Arne Pusch[1], Niko Moritz[1], Jan Rennies[1,2], Henning Schepker[3], Bernd T. Meyer[4]

[1]Fraunhofer IDMT, Hearing, Speech and Audio Technology and Cluster of Excellence Hearing4All, Oldenburg, Germany
[2]Boston University, Department of Speech, Language and Hearing Sciences, Boston, MA, USA
[3]University of Oldenburg, Department of Medical Physics and Acoustics, Signal Processing Group, and Cluster of Excellence Hearing4All, Oldenburg, Germany
[4]University of Oldenburg, Department of Medical Physics and Acoustics, Medical Physics Group, and Cluster of Excellence Hearing4All, Oldenburg, Germany

Email: {rainer.huber, arne.pusch, niko.moritz, jan.rennies-hochmuth}@idmt.fraunhofer.de,
{henning.schepker, bernd.meyer}@uni-oldenburg.de

Abstract

A single-ended method for the prediction of perceived listening effort based on an automatic speech recognition system was adopted from the literature and modified to evaluate a near-end listening enhancement (NELE) scheme. The listening effort prediction method employs a deep time delay neural network (TDNN) that was trained as part of an automatic speech recognizer. The TDNN computes phoneme posterior probabilities (or "posteriorgrams"), which degrade in the presence of noise or other distortions. The degree of posteriorgram degradation is quantified by a performance measure and serves as a predictor for mean subjective listening effort ratings of normal-hearing listeners. The modification of the original method consists of the usage of a TDNN (in contrast to a regular feed-forward DNN used before), which was trained on a much bigger speech corpus. Without any task-specific training or optimization, the modified method achieves a very high correlation with subjective listening effort ratings from the used test data set of unprocessed and NELE-processed speech in two types of background noise (r = 0.98), generalizes to unseen noise conditions, and produces consistent predictions across these conditions that can be directly compared.

1 Introduction

The evaluation of speech enhancement schemes usually involves subjective listening tests to measure a possible improvement of speech intelligibility or a reduction of listening effort, respectively. In contrast to speech intelligibility, listening effort can still be affected by further reduction of noise levels at such SNRs (e.g., [1, 2]). Consequently, Rennies et al. [2] conclude that "intelligibility is an insensitive measure to evaluate many everyday listening conditions". In principle, the assessment of listening effort allows for the evaluation of algorithm performance even at SNRs where speech intelligibility is at ceiling already in the unprocessed reference condition, i.e., at SNRs at which little or no room to quantify algorithm benefit in terms of intelligibility is possible. However, the measurement of speech intelligibility or listening effort by means of formal subjective listening tests is time consuming and costly and cannot be used for, e.g., online-monitoring. Hence, instrumental methods to predict speech intelligibility or perceived listening effort would be valuable tools for the automatic evaluation of speech enhancement schemes. Moreover, such tools could be used for online-steering of speech enhancement schemes, being a "model in the loop".

Signal-based instrumental methods for the prediction of, e.g., speech quality, speech intelligibility or listening effort can be classified into single-ended (or "reference-free", "non-intrusive") and double-ended (or "reference-based", "intrusive") methods. Double-ended methods (such as [3, 4] for speech/audio quality assessment) typically achieve more accurate predictions than single-ended methods (such as [5, 6]), but have the disadvantage that they need a clean or nearly clean reference signal, which is often not available. Consequently, we want to focus on single-ended methods for the prediction of perceived listening effort in the following.

In a recent paper [7], Huber et al. introduced a single-ended approach for listening effort prediction from acoustic parameters (LEAP) based on an automatic speech recognition (ASR) system. The ASR system employs a deep neural network (DNN) to compute phoneme posterior probabilities (or "posteriorgrams") of input speech. Distortions or additive noise increase the uncertainty of the ASR system, which is reflected by smeared posteriorgrams. The degree of posteriogram degradation is quantified by a performance measure, i.e., the *M*-Measure proposed by Hermansky et al. [8]. It has been found that the *M*-Measure correlates well with measured listening effort data of several data sets [7]. In related work, this modeling approach was also explored in the context of speech quality prediction [9, 10]. One limitation of the original method presented in [7] was that the ASR system was trained on the specific background noises of the test data sets. This issue is addressed with a modified LEAP model [11], which investigates diverse training data sets and multiple noise types with the aim of generalizing to unseen noise conditions during testing. The average correlation for three different databases was found to be high (r = 0.88) although not as high as for the original approach with known test noises (r = 0.96). These models exhibit different output values for different conditions. When pooling the data points from different acoustic scenarios, lower correlations are obtained, which prevents them from being directly applied as model-in-the-loop since they would need to know the specific acoustic scene *a priori*. In this study, we investigate deep time delay neural networks (TDNNs) trained with a much bigger speech corpus with speaker-independent training and mismatched noises (see Section 2.2 for details). The test data consist of noisy speech signals, with clean speech signal components being either unprocessed or processed

ISBN 978-3-8007-4767-2

by an adaptive, non-linear algorithm employing frequency-shaping and dynamic range compression (AdaptDRC) [12] before mixing them with noises. The AdaptDRC algorithm aims at improving the intelligibility of speech in noise by increasing the local SNRs in the frequency bands most important for intelligibility, while preserving the broadband SNR (see Section 2.1.3 or [12] for details).

2 Methods

2.1 Listening effort data

The set of signals with corresponding subjective listening effort ratings was adopted from a study of Pusch et al. [13] and will be described briefly in the following.

2.1.1 Stimuli

German sentences were used as speech material, which were taken from the Oldenburg sentence test [14]. The speech was either unprocessed or processed by the AdaptDRC algorithm [12]. While the speech level was fixed at 60 dB SPL, the noise levels were varied to obtain the desired SNRs. Two different noise types were used: A stationary, speech-shaped noise, which had the same long-term spectrum as the average unprocessed speech material (OLNOISE), and a cafeteria noise, which contained more envelope fluctuations. Based on a pilot measurement, a wide range of SNRs was chosen (-15, -10, -5, 0, 2.5, 5, 7.5, 10 dB). The signals were presented diotically to the subjects via Sennheiser HD650 headphones in a sound-attenuated booth.

2.1.2 AdaptDRC algorithm

For enhancing speech intelligibility in given noisy conditions, speech can be pre-processed by speech enhancement algorithms before it is played back via loudspeakers (e.g., when playing back an announcement in a noisy room). This is commonly referred to as near-end listening enhancement (NELE). The AdaptDRC algorithm is one of these algorithms. It processes the speech by applying frequency-shaping and dynamic range compression (DRC) in time frames of 20 ms length. First, each time frame is split into eight octave subbands centered at 125 Hz to 16 kHz using a non-decimating filterbank. Based on a simplified version of the Speech Intelligibility Index (SII) [15], the subbands are weighted. For low SII values, the weighting leads to an increase of energy in high frequencies bands, whereas for SII values close to 1, the spectral shape is not modified. The DRC stage aims at increasing the audibility of softer parts. For low subband SNRs, a higher compression ratio is applied than for high SNRs. A typical constraint for NELE algorithms is to maintain the broadband RMS power of the input speech signal. To ensure this, a normalization of the processed speech signal is applied. This normalization implies that both unprocessed and processed speech are presented at the same broadband SNR in a given background noise.

2.1.3 Subjects and procedure

Eleven normal-hearing subjects (nine male and two female) participated in the experiment. Their median age was 25.5 years (ranging from 24 to 36 years). All had normal audiograms with pure-tone averages lower than 25 dB HL. The subjects listened to the stimuli and rated their perceived listening effort on a scale with 13 categories ranging from “no effort” (1 Effort Scaling Categorical Unit, ESCU) to “extreme effort” (13 ESCU) [1]. In addition, a 14th category ('only noise') was available for conditions in which subjects could not detect any speech. For each subject and trial, a randomly selected noise start sample and sentence were used. Each combination of noise type, processing type, and SNR was measured six times by each subject, and the combinations were randomly spread in the experiment.

2.2 Posteriorgram generation

An acoustic model for automatic speech recognition (ASR) was trained prior to the extraction of context-dependent triphone posteriograms. For ASR training, we used about 1.000 hours of unprocessed German speech data of an in-house training data set and inflated it up to about 8.000 hours in a multi-condition training setup. A deep time-delay neural network (TDNN) [16, 17], which is also known as a one-dimensional temporal convolutional neural network, was trained with the lattice-free maximum mutual information (LF-MMI) criterion [18]. To save computational time, the LF-MMI trained neural network modeled output posterior probabilities at one third of the frame rate of conventional acoustic ASR models, which usually run at a 100 Hz frame rate. The TDNN topology was similar to a setup described in [19] that had a total context size of +/- 15 input feature frames (equal to 310 ms), which were analyzed over 7 hidden neural network layers of 700 rectified linear units (ReLU) dimensions each. The dimensionality of the final output layer amounted to 6448, which was the result of decision tree clustering of context-dependent Hidden Markov Model output distributions. As acoustic features input to the TDNN, we used 40-dimensional log-Mel filterbank energies. Note that during training, the TDNN used here had two output layers, one that followed the LF-MMI objective function and one that followed a cross-entropy (CE) objective function. The latter one is usually used to regularize training only, while the former one is used for ASR purposes. In this work we used the CE output layer for generating posteriorgrams instead, due to better results.

2.3 Performance measure

From the posteriorgrams, the mean temporal distance (or *M*-Measure) as proposed by Hermansky et al. [8] was 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 (which is equal to the length of the analyzed

speech file, i.e. around 3s in the present study), and D being the symmetric Kullback-Leibler divergence between two vectors x and y with components $x(i)$ and $y(i)$:

$$D(x,y) = \sum_{i=1}^{N} x(i)\log(\frac{x(i)}{y(i)}) + \sum_{i=1}^{N} y(i)\log(\frac{y(i)}{x(i)})$$

In the present study, N equals the dimensionality of the TDNN output layer (6448) and M was computed for Δt =350 to 800 ms (in 50 ms steps) and averaged to yield the final listening effort predictor $\bar{M}$.

2.4 SNR estimator

As the perceived listening effort might be dominated by the SNR of the noisy speech signals used in this study, an SNR estimator was also applied as a baseline measure. For a fair comparison with the single-ended ASR-based approach, the SNR estimator had to be "blind" in the sense that it only works on the mixed signal and does not have access to the separate clean speech and noise signal. The SNR estimation method employed here was proposed by Denk et al. [20]. Their method performs an iterative, threshold-based combined Voice Activity Detection (VAD) and broadband SNR estimation. From a first threshold (using the dynamics of the signal) the speech is estimated. From this estimation, the SNR is calculated, and is then used to define a new speech threshold leading to a new speech detection to serve as basis for a new SNR estimation. This is continued until two consecutive SNR estimates do not differ more than a certain value (for details, see [20]).

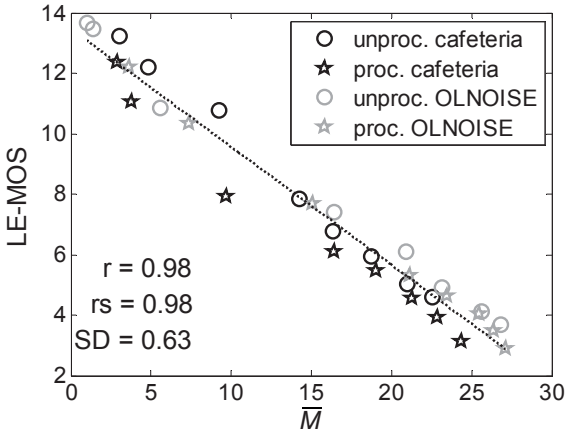

Fig. 1: Comparison of measured listening effort (mean subjective ratings over trials and subjects – LE-MOS) with corresponding averaged M-Measure values $\bar{M}$. Black circles: unprocessed speech with cafeteria noise; black stars: speech processed by the AdaptDRC algorithm, mixed with cafeteria noise; gray circles: unprocessed speech with speech simulating OLNOISE; gray stars: speech processed by the AdaptDRC algorithm, mixed with OLNOISE. Dotted line: linear regression fit. r: linear correlation coefficient after Pearson. rs: rank correlation coefficient after Spearman. SD: standard deviation (in LE-MOS units) from linear fit.

3 Results

3.1 Results obtained with the *M*-Measure

The noisy speech signals of the test data set were fed into the ASR system described earlier and the averaged M-Measure $\bar{M}$ was calculated for all resulting 2112 posteriorgrams (2 noise types x 8 SNRs x 2 processings (AdaptDRC/unprocessed) x 11 subjects x 6 trials per condition). Subsequently, $\bar{M}$ was averaged across all trials per SNR and processing condition (i.e. unprocessed vs. processed by AdaptDRC) and across all subjects. The subjective listening effort ratings were averaged in the same way to yield the Listening Effort Mean Opinion Score (LE-MOS). The scatter plot shown in Fig. 1 compares the subjective LE-MOS values with corresponding objective mean $\bar{M}$ values. In the plot, the correlation between subjective and objective data is quantified by the Pearson correlation coefficient r and the Spearman rank correlation coefficient rs. Additionally, the standard deviation (SD) of the LE-MOS data from a linear regression fit is given. The scatter plot and correlation coefficients show that a very high correlation between measured LE-MOS and the computed $\bar{M}$ data is achieved (r = 0.98). The relation between LE-MOS and $\bar{M}$ data is independent of the noise type. There is no systematic difference between the $\bar{M}$–LE-MOS relations with regard to whether the speech signals were processed by the AdaptDRC algorithm or not. This means that the effect of the AdaptDRC processing in terms of listening effort reduction is correctly predicted.

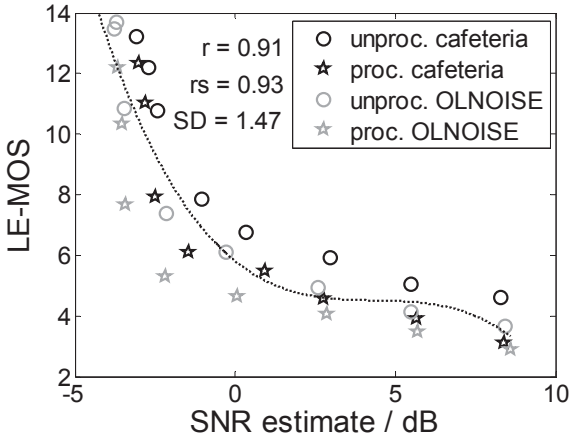

Fig. 2: As Fig. 1, but for the SNR estimator. Dotted line: 3rd-order polynomial regression fit.

3.2 Results obtained with the SNR estimator

The relation between LE-MOS data and corresponding SNR estimates is shown in Fig. 2. A floor effect of the SNR estimate at about -4 dB is apparent (remember that the actual SNRs ranged down to -15 dB). As a consequence, the regression curve (dotted line) becomes much steeper at lower SNRs. Moreover, the effects of background noise type and of processing type on the perceived listening effort are not accounted for by the SNR estimate. This is particularly apparent at positive SNRs: While subjective listening effort ratings of the four dif-

ferent conditions (processed / unprocessed, OLNOISE / cafeteria noise) spread across a certain LE-MOS range per SNR condition, the corresponding SNR estimates are nearly constant. Hence, the overall correlation with the subjective data is clearly smaller for the SNR estimate than for the averaged *M*-Measure: When a 3rd-order polynomial regression fit is applied, the Pearson correlation coefficient amounts to $r = 0.91$.

4 Discussion

The presented results confirm the qualification of the ASR-based, single-ended method proposed earlier [7] to predict the perceived listening effort of noisy speech in the presence of different maskers with current deep learning algorithms. It appears that deep systems trained with an appropriate amount of data have become robust enough to be similarly affected by speech distortions as human listeners are, which is supported by their use for conversational telephone speech [21] or their use for predicting speech intelligibility of normal-hearing listeners [22]. For listening effort, correct predictions are obtained without any task-specific training or optimization of the method, the influences of the SNR, the noise type and the non-linear speech enhancement processing in the used dataset. In contrast to the original method [7], the system investigated here used a deep TDNN that was trained with a larger dataset (several thousand hours of data) for speaker-independent and noise-mismatched recognition. In related work [11], it was shown before that the general approach is principally suited to predict listening effort in unknown noise scenarios. However, the comparison of results with the current net (8000 hours of training data, $r = 0.98$) with the one from [11] (80 hours of training data, $r = 0.88$) suggests that our model produces good results with comparatively small ASR training sets, but the predictive power is increased when increasing the amount of speech seen during training.

In the current study, very high correlations were obtained in unseen noise types which were pooled over different noise conditions. The latter point is especially important since it implies that consistent model predictions were obtained across the acoustic scenes investigated here. In [7] and [11], different mapping functions between the averaged *M*-Measure and measured listening effort data were observed for different noise conditions, i.e., the acoustic scene needed to be known *a priori* for accurate predictions. This *a priori* knowledge seems not to be required by the TDNN model presented in this paper, which could make it attractive for listening effort predictions of processed and unprocessed signals for assisted hearing. In future work, the model will be tested in a wider range of noise types and processing strategies.

The correlation between subjective and objective data achieved in the present study appears to be close to the theoretical maximum determined by the reliability of the subjective data. The average inter-individual standard deviation of the subjective listening effort data amounts to 1.5 Effort Scaling Categorical Units (ESCUs), which has to be compared to the standard deviation of 0.7 ESCUs between the subjective LE-MOS and the fitted objective $\overline{M}$ data. If the subjective data are divided into two randomly selected subsets consisting of ratings averaged across three of the total six trials per condition and subject (which can be interpreted as test and retest datasets), the (test-retest) correlation between these two subsets is $r = 0.995$.

A possible shortcoming of the ASR-based listening effort prediction method could be an underestimation of listening effort due to interfering speech. First own experiences indicate that phoneme activations are hardly smeared by interfering speech. Instead, the phoneme density in the posteriorgrams increases, but this does not affect the *M*-Measure. A possible remedy for this shortcoming could be the development and usage of a detector of background speech, which could be implemented by collecting few data points from the target or background speaker to perform a speaker-specific adaptation of the model during runtime. Whenever background speech is detected, a correction factor could be applied to increase the listening effort estimate.

So far, no alternative single-ended measures for listening effort prediction with similar prediction accuracies are known to the authors. The blind SNR estimator investigated in the present study fails to estimate the true SNRs, at least for negative SNRs. Moreover, even if it could, the influence of the type of background noise and of the speech enhancement processing on the perceived listening effort would not be taken into account, which can be seen in the results at positive SNRs presented in this study (Fig. 2). In [7], two standard single-ended speech quality measures (i.e., ITU-T P.563 [5] and ANIQUE+ [6]) were tested for comparison as well and showed high correlations for one of the three tested datasets, although not as high as the ones achieved with the ASR-based method. However, the prediction accuracies for the other two datasets were poor.

5 Conclusions

The single-ended prediction method for perceived listening effort based on an ASR-based DNN was adopted from [7] and modified using an extensively trained deep TDNN. This allowed the method to better generalize with regard to unknown noise types as well as across noise types, i.e., it produced accurate listening effort predictions without using the test noise during training (a limitation of the study presented in [7]) or by using individual fits that are required for each test noise as reported in [18]. The method was applied to human listening test data obtained from an evaluation study [13] of a speech enhancement scheme (AdaptDRC [12]). A very high correlation between subjective and objective listening effort ratings was found ($r = 0.98$). This proves the qualification of the proposed method to be used as an instrumental tool for the evaluation of, e.g., adaptive and nonlinear speech enhancement schemes or as a "model in the loop" for online steering of speech enhancement algorithms.

6 Acknowledgment

This study was supported by the DFG Cluster of Excellence EXC 1077/1 "Hearing4all".

References

[1] M. Krueger, M. Schulte, M. Zokoll, K. Wagener, M. Meis, T. Brand, and I. Holube, “Relation between listening effort and speech intelligibility in noise,” *Am. J. Audiol.,* vol. 26, pp. 378-392, Oct. 2017.

[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, no. 5, pp. 2642-2653, Nov. 2014.

[3] ITU-T, 2014. Methods for Objective and Subjective Assessment of Speech Quality (POLQA): Perceptual Objective Listening Quality Assessment. Recommendation P.863, International Telecommunication Union, Geneva, Switzerland.

[4] R. Huber and B. Kollmeier, “PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902-1911, Nov. 2006.

[5] ITU-T, 2004. Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications. Recommendation P.563, International Telecommunication Union, Geneva, Switzerland.

[6] D. S. Kim and A. Tarraf, “ANIQUE+: a new American national standard for nonintrusive estimation of narrowband speech quality,” *Bell Labs Tech. J.* vol. 12, no. 1, pp. 221-236, May 2007.

[7] R. Huber, M. Krüger, and B. T. Meyer, “Single-ended prediction of listening effort using deep neural networks”, *Hear. Res.*, vol. 359. pp. 40-49, Mar. 2018

[8] H. Hermansky, E. Variani, and V. Peddinti, “Mean temporal distance: predicting ASR error from temporal properties of speech signal,” in *Proc. IEEE Conf. Acoust. Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May. 2013, pp. 7423-7426.

[9] R. Huber, J. Ooster, Meyer, “Single-ended Speech Quality Prediction Based on Automatic Speech Recognition”, J. Aud. Eng. Soc., in press

[10] J. Ooster and B.T. Meyer. "Prediction of Perceived Speech Quality Using Deep Machine Listening," submitted to the ITG Conference on Speech Communication, 2018.

[11] P. Kranzusch, R. Huber, M. Krüger, B. Kollmeier, B.T. Meyer, “Prediction of Subjective Listening Effort from Acoustic Data with Non-Intrusive Deep Models,” submitted to Interspeech 2018.

[12] H. Schepker, J. Rennies, and S. Doclo, “Speech-in-noise enhancement using amplication and dynamic range compression controlled by the speech intelligibilility index,” *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 2692-2706, Nov. 2015.

[13] A. Pusch, J. Rennies, H. Schepker, and S. Doclo, „Höranstrengung als Messverfahren für die Evaluierung von Near-End Listening Enhancement Algorithmen,“ in *Proc. Fortschritte der Akustik - DAGA 2018*, Munich, Germany, Mar. 2018, pp. 543-546.

[14] K. C. Wagener, V. Kühnel, and B. Kollmeier, „Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: design des Oldenburger Satztests (Development and evaluation of a German sentence test I: design of the Oldenburg sentence test),“ Z. Audiol, vol. 38, pp. 4-15, Sep. 1999.

[15] ANSI: ANSI S3.5-1997. American National Standard Methods for the Calculation of the Speech Intelligibility Index. New York: ANSI, 1997.

[16] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transaction on Acoustics, Speech, and Language Processing, vol. 37, no. 3, pp. 328-339, 1989.

[17] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of a long temporal contexts," in Interspeech, Dresden, 2015, pp. 2440-2444.

[18] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in Interspeech, San Francisco, 2016.

[19] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," IEEE Signal Processing Letters, vol. 25, no. 3, pp. 373-377, 2018.

[20] F. Denk, J. P. C. L. da Costa, and M. A. Silveira, “Enhanced forensic multiple speaker recognition in the presence of coloured noise,” in Proc. Int. Conf. Signal Process. Comm. Syst. (ICSPCS), Gold Coast, Australia, Dec. 2014, pp. 1-7.

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

[22] C. Spille, S.D. Ewert, B. Kollmeier, B.T. Meyer "Predicting speech intelligibility with deep neural networks," Comput. Speech Lang 48, pp. 51-66. doi:10.1016/j.csl.2017.10.004.