Interspeech 2018
2-6 September 2018, Hyderabad

Prediction of Perceived Speech Quality Using Deep Machine Listening

Jasper Ooster[1,3], *Rainer Huber*[2,3], *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

jasper.ooster@uni-oldenburg.de, rainer.huber@idmt.fraunhofer.de,
bernd.meyer@uni-oldenburg.de

Abstract

Subjective ratings of speech quality (SQ) are essential for evaluating algorithms for speech transmission and enhancement. In this paper we explore a non-intrusive model for SQ prediction based on the output of a deep neural net (DNN) from a regular automatic speech recognizer. The degradation of phoneme probabilities obtained from the net is quantified with the mean temporal distance proposed earlier for multi-stream ASR. The SQ predicted with this method is compared with average subject ratings from the TCD-VoIP speech quality database that covers several effects of SQ degradation that can occur in VoIP applications such as clipping, packet loss, echo effects, background noise, and competing speakers. Our approach is tailored to speech and therefore not applicable when quality is degraded by a competing speaker, which is reflected by an insignificant correlation between model output and subjective SQ. In all other conditions mentioned above, the model reaches an average correlation of $r = 0.87$, which is higher than the correlation achieved with the baseline ITU-T P.563 ($r = 0.71$) and the American National Standard ANIQUE+ ($r = 0.75$). Since the most robust ASR system is not necessarily the best model to predict SQ, we investigate the effect of the amount of training data on quality prediction.

Index Terms: single-ended speech-quality prediction, deep learning, mean temporal distance

1. Introduction

The perceived speech quality (SQ) is an important measure in applications that range from telecommunication over speech enhancement algorithms to the design of hearing aid processing. Several reference-based models such as PESQ [1] and POLQA [2] have been proposed that provide good estimates of SQ, but require a separate input of the original speech signal and a degraded version of that signal, which is not available in many real-life applications. Several reference-free (or non-intrusive, or single-ended) estimators for perceived SQ have been proposed which only require a potentially distorted speech signal as input. Two algorithms that were shown to produce accurate predictions of subjective SQ are the ITU standard P.563 [3] and ANIQUE+ [4], which is a standard of the American National Standard Institute. P.563 estimates separate quality features from signal characteristics such as the SNR, linear prediction coefficients, and interruption indicators, and combines them into the SQ prediction. ANIQUE+ estimates the perceived speech quality by combining three intermediate measures of distortion, i.e., mute and non-speech distortion, as well as frame distortion. The latter is quantified by performing a spectral modulation analysis based on a perceptual model. SQ models based on machine learning have been proposed by Falk and Chan [5], which exploit intermediate features produced by Gaussian mixture models, support vector machines and random forest classifiers, which are integrated in an additional classification step. A comprehensive overview speech quality prediction algorithms is presented in [6].

In this paper, we explore a model that has been proposed for the prediction of subjective listening effort of normal-hearing and hearing-impaired listeners [7, 8] and for SQ prediction [9]. The model is based on Deep machine listening for Estimating Speech Quality (DESQ). In contrast to our previous work [7, 8] the model is blind with respect to speech signals, noise types, and artefacts from speech enhancement algorithms. Further, all previous approaches used a limited training set, i.e., the influence of the amount of training data on model prediction performance remained unclear.

The model is based on a regular automatic speech recognition (ASR) system that combines a DNN with a hidden Markov model (HMM). To produce SQ predictions, the output of the DNN (representing phoneme probabilities) is quantified, since it is potentially degraded in the presence of speech distortions, which could relate to subjective SQ. As measure for quantification, the mean temporal distance as proposed by Hermansky and colleagues is used that was shown to accurately predict phoneme error rates [10] and was later used for selecting the optimal stream in a multi-stream ASR system [11].

Our model is evaluated using the TCD-VoIP database that contains subjective ratings for signals distorted by VoIP artifacts [12]. We analyze the correlation between the mean objective scores (MOS) [13] and the model output, and compare the results to two baseline measures (ANIQUE+ and ITU P.563). To obtain good SQ estimates, the ASR-based model should be affected by signal distortions similar to listeners, which relates to the robustness of the ASR system. We therefore analyze the influence of the amount of training data as well as techniques such as state-level minimum Bayes risk (sMBR) [14] on the predictive power of the model.

The remainder of this paper is structured as follows: The general concept of the ASR-based model is described in the next section, along with the ASR architecture, the corresponding training data, as well as the SQ database. The results section presents model performance in five different types of distortion. Discussion and Summary are presented in Sections 4 and 5, respectively.

2. Methods

2.1. Speech quality prediction system

The SQ model is created by first training a standard ASR system that combines a feed-forward DNN (which serves as acoustic model) with an HMM. The training procedure is described

10.21437/Interspeech.2018-1374

in the next section in detail. We assume that phoneme posterior probabilities from the DNN degrade in the presence of factors that negatively affect speech quality, and quantify this degradation by using the mean temporal distance (MTD) [10]. It is based on the observation that signal distortions can result in phoneme activations that are temporally smeared (cf. lower left panel in Fig. 1, in which the class *AH* is over-proportionally activated, and overall activations are less distinct than for the clean case). Hence, the average difference $\mathcal{D}$ between two phoneme vectors p with a temporal distance Δt should be higher for clean than for noisy vectors. As in the original proposal [10], we use the symmetric Kullback-Leibler divergence as distance measure $\mathcal{D}$, and obtain the difference for a given Δt for an segment of length T between two phoneme vectors $p(t)$ according to

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

Since [10] reported the curve of $\mathcal{M}$ over Δt to saturate after 200 ms (which was attributed to coarticulation effects), we obtain a scalar value by calculating and averaging $\mathcal{M}$ from 350 to 800 ms, i.e., the saturated and stable part of the curve. In the following, the resulting scalar value is referred to as mean temporal distance (MTD). Before calculating the MTD, the context-dependent triphones from the DNN are grouped to approximately 40 monophones. This allows to visualize the output (Figure 1), is computationally cheaper, and produces similar results than using triphone activations [15]. Note that a forward-run of the model does not require a decoding step with the HMM or a word transcript, since it relies on the DNN output alone.

2.2. ASR system

The DNN was trained on 40-dimensional log-Mel-spectral coefficients features with a splicing of ± 5 frames using the *nnet1* recipe from the open source ASR software Kaldi [16]. The DNN had six layers, each with 2048 neurons, a softmax output-layer and a sigmoid-nonlinearity. Before training the DNN, alignments for ≈3,400 triphones were created as training targets. This was done using a Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) system with Mel-Frequency Cepstral Coefficients (MFCCs) features with 13 components per frame, to which the first and second numerical derivatives (delta and double delta features) were appended. The MFCC for the GMM features were adapted to each speaker with a Feature Space Maximum Likelihood Linear Regression (fMLLR) [17] on top of a Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) [18].

The DNN was initialized with a layer-wise Restricted Boltzmann Machine (RBM) pre-training [19]. This pre-trained DNN was fine-tuned with frame cross-entropy (CE) training [20]. With this fine-tuned DNN, new phonetic alignments were created on which the pre-trained network was fine tuned again. This complete procedure is a standard approach for training ASR models in kaldi and has not been optimized for SQ prediction. Additionally, we also tried five iterations of sMBR [14] sequence-discriminant training on top of alignments generated by the second fine-tuned DNN. During the training procedure new alignments are created after the first iteration of the sMBR sequence-discriminant training.

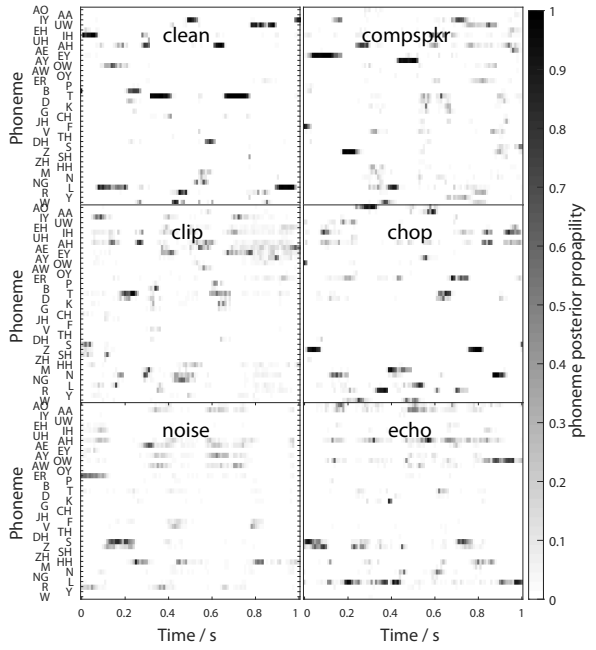

Figure 1: *Phoneme posteriorgrams for a clean speech segment (upper left), and the five conditions of the TCD-VoIP database.*

2.3. ASR training data

The WSJ1 speech corpus is used as basis training set. To investigate the effect of the size of the training data (and potentially the trade-off between robustness of the ASR system and performance of the SQ model), we created four additional training sets by random selection of 20 to 80% of the original utterances in steps of 20%. The full SI284 set contains 37,416 utterances and 81.27 h from 283 speakers. To create training sets which are similar to the Aurora4 multi-condition training set, we added additive noise at random SNRs in the range of 10 dB to 20 dB to 75% of the utterances of each training set. Finally, all files were filtered according to the ITU-T recommendation P.341 [21]. We used the original Aurora4 maskers as additive noise (airport, car, restaurant, subway, babble, exhibition, street, train). To ensure that the DNN does not overfit to the relatively short noise files from the Aurora 4 multi-condition training set, we also added additional noises from the Bits and Pieces sound effects library (http://www.bitsandpieces.co.uk/) that are similar to the original selection. These include *trade show atmosphere, shopping centre atmosphere, crowds chatting on street, town skyline, London street, self service restaurant atmosphere, book market atmosphere, airport arrivals hall atmosphere, large theatre foyer atmosphere, motorway, networker train,* and *commuter train*.

2.4. Subjective listening data

We used the TCD-VoIP corpora [12] to evaluate our model. This database contains subjective quality ratings in the presence of different degradations that can occur in VoIP applications, which are not limited to narrowband data. While the clean speech is also available from in the database, only the corrupted signals were used in our experiments. In the following all conditions are briefly described:

- Clipping effects: The wav signal is multiplied with a fac-

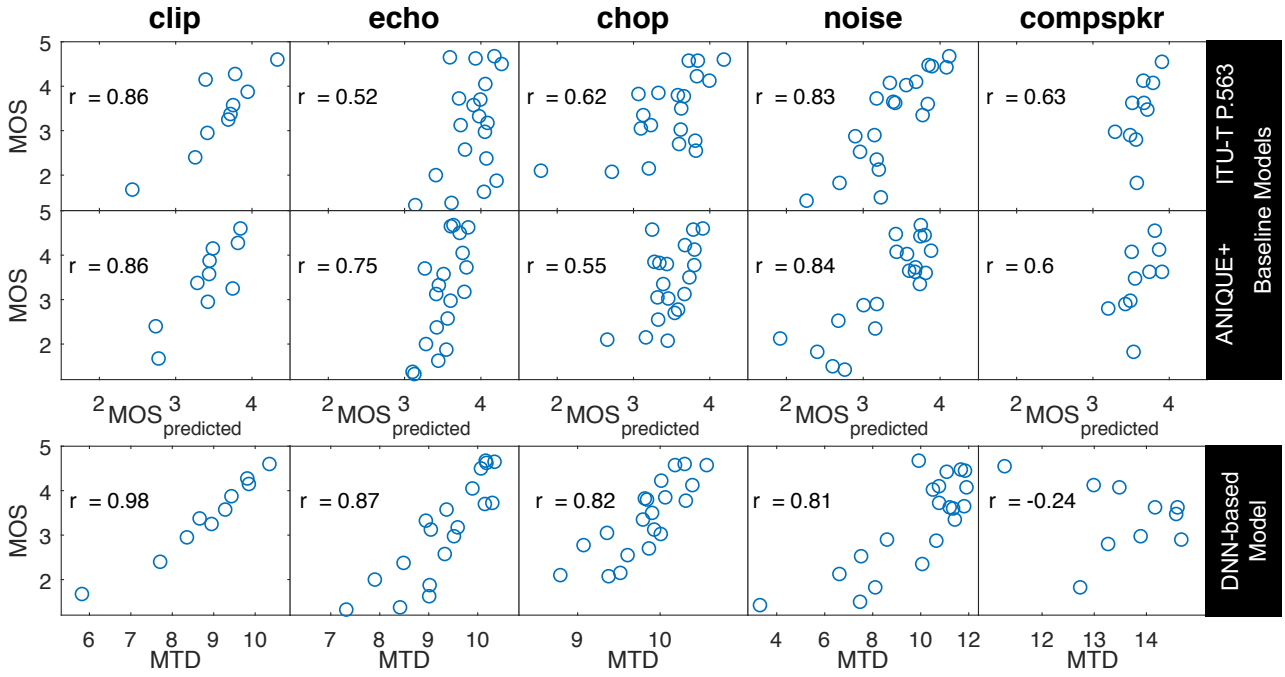

Figure 2: *Model prediction for baseline models (top rows) and the proposed model (bottom row) and relation to mean objective score of subjective listening effort. Each MOS data point represents the average scoring from 24 subjects on four files, resulting in 96 ratings.*

tor between 1 and 55, causing some portion of the samples to be clipped (i.e. set to 1 or -1).

- Echo effects: One ore more copies of the signal are added to the original signal with a delay between 0 and 220 ms and a relative amplitude of the first delayed version related to the original between 0 and 0.5.
- Chopped speech: Samples with a length between 20 and 40 ms are either replaced with zeros, deleted entirely or overwritten with the previous portion of samples at a rate of 0 to 6 chops/s.
- Background noise: Car, street, office and babble noise are additively mixed to the signal at SNRs between 5 and 55 dB. The noise files are taken from [22].
- Competing speakers: Two speakers (female/male, female/female, or male/male) talking in the background at SNRs between 10 and 50 dB. The target speech starts 500 ms before the competing speakers.

All subjective data is recorded accordingly ITU-T Rec. P.800 [13] with 13 male and 11 female normal-hearing subjects (except for the echo condition with 17 males and 7 females).

3. Results

Examples of the posteriorgrams computed with the DNN trained with the full training data (without sMBR training) are shown in Figure 1. This figure shows a sample of the audio file with the worst subject rating for each condition to highlight differences between conditions. The relation of the MOS to the MTD, as well as to the baseline predictions are shown is Figure 2.

The DNN-based model shows high correlations (Pearson's r) between MTD with the MOS for the first four conditions. All correlations are highly significant ($p \leq 0.0002$, two-sided distribution), with the exception of the fifth condition (competing speaker) for which the correlation is low and insignificant (p=0.5). This can be attributed to the DNN not distinguishing between the target and the competing speakers since the overall approach is based on speaker-independent ASR. We assume that the DNN activates the currently dominant phoneme from the multi-speaker mixture — in this case a three-speaker mixture — which would result in clear activations (which are also visible in Figure 1, top right panel). At the same time, a high word error rate would be obtained in ASR tests since the phoneme sequences are incoherent. Since the DNN-based system is not in principal suitable to predict the speech quality for this condition, it is excluded from all further analyses with the DNN-based model.
The baseline systems correlation are all significant ($p \leq 0.05$) except the correlation for the competing speaker condition with the ANIQUE+ model ($p = 0.069$). For the additive noise condition, both baseline systems reach slightly higher correlations, but for the remaining three conditions (clip, echo and chop) the DNN-based system clearly outperforms the two baseline systems in terms of correlation with the subjective ratings. This results in an average correlation of $\bar{r} = 0.71$ with the ITU-T P.563, $\bar{r} = 0.75$ with ANIQUE+ and $\bar{r} = 0.87$ with the DNN-based model.

Figure 3 shows the effect of the amount of training data on model predictions. The 100% data points correspond to the correlation values shown in Figure 2. While the ASR-based systems performance for the clipping condition is already on a very high level with 20% of the training data, we find an average improvement with increasing training data for the remaining three conditions (echo, chop, noise). The echo condition has the biggest performance increase in the first step from 20% to 40% with 5.7% rel. improvement. For the chop condition the

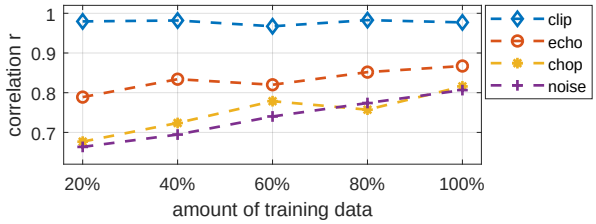

Figure 3: *SQ correlation between the perceived speech quality and the MTD in dependency of the amount of ASR training data.* 100% *corresponds to the full WSJ1 data with* $81.27h$.

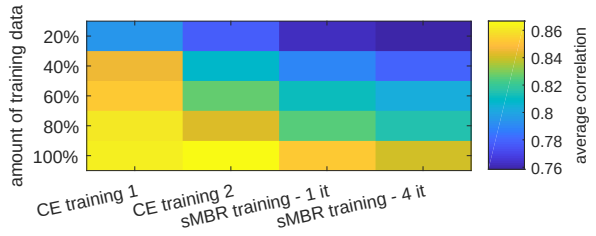

Figure 4: *Average correlation over the four working conditions (clip, echo, chop, noise) in dependency of the amount of data during training and the different training steps of the neural network. Between every column of the matrix the alignments are re-generated with the previous trained model.*

performance increases the most until the 60% dataset and for the noise condition the performance increases constantly over all data sets with overall 21.6% rel. improvement.

The results with modified training procedures are presented in Figure 4. The averaged values from Figure 3 correspond to the second column of this figure, which corresponds to standard training as described above.

In all training steps, $\bar{r}$ increases (10.6% relative improvement on average) with increasing training data size. Furthermore, the prediction performance *de*creases with ongoing training steps of the neural network. The only exception from this trend is the second frame cross entropy (CE) training with 100% of the data, which results in a slightly better performance. Nevertheless, the discriminant sMBR training on this data also degrades the correlation of the MTD with the MOS.

Note that the worst performing DNN-based model (20% data, five iterations of sMBR discriminant training) still reaches a better average correlation $\bar{r} = 0.76$ than the two baseline systems (ITU-T P.563: $\bar{r} = 0.71$, ANIQUE+: $\bar{r} = 0.75$).

4. Discussion

In this study, both the amount of training data as well as the type of training were found to influence the ASR-based SQ model. The best model performance is obtained for CE training 2 with 100% of the data which uses DNN-based alignments (Figure 4. For smaller amounts of data, the best models are found with GMM-based alignments (CE training 1) which can be attributed to the fact that good GMM performance can be obtained with smaller amounts of data which is in contrast to the DNN model employed here. The application of sMBR training consistently degrades model performance independently of the amount of training data. The sequence-discriminant training tries to increase the likelihood ratio between the correct and incorrect class (in this case triphone state). Hence, this results in fewer false alarms in the posteriorgram. The basic assumption of our model is that in noisy or otherwise degraded conditions the posteriorgram smears and the KL-divergence decreases consequentially. With sequence-discriminant training, this effect is reduced and the correlation of the MTD with the MOS decreases. Nevertheless, the best average correlations in this study were obtained by using the full data set (100% Wall Street Journal). It seems likely that the correlation curve shown in Figure 3 is not saturated and therefore higher correlations could be obtained with a larger amount of training data. Prediction in the noise condition has the largest potential for improvement since it exhibits the lowest absolute MTD values and the lowest average correlation when omitting the condition with the competing speaker, which is per se problematic for our model, as described in Section 3.

Additional studies should be conducted to analyze if the correlation values saturate when increasing the size of the training set, or if a local optimum in terms of model quality is reached for a specific amount of data. The latter case could occur if the model becomes too robust to certain distortions so that the posteriorgram is not easily degraded and therefore less informative than a model that was exposed to fewer training samples. Generally, it seems that the development of deep learning has brought speech technology to a point at which it is very interesting for modeling human speech perception. Xiong and colleagues have closed the human-machine gap for conversational telephone speech [23], which indicates that humans and machines are similarly affected by distortions in single-channel data. Our results also show that the DNN output is robust enough to differentiate between severely distorted signals but (still) sufficiently fragile so that small degradations affect the DNN output.

A limitation of our current approach is that it is not suited to predict SQ in the presence of a competing talker, since it is tailored to speech and based on a speaker-independent training set. Potentially, this limitation could be addressed by creating speaker-specific acoustic models and adapting the DNN to a target speaker, which in turn would require collecting target speech samples before test time.

5. Summary

This paper presented Deep machine listening for Estimating Speech Quality (DESQ). The model is trained as regular ASR system, but does not require a decoding step with an HMM when it is applied. Instead, the phoneme representations obtained from the DNN are quantified, which was shown to relate to the perceived SQ for distorted VoIP communication as represented in the TCD-VoIP database. The DNN-based model fails to predict SQ in the presence of background speakers, which can be attributed to its speech-specific, speaker-independent design. In four other conditions, it produces an average correlation of $r = 0.87$ and outperforms two baseline SQ models, i.e., ITU-T P.563 and ANIQUE+. We found its predictive power to increase with larger amounts of training data, while sMBR sequence-discriminant training was not beneficial for SQ prediction with the proposed model.

6. Acknowledgements

This research was supported by the DFG (Cluster of Excellence 1077/1 Hearing4all; URL: http://hearing4all.eu and the CRC TRR 31, Transfer Project T01).

7. References

[1] ITU-T, "Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," *International Telecommunication Union, CH-Geneva*, 2011.

[2] ——, "Recommendation P.863: Perceptual Objective Listening Quality Assessment (POLQA)," *International Telecommunication Union, CH-Geneva*, 2011.

[3] ——, "Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," *International Telecommunication Union, Geneva*, 2004.

[4] D.-S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007. [Online]. Available: https://doi.org/10.1002/bltj.20228

[5] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1935–1947, 2006.

[6] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech Quality Estimation: Models and Trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, nov 2011. [Online]. Available: http://ieeexplore.ieee.org/document/6021874/

[7] R. Huber, C. Spille, and B. T. Meyer, "Single-ended prediction of listening effort based on automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1168–1172, 2017. [Online]. Available: https://doi.org/10.21437/Interspeech.2017-1360

[8] R. Huber, M. Krüger, and B. T. Meyer, "Single-ended prediction of listening effort using deep neural networks," *Hearing Research*, vol. 359, pp. 40–49, 2018. [Online]. Available: https://doi.org/10.1016/j.heares.2017.12.014

[9] R. Huber, J. Ooster, and B. T. Meyer, "Single-ended Speech Quality Prediction Based on Automatic Speech Recognition," *Journal of the Audio Engineering Society*. [Online]. Available: https://doi.org/10.17743/jaes.2018.0041

[10] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, may 2013, pp. 7423–7426. [Online]. Available: http://ieeexplore.ieee.org/document/6639105/

[11] S. H. Mallidi, T. Ogawa, and H. Hermansky, "Uncertainty estimation of DNN classifiers," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 2015, pp. 283–288. [Online]. Available: https://doi.org/10.1109/ASRU.2015.7404806

[12] N. Harte, E. Gillen, and A. Hines, "TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications," *2015 7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, 2015. [Online]. Available: https://doi.org/10.1109/QoMEX.2015.7148100

[13] ITU-T, "Recommendation P.800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, 1996.

[14] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, Lyon, 2013, pp. 2345–2349.

[15] B. T. Meyer, S. H. Mallidi, A. M. Castro Martinez, G. Paya-Vaya, H. Kayser, and H. Hermansky, "Performance monitoring for automatic speech recognition in noisy multi-channel environments," in *IEEE Workshop on Spoken Language Technology*, 2016, pp. 50–56. [Online]. Available: https://doi.org/10.1109/SLT.2016.7846244

[16] 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. Veselý, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011, pp. 1–4. [Online]. Available: https://doi.org/10.1017/CBO9781107415324.004

[17] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998. [Online]. Available: https://doi.org/10.1006/csla.1998.0043

[18] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2. IEEE, 1998, pp. 661–664. [Online]. Available: https://doi.org/10.1109/ICASSP.1998.675351

[19] A. R. Mohamed, G. G. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012. [Online]. Available: https://doi.org/10.1109/TASL.2011.2109382

[20] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and Others, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. [Online]. Available: https://doi.org/10.1109/MSP.2012.2205597

[21] ITU-T, "Recommendation P.341: Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals," p. 30, 2011. [Online]. Available: http://www.itu.int/rec/T-REC-P.341-201103-I/en

[22] European Telecommunications Standards Institute, "Speech quality performance in the presence of background noise - part 1: Background noise simulation technique and background noise database," Tech. Rep. ETSI EG 202 396-1, 2008.

[23] 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*, no. February, 2016. [Online]. Available: http://arxiv.org/abs/1610.05256