INTRUSIVE AND NON-INTRUSIVE PERCEPTUAL SPEECH QUALITY ASSESSMENT USING A CONVOLUTIONAL NEURAL NETWORK

Hannes Gamper, Chandan K A Reddy, Ross Cutler, Ivan J. Tashev, Johannes Gehrke

Microsoft Corporation,
One Microsoft Way, Redmond, WA 98052, USA
{hannes.gamper, chkarada, rcutler, ivantash, johannes}@microsoft.com

ABSTRACT

Speech quality, as perceived by humans, is an important performance metric for telephony and voice services. It is typically measured through subjective listening tests, which can be tedious and expensive. Algorithms such as PESQ and POLQA serve as a computational proxy for subjective listening tests. Here we propose using a convolutional neural network to predict the perceived quality of speech with noise, reverberation, and distortions, both intrusively and non-intrusively, i.e., with and without a clean reference signal. The network model is trained and evaluated on a corpus of about ten thousand utterances labeled by human listeners to derive a Mean Opinion Score (MOS) for each utterance. It is shown to provide more accurate MOS estimates than existing speech quality metrics, including PESQ and POLQA. The proposed method reduces the root mean squared error from 0.48 to 0.35 MOS points and increases the Pearson correlation from 0.78 to 0.89 compared to the state-of-the-art POLQA algorithm.

Index Terms— Speech quality, mean opinion score, PESQ, POLQA

1. INTRODUCTION

In communication systems the speech signal is affected by the noise and reverberation in the transmitting room. Most systems use some form of speech enhancement algorithms, including noise reduction, echo cancellation, and de-reverberation. While these algorithms aim primarily at enhancing the speech signal, they may introduce distortions. Compressing and transmitting the speech signal may cause additional artifacts. The resulting combination of noise, reverberation, distortions and artifacts affects the perceived speech quality. Numerical methods for evaluating speech quality, including the mean squared error (MSE), signal-to-noise ratio (SNR) or the signal-to-distortion ratio (SDR), provide a quantitative metric for the reproduction error, but may not correlate well with human perception of speech quality [1].

The Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) standardized the perceptual evaluation of speech quality in its Recommendation P.800, introducing the mean opinion score (MOS) [2]. A MOS is obtained by asking listeners to evaluate the quality of an audio signal on a scale from one to five and subsequently averaging their scores. This method is slow and expensive and the research community designed several approaches for automating these tests and to serve as computational proxies for MOS listening tests. In *non-intrusive* approaches the signal processing algorithm uses only the contaminated signal for evaluation of sound quality, similarly to human listeners. An example is 3SQM, ITU-T Recommendation P.563 [3]. *Intrusive* methods calculate a perceptually weighted distance between the clean reference and the contaminated signal to estimate perceived sound quality. Intrusive methods are considered more accurate as they provide a higher correlation with subjective evaluations. Representative of this approach is the Perceptual Evaluation of Speech Quality (PESQ), ITU-T Recommendation P.862 [4,5]. In 2005, an extension to ITU-T Recommendation P.862 was proposed due to the advent of wideband telephone services [6, 7]. In 2011, ITU-T Recommendation P.863 introduced the Perceptual Objective Listening Quality Assessment (POLQA) as an update to PESQ to address super-wideband speech services [7, 8].

One common problem of standardized objective algorithms is that they may become obsolete with the emergence of new scenarios, e.g., far field sound capture, new audio compression algorithms, and new speech enhancement models and artifacts. The advances of deep neural networks and their applications in signal processing pave the way to derive objective quality evaluation models that are both accurate and rapidly re-trainable. Prior art includes using a neural network to predict the effect of compression and transmission artifacts on conversational quality [9], using tree-based regression for non-intrusive estimation of speech quality and intelligibility [10], and using a fully connected network to blindly estimate the speech transmission index [11]. Here we build on our work on non-intrusive speech quality estimation [12] and propose a convolutional neural network (CNN) to evaluate the perceptual quality of noisy, reverberant speech samples both intrusively and non-intrusively. Results indicate that for the tested scenarios, the proposed method outperforms existing quality metrics, including PESQ with wideband extension and POLQA.

2. DATA GENERATION AND SUBJECTIVE RATING

The data generation and labeling procedure is outlined in [12]. We use a corpus of 2010 clean speech samples with an equal distribution of male, female, and child voices and a sampling rate of 16 kHz. Each sample is approximately 20 seconds long and consists of three utterances. The samples are normalized to -23 dB FS, before applying a random Gaussian gain with a standard deviation of 8 dB to simulate different talker levels. To simulate reverberation, each clean sample is convolved with a single-microphone room impulse response (RIR) drawn randomly from a library of 120 RIRs with reverberation times between 300 and 500 ms and source-to-microphone distances between 0.5 and 3 meters. Some anechoic and close-talk samples were included in the data corpus as well. Ambient recordings of office, home and other environments serve as additive noise for the reverberant speech samples. A randomly selected noise recording is normalized to -43 dB FS before apply-

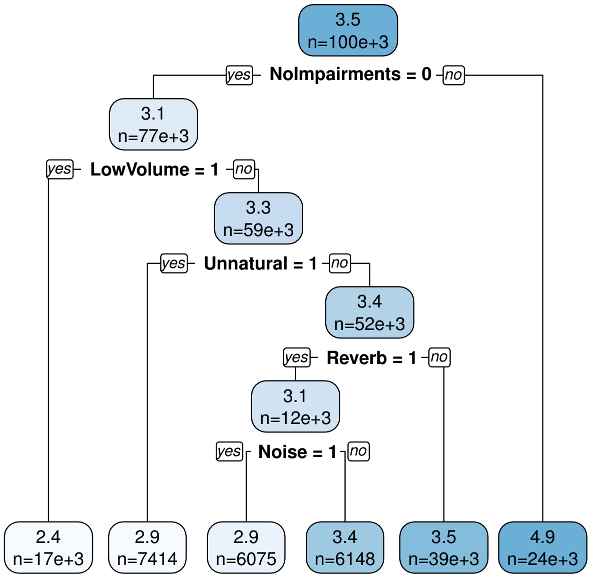


Figure 1: Regression tree of questionnaire responses as predictors of perceived quality judgments. Each leaf node indicates the MOS calculated as the average of all judgments in that node.

ing a random Gaussian gain with a standard deviation of 15 dB. The resulting signal-to-noise ratio (SNR) is limited to [0, 50] dB.

The data corpus totals approximately 10 000 noisy, reverberant samples. Half of these are post-processed with a proprietary noise suppressor and automatic gain control, to account for the effects a speech processing pipeline may have on the perceived quality. All samples were rated subjectively by human judges on a discrete scale from 1 (lowest perceived quality) to 5 (highest perceived quality) according to ITU-T Recommendation P.800 [2]. The ratings of ten judges were averaged to derive the mean-opinion score (MOS) for each sample. A crowd-sourcing platform was used to recruit a total of 654 judges providing on average about 150 ratings each. The judges went through a training phase to familiarize themselves with the types of artifacts present in the data. To eliminate potentially unreliable judges, which may be a bigger concern in crowd-sourced experiments than in laboratory settings [13], a qualification step was used [12]. The judges listened to the samples either via loudspeakers or diotically via headphones.

While the crowd-sourced online labeling process might mimic a more practical listening scenario than a controlled lab experiment, it introduces unknowns, including the level of audio expertise of the judges and the playback setup they use. To better understand what variables might affect the individual judgments, the judges were asked to answer the following multiple choice questions along with their subjective ratings:

- Which device are you using?
  - Headphones
  - Speakers
- Which impairments did you hear?
  - I heard reverb in the call
  - Speech was not natural or sounded distorted
  - I heard noise in the call
  - No impairments
  - Volume was low
  - I could not hear any sound

Judgments that were marked with “I could not hear any sound” were removed from further processing. We performed a regression tree analysis [14] to predict the average perceptual ratings using the multiple choice answers as covariates. Fig. 1 illustrates the resulting regression tree. As can be seen, the analysis did not find the type of playback device to be a determinant affecting the subjective ratings. As expected, samples marked with ”No impairments” received very high average ratings. Samples marked with ”Volume was low”, on the other hand, were rated especially poorly, indicating that the loudness of the speech was an important predictor for perceived quality. Note that while judges were free to choose their preferred playback level at the start of the experiment, they were instructed to keep the level constant for all ratings. “Not natural” sounding samples were rated poorly as well, perhaps indicating that judges used this choice as a “catch-all” for various types of distortions or artifacts. Interestingly, the average rating for samples marked as having impairments but without further markings of specific types of impairments is 3.5, which corresponds to the average MOS of the entire data set. This might suggest that judges were on average satisfied with the speech quality of the samples, while also indicating that many contained certain “impairments” that they could not identify or that were not available among the choices provided.

3. MOS ESTIMATION USING A NEURAL NETWORK

We use a convolutional neural network to estimate the mean-opinion score (MOS) of speech samples with noise, reverberation and distortions.

3.1. Features

The audio signals are processed in frames of 512 samples with a hop size of 160 samples. For each frame exceeding a voice activity threshold, we extract pitch, voice activity, frame energy, and 26 Mel-frequency coefficients, as well as their deltas, for a total of 58 features per frame. Combining the features of the 12 preceding and succeeding frames for both the clean reference signal and the noisy, reverberant test signal yields a feature matrix of size $2 \times 25 \times 58$ per frame.

3.2. Neural network architecture

The spectro-temporal nature of the input features is well-suited for a convolutional neural network (CNN). The proposed architecture is shown in Fig. 2. It consists of four convolutional layers with batch normalization and a kernel size of 2×2, followed by two fully connected layers with a dropout rate of 0.5 and 128 hidden units. The first two CNN layers are followed by max-pooling layers. Their respective kernel sizes and strides are 1×3 and (1, 2) for the first layer, and 3×3 and (2, 2) for the second layer. Rectified linear unit (ReLU) activation is used throughout the network. The network parameters were obtained experimentally via a non-exhaustive search.

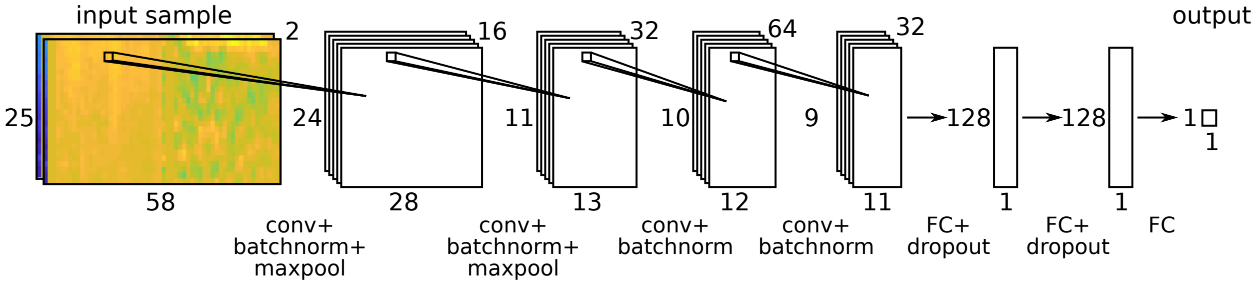
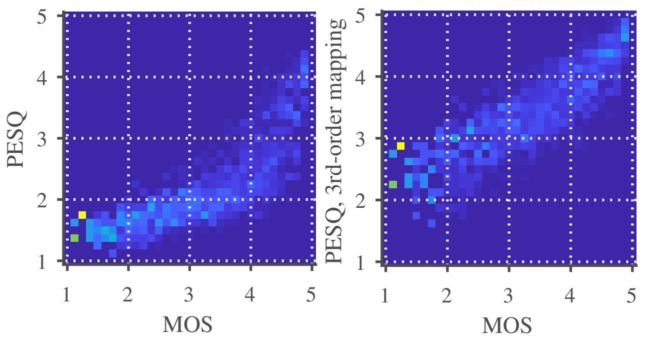
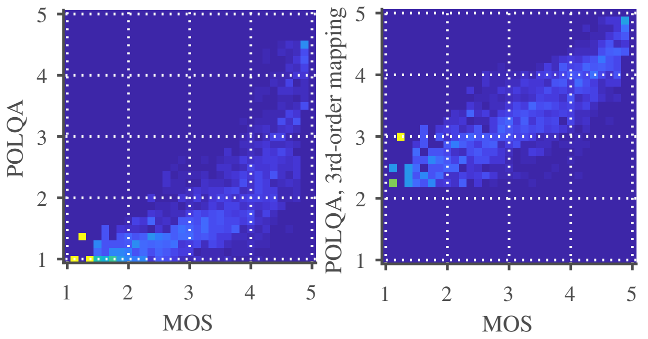


Figure 2: CNN network architecture.



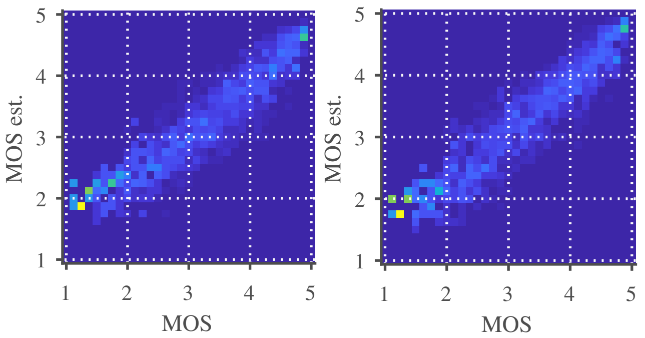
a) b)

Figure 3: Confusion matrix for (a) PESQ and (b) PESQ after 3rd-order polynomial re-mapping.



a) b)

Figure 4: Confusion matrix for (a) POLQA and (b) POLQA after 3rd-order polynomial re-mapping.



a) b)

Figure 5: Confusion matrix for the proposed method: (a) non-intrusive CNN, (b) intrusive CNN.

4. EXPERIMENTAL EVALUATION

We evaluated the proposed MOS estimator using the data corpus described in Section 2. The feature extraction (see Section 3.1) resulted in a total of 5 420 457 feature frames. The data were split by utterances into 70% for training, 15% for validation, and 15% for testing. The network was implemented using the Microsoft Cognitive Toolkit (CNTK) [15]. We formulated the estimation as a regression problem, using a squared error loss function with stochastic optimization [16] and a learning rate of 0.0004. The mean and variance of the input features was normalized using estimates derived from the training set. The network was trained on minibatches of 5000 samples over a total of 500 epochs on a cluster of four GPUs. We explored various strategies for deriving utterance-level estimates from the frame-level output of the network, including taking the mean and median of the frame-level estimates, as well as using a long short-term memory (LSTM) output layer. The best results were achieved by training a classifier to aggregate the frame-level estimates [17].

To compare the performance of intrusive and non-intrusive speech quality estimation, the CNN was trained intrusively on all available data and non-intrusively by discarding the features of the reference signal, i.e., using feature matrices of size $1 \times 25 \times 58$ as input to the network (see Section 3.1). Note that the human judges were not presented with a reference signal, i.e., they judged the speech quality non-intrusively. As evaluation parameters we chose the root-mean-squared error (RMSE) and Pearson's correlation coefficient, ρ [18].

Several objective metrics were calculated for all utterances in the data corpus as a baseline. Fig. 3a illustrates the performance of PESQ with wideband extension as a MOS estimator. For our test utterances, PESQ produces much lower scores than human judges, which points to a mismatch between the assumptions underlying PESQ and the type of distortions present in our data. Fig. 4a presents the results for POLQA, which exhibits similar behaviour to PESQ, likewise underestimating the MOS. ITU-T Recommendation P.862 suggests applying a 3rd-order polynomial to map from PESQ scores to subjective ratings in the case of mismatches [4, 5, 18]. Fig. 3b and Fig. 4b show the confusion matrices between PESQ and POLQA scores and MOS after applying a

Table 1: Utterance-level results for predicting PESQ.

Model	RMSE	ρ
Non-intrusive, CNN + ELM	0.1656	0.9727
Intrusive, CNN + ELM	0.1378	0.9809

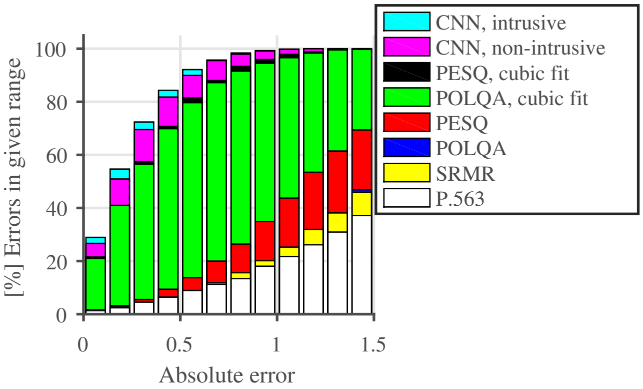


Figure 6: MOS estimation error distribution for P.563 [3], SRMR [19], POLQA [8], PESQ [4], and the proposed CNN+ELM model. Figure adapted from [20].

3rd-order polynomial mapping derived from the test data. For both metrics this mapping seems to improve performance substantially.

The CNN estimates MOS and PESQ jointly by minimizing the sum of their squared errors. PESQ was estimated to ensure that the proposed CNN has sufficient capacity to reproduce a deterministic model like PESQ. Table 1 summarizes the results for predicting PESQ. Interestingly, even the non-intrusive model achieved a relatively low RMSE of 0.1656, and a correlation coefficient of $\rho = 0.9727$, despite the fact that PESQ is an intrusive metric. The intrusive CNN model performed better still, with an RMSE of 0.1378 and $\rho = 0.9809$.

The MOS estimation results for the test data using the proposed CNN are shown in Fig. 5. Both the non-intrusive (Fig. 5a) and intrusive (Fig. 5b) approaches show better correlation with the ground truth than the MOS estimation based on PESQ (cf. Fig. 3) and POLQA (cf. Fig. 4), indicating that the proposed approach offers a better estimate of perceived speech quality for the conditions contained in the test corpus.

Fig. 6 illustrates error ranges for various existing objective quality metrics as well as the proposed CNN estimators. For the proposed approach, close to 100% of estimates fall within an absolute error of approximately 1. PESQ and POLQA obtain acceptable results after re-mapping the raw scores to the ground truth MOS values.

Table 2 summarizes the MOS estimation results in terms of the RMSE and the Pearson correlation coefficient, ρ. For comparison, the results for an individual human judge are estimated by comparing the judgments of each judge to the MOS calculated excluding that respective judge. As can be seen, the individual judgments are relatively noisy when compared to the MOS, with an RMSE of about 1 MOS and a correlation factor just above 0.5. In terms of RMSE, individual judges performed worse than a dummy judge that simply outputs 3.53, i.e., the mean MOS of the test set. Existing objective metrics perform relatively poorly on our data set in terms of RMSE, except for PESQ and POLQA scores after 3rd-order polynomial re-mapping [4, 8]. It should be noted that PESQ is not intended to evaluate distortions due to noise reduction algorithms [4]. The proposed CNN achieves the highest performance in terms of both RMSE and the correlation with the ground-truth MOS, slightly outperforming the previously proposed MLP-based architecture [12]. A statistical significance test for comparing correlation coefficients, as outlined in ITU-T Recommendation P.1401 [18], indicates that the correlation coefficients of PESQ after 3rd-order mapping, $\rho = 0.7824$, and the non-intrusive MOS estimation proposed here, $\rho = 0.8792$ are statistically significantly different ($N = 1499$, $p < 0.0001$). Intrusive estimation seems to outperform the non-intrusive approach, albeit only marginally, perhaps due to the fact that the human judges operated non-intrusively. A statistical significance test [18] does not indicate a statistically significant difference between the correlation coefficients of the proposed intrusive and non-intrusive methods ($N = 1499$, $p = 0.1580$).

Table 2: Utterance-level results for predicting MOS.

Model	RMSE	ρ	
Individual human judge	1.0304	0.5344	
Dummy judge (MOS = 3.53)	0.7734	0	
P.563 [3]	1.9069	0.3658	
P.563, 3rd-order mapping	0.7165	0.3765	
SRMR [19]	1.6273	0.6237	
SRMR, 3rd-order mapping	0.6011	0.6292	
POLQA [8]	1.6306	0.7247	
POLQA, 3rd-order mapping	0.4986	0.7644	
PESQ [4]	1.3118	0.7441	
PESQ, 3rd-order mapping	0.4816	0.7824	
Non-intrusive, MLP + ELM [12]	0.3878	0.8668	$*$
Non-intrusive, CNN + ELM [**proposed**]	0.3742	0.8792	
Intrusive, CNN + ELM [**proposed**]	**0.3546**	**0.8904**	$+$

$^*p < 0.0001$, $^+p >= 0.05$

We hypothesize that further performance gains could be achieved by using more training data to train a more powerful model and additional human judges to stabilize the ground truth MOS estimates.

5. CONCLUSION

We propose a convolutional neural network (CNN) for estimating the perceptual quality of speech in noise and reverberation emulating practical telephony and voice scenarios. A comparison with existing objective metrics illustrates their potential shortcomings when tested on our data set. The proposed CNN is shown to estimate the mean opinion score (MOS) successfully both intrusively and non-intrusively. For our data set, the proposed model achieves a root-mean-squared estimation error of less than 0.4, and a Pearson correlation coefficient of 0.89, thus outperforming PESQ and POLQA, both of which operate intrusively. A comparison with alternative metrics, including PEMO-Q [21] and HASQI [22], as well as an analysis of the proposed model's ability to generalize to unseen types of speech distortion is left for future work.

6. REFERENCES

[1] E. Paajanen, B. Ayad, and V. Mattila, "New objective measures for characterisation of noise suppression algorithms," in *Proc. IEEE Workshop on Speech Coding*, Sep. 2000, pp. 23–25.

[2] ITU-T, *Recommendation P.800: Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, Feb. 1998.

[3] ITU-T, *Recommendation P.563: Single-ended method for objective speech quality assessment in narrowband telephony applications*, ITU-T Recommendation P.563, 2004.

[4] 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*, ITU-T Recommendation P.862, Feb. 2001.

[5] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2001, pp. 749–752.

[6] ITU-T, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Recommendation P.862.2, 2005.

[7] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I – temporal alignment," *J. Audio Eng. Soc*, vol. 61, no. 6, pp. 366–384, 2013.

[8] ITU-T, *Recommendation P.863: Perceptual Objective Listening Quality Assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals*, ITU-T Recommendation P.863, Jan. 2011.

[9] A. P. C. da Silva, M. Varela, E. d. S. e Silva, R. M. Leão, and G. Rubino, "Quality assessment of interactive voice applications," *Computer Networks*, vol. 52, no. 6, pp. 1179–1192, 2008.

[10] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84 – 94, 2016.

[11] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, April 2018, pp. 591–595.

[12] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2019, pp. 631–635.

[13] M. D. Smucker and C. P. Jethani, "The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior," in *Proc. SIGIR Workshop on crowdsourcing for information retrieval*, 2011.

[14] T. Therneau, B. Atkinson, B. Ripley, and M. B. Ripley, "Package 'rpart'," *Available online: cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed April 2019)*, 2015.

[15] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[17] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme Learning Machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[18] ITU-T, *P.1401: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, ITU-T Recommendation P.1401, 2012.

[19] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[20] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II: Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, 2002.

[21] 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, Nov 2006.

[22] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, 2014.