Non-Intrusive POLQA Estimation of Speech Quality using Recurrent Neural Networks

Dushyant Sharma*, Aidan O. T. Hogg†, Yu Wang‡, Amr Nour-Eldin* and Patrick A. Naylor†
* Nuance Communications
† Department of Electrical and Electronic Engineering, Imperial College London, UK
‡ Department of Engineering, University of Cambridge, UK
Email: dushyant.sharma@nuance.com

***Abstract*—Estimating the quality of speech without the use of a clean reference signal is a challenging problem, in part due to the time and expense required to collect sufficient training data for modern machine learning algorithms. We present a novel, non-intrusive estimator that exploits recurrent neural network architectures to predict the intrusive POLQA score of a speech signal in a short time context. The predictor is based on a novel compressed representation of modulation domain features, used in conjunction with static MFCC features. We show that the proposed method can reliably predict POLQA with a 300 ms context, achieving a mean absolute error of 0.21 on unseen data. The proposed method is trained using English speech and is shown to generalize well across unseen languages. The neural network also jointly estimates the mean voice activity detection (VAD) with an F1 accuracy score of 0.9, removing the need for an external VAD.**

***Index Terms*—speech quality estimation, POLQA estimation, deep neural networks.**

I. Introduction

In recent years, speech quality assessment has gained increasing importance in multi-media and telecommunication applications, as well as performance evaluation of speech coders [1], text-to-speech synthesis [2] and automatic speech recognition (ASR) [3]. Speech quality can be defined as the subjective opinion of a given speech signal, as determined by the listener [4]. This definition highlights the subjective nature of this measure. Factors contributing to degradation of speech quality include additive noise, reverberation due to convolution with a Room Impulse Response (RIR) [5], coding artefacts, clipping and transmission errors.

Traditionally, extensive listening tests are required for speech quality assessment. Such tests are expensive and time intensive procedures [6]. The ability, therefore, to be able to carry out speech quality assessment in an objective manner without the need for subjective testing is of great interest. The quality of a speech signal can be estimated objectively using either intrusive or non-intrusive methods. The more common approach uses intrusive methods which rely on both the clean and degraded signal to calculate the speech quality. The previous ITU-T recommendation was Perceptual Evaluation of Speech Quality (PESQ) [7]. However, this has now been superseded by the Perceptual Objective Listening Quality Analysis (POLQA) [8] model. These models both work by aligning the clean and degraded signals before performing an auditory transform. The dissimilarity between the clean and degraded signals is then evaluated before being mapped onto the subjective mean opinion score (MOS) [9] scale. The disadvantage of intrusive methods is that a clean reference signal is required whereas, in practice, the clean signal is typically unavailable [10].

Non-intrusive methods work without the need for a clean reference signal, and the current ITU-T industry-standard recommendation for non-intrusive speech quality assessment is P.563, which was designed for narrow-band telephony applications [11]. This standard uses a number of features from the degraded speech to estimate the quality score directly on the MOS scale. A number of data-driven methods currently exist that extract features from the speech signal and use a previously trained model to map the features to a quality score. Dubey *et al.* [12] train a Gaussian Mixture Model. The work in [13], [14] deploys a classification and regressions tree (CART) based model using short-term features characterized by their statistics along with long-term spectral deviation features over an entire utterance. More recent methods rely on deep learning [15] [16] [17]. Results demonstrate that a mapping function can be learnt effectively for the task of speech quality assessment. Yang et al. [18] have also shown that a deep neural network (DNN) is capable of predicting POLQA scores using real-time control protocol (RTCP) features.

This paper proposes training a recurrent neural network (RNN) on a large training data set to estimate the POLQA score non–intrusively, on a short-time basis, using speech features. To the best of our knowledge, this is a first non-intrusive POLQA estimator that predicts POLQA on a short-time basis. Unlike previous work [18] that requires RTCP features, our method works with speech features extracted from the decoded speech signals, that is, without the need for the transmission packet information. Another novel contribution of this work is the joint estimation of voice activity and POLQA using a multi-task RNN.

The remainder of the paper is organised as follows. In Section II we present the proposed algorithm and the baseline method from [13], followed by a description of the data sets and evaluation metrics in Section III. We finish with results in Section IV and conclusions in Section V.

978-9-0827-9703-9/19/$31.00

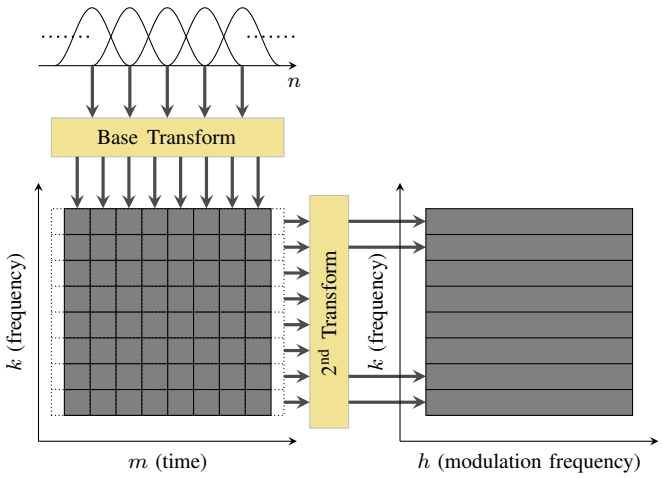

Fig. 1. Overview of the procedure for computing the modulation spectrogram representation of a speech signal.

II. Methods

A. Baseline

As a baseline for this paper and given that there are no other recent comparable approaches, we consider our previous work on non-intrusive PESQ estimation [13]. There we used a CART regression tree and a set of 112 features to model the PESQ score of an utterance of speech. In this paper, we re-train this baseline method for the task of non-intrusive POLQA prediction, on the same training data as our proposed method and refer to it as QOS.

B. MFCC Features

The Mel-Frequency Cepstral Coefficients (MFCC) are a common feature set for a number of speech processing applications, including ASR [19]. In our method, we use a sample rate of $f_s = 8\,\text{kHz}$ with a pre-emphasis factor of 0.97 and extract 24 MFCC features every 10 ms using a 25 ms analysis window. These are appended to 24-dimensional MDCC features described and defined in the following subsection.

C. MDCC Features

Here we describe the modulation domain feature set used in our algorithm (denoted as MDCC). There is much interest in amplitude modulation domain processing of speech because low-frequency modulations of speech are the fundamental carrier of linguistic information. This representation has been exploited in areas such as speech coding [20], recognition [21], [22], enhancement [23], [24] and speech intelligibility modelling [25], [26]. This approach is also motivated by studies of the human auditory system [27] that point to analyses and separation of different acoustic objects in this domain. A number of modulation domain methods have been proposed in the fields of emotion detection [28] and speech quality estimation [29], [30]. In the field of speech and acoustics, there are a number of definitions of the amplitude modulation spectrum that differ in the sub band decomposition. The procedure we use to define the modulation domain representation of the signal is shown in Fig. 1. The first transform in the figure is applied to the time-domain signal to decompose it into sub band signals (using linear frequency spacing). The temporal envelope within each band is then computed. Without loss of generality, we will use Fourier transform to mean a Discrete Fourier transform and compute it using the Fast Fourier Transform (FFT) algorithm in the following. Denoting the length N time domain signal as $s(n)$, its short-time Fourier transform (STFT) is calculated as

$$S_k(m) = \sum_{n=0}^{N} s(n) w_a(n - mL) e^{-j\frac{2\pi}{N}k}, \tag{1}$$

where m is the short-time frame (typically 30 ms duration), which in the context of modulation domain processing is defined as an 'acoustic frame' [20], $w_a(n)$ is the window applied on each frame and L is the acoustic frame increment in samples. After the first STFT, the temporal envelope of each acoustic frequency band k (also called the modulating signal) is obtained as the magnitude of the transformed signal, $|S(m,k)|$. If the increment of the acoustic frame is $t = 1/f_s$ s then the sampling frequency of the modulating signal is $1/t$ Hz. Since the highest amplitude modulation frequency for the human auditory system is 256 Hz for the cochlear nucleus [31], $t = 4\,\text{ms}$ could be chosen. To obtain the modulation spectrum, a window function $w_m(n)$ is used to segment the amplitude envelope of each frequency bin and a second STFT is performed on each modulation frame, as

$$S_l(k,h) = \sum_{m=0}^{M} |S_k(m)| w_m(m - lL) e^{-j\frac{2\pi}{H}h}, \tag{2}$$

where H is the number of modulation frequency bins and m is the modulation frame index. Because the features are extracted independently for each modulation frame, in the following description we omit the index of the modulation frame, l, for clarity. The modulation spectrogram is thus given by $P(k,h) = |S(k,h)|^2$. An example of the modulation spectrum for one frame of a speech utterance is shown in Fig. 2 (top). In order to compress the information in the modulation spectrum, we propose a novel final step. We apply a two dimensional DCT-II (2D-DCT) on the modulation spectrogram $P(k,h)$ to produce a set of DCT coefficients D as follows.

$$D(\Omega,\Phi) = \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} W P(k,h) C(k,h,\Omega,\Phi),$$

where $C(\pi,k,h,\Phi) = \cos\left[\frac{\pi}{K}\left(k+\frac{1}{2}\right)\Omega\right]\cos\left[\frac{\pi}{H}\left(h+\frac{1}{2}\right)\Phi\right]$ and $W = \sqrt{(1/K)}\sqrt{(1/H)}$ for $\Omega = 0,...,K-1$ and $\Phi = 0,...,H-1$. An example compressed spectrum is shown in Fig. 2 (bottom). From empirical experiments, it was found that only a few coefficients from the upper triangle of $D(\Omega,\Phi)$ are sufficient in capturing most of the variation in the modulation spectrum. 24 MDCC coefficients are the following set, $[D(1,1:21), D(2,1:3)]$.

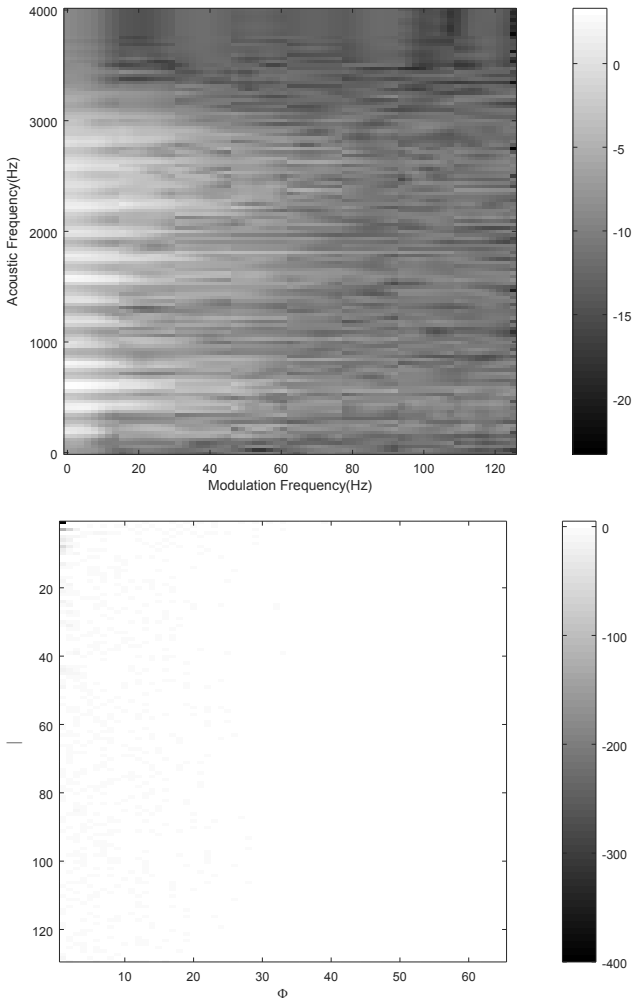

Fig. 2. Modulation spectrum (top) of a 128 ms modulation frame of speech from [32]. An example of the modulation spectrum after applying the 2D-DCT (bottom).

D. Recurrent Neural Network Model

The 24 MFCC and 24 MDCC features are standardized (zero mean and unit standard deviation) before being used to train an RNN that jointly estimates VAD and POLQA as depicted in Fig. 3. The context length is a parameter that can be varied from 10 ms up to the maximum length of an utterance in the data and represents the context available as input to the multi-task RNN. For this paper we experimented with lengths from 100 ms (10 frames) to 1.2 s (120 frames). It will be shown in the results section that the proposed method can accurately estimate the POLQA scores and the VAD scores with a context greater than or equal to 30 frames. The RNN for this context size has an input layer of size 48×30 followed by three layers of long short term memory (LSTM) cells [33] in a $40 \times 21 \times 16$ topology (for each time step). In the last hidden layer, the activations are averaged over a window of 10 frames, which is then followed by an output layer with two nodes. One of the output nodes estimates the POLQA score and the second node estimates the Mean Voice activity Posterior (MVP). The MVP is obtained by running a simple

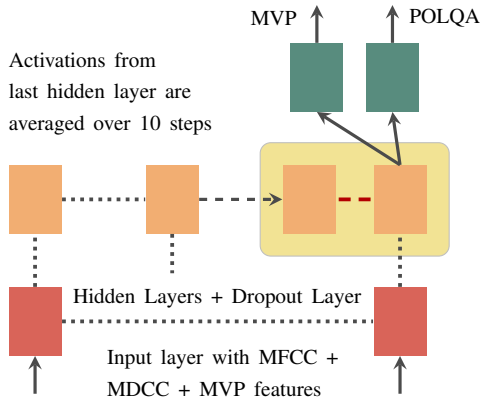

Fig. 3. Multi-task RNN topology used for the joint estimation of the mean voice posterior (MVP) and POLQA scores.

VAD from [34] on the clean speech used in the synthesis of the training data (described in more detail in Section III). The VAD outputs a posterior probability of the frame being voiced and these scores are averaged over the context window used for the RNN model, resulting in a mean voice activity posterior, which becomes the target of the second output of the RNN model. During test time, the MVP can be used as a confidence measure and used to prune the POLQA scores, accordingly when the MVP is lower than a threshold, we ignore the POLQA score on the grounds that there isn't enough speech in the context window to reliably estimate POLQA. In our experiments, we found a threshold of 0.5 to be a good choice. This parameter is not tuned. The multi-task RNN is trained for 10 epochs on the training data using a mini-batch size of 1000 observations with an Root Mean Square (RMSE) error metric ($RMSE_{POLQA} + RMSE_{MVP}$). The Adam [35] optimiser is used with an initial learning rate of 0.001. In order to avoid over-fitting the model parameters, dropout is applied before the last hidden layer of the RNN. In this paper, none of the RNN hyper-parameters were tuned for this particular problem or data and one can expect further gains in performance from doing such tuning.

III. Data and Metrics

A. Training Data

The training data for the proposed and baseline methods is based on the clean 100 hrs training partition of the Librispeech corpus [36], which is derived from audiobooks read by a large number of speakers. This forms the base material for the training data set. The base material is artificially corrupted by RIR convolutions (measured and artificially generated) and additive noise. This noisy and reverberant set forms a block of data, with each block sampling the available noise and RIR using uniform sampling. In each block, 128 RIR are sampled with a T60 in the range [0.1—1.25 s] and C50 [37] in the range [0—30 dB]. Also, 100 noise realizations including babble, household, ambient, street and vehicle are included with SNR in the range [0—30 dB]. Care is taken to ensure that street and vehicle noise types are not applied to speech that has

an RIR convolution applied. Each block is then processed by 10 CODEC conditions, covering the following: G.711 a-law, GSM-FR, G.729 [A and B], GSM-AMR [4.75, 6.70, 7.40 and 12.2 kbps] and linear PCM. The resulting training set contains nearly 100 hrs of processed data. All signals are down-sampled to 8 kHz and filtered appropriately to represent narrowband telephony data.

B. Test Data

The test data follows the same processing procedure as described in the preceding subsection on training data but with a different set of speech base material, RIR and noise sources, ensuring no overlap between the training and test data sets in terms of speaker, noise or reverberation. In order to better understand the generalization performance of the proposed method, we created test sets in three different languages—English, French and Japanese. For English, we created two sets of base material, the first is the test-clean partition of Librispeech [36] and the second is the clean English partition of the P.23 [38] database. We also used clean French and Japanese speech material from the P.23 [38] database. These test sets were processed similarly to the training data, with each block constructed with a smaller set of 32 RIR and 36 noise sources. The same 10 CODEC conditions were applied, resulting in approximately 15 hrs of test data per set.

C. Evaluation Metrics

In the following, P_e and P_t are the estimated and true POLQA scores, respectively and the error in estimating a sample is defined as $E(n) = P(n)_e - P(n)_t$.

Metric	Description
Pearson Correlation Coefficient (R)	Measures the dependence between P_e and P_t and takes a value in the range [-1,1].
Root Mean Square Error (RMSE)	$\text{RMSE} = \sqrt{\sum_{n=1}^{N} \frac{1}{n} E(n)^2}$
Mean Absolute Difference (MAD)	$\text{MAD} = \sum_{n=1}^{N} \frac{1}{n} \lvert E(n) \rvert$
F1 Score (F1)	Measures the accuracy of the MVP estimation (which becomes a classification task when an appropriate threshold is applied on a segment) and is defined as $F1 = \frac{2TP}{2TP + FP + FN}$, where TP is the true positive rate, FP and FN are the false positive and negative rate respectively.

IV. Results

A summary of the mean performance of the proposed method on all the test sets for different context lengths is presented in Fig. 4. It can be observed that the proposed method outperforms the baseline QOS method for all context lengths evaluated (from 100 ms to 1.2 s). The average RMSE across all test sets is 0.29 for the proposed method for a 300 ms context window and this represents a 38.3% relative reduction in error over the baseline. Even with a context of just 100 ms, our method has an RMSE of 0.38, that is 19.1% relative lower than QOS. Also it can be seen that the performance is very close to the performance obtained when using an oracle VAD during testing. Here the RMSE across all test sets with ground truth VAD is 0.26 for a 300 ms window and the proposed estimate is only 11.5% worse. The addition of the proposed MDCC features helps reduce the RMSE by nearly 31% relative to the MFCC only based system (see Tables I and II). Another important question is the ability of the proposed method to generalize to unseen languages. As described in Section II, the proposed method is trained on English speech material. Table II presents the detailed results for POLQA and MVP estimation for a 30 frame context for the different test sets using the full feature set. It can be seen that the performance is consistent across English, French and Japanese languages, both in terms of POLQA and VAD estimation.

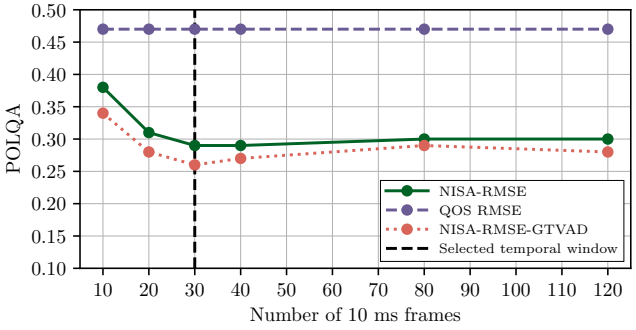

Fig. 4. POLQA prediction RMSE (averaged over all test sets) using the proposed method (middle line), the baseline QOS method (top line) and the proposed method with oracle VAD used for testing (bottom line).

TABLE I
DETAILED RESULTS FOR POLQA ESTIMATED USING A 300 MS TEMPORAL WINDOW AND SEGMENT PRUNING BASED ON THE ESTIMATED VAD POSTERIORS FOR THE DIFFERENT TEST SETS, USING ONLY THE MFCC FEATURES.

	MAD	RMSE	R	VAD F1
Libre	0.35	0.54	0.70	0.86
P23-EN	0.28	0.43	0.67	0.91
P23-FR	0.22	0.31	0.69	0.89
P23-JP	0.24	0.38	0.70	0.89
Mean	0.27	0.42	0.69	0.89

TABLE II
DETAILED RESULTS FOR POLQA ESTIMATED USING A 300 MS TEMPORAL WINDOW AND SEGMENT PRUNING BASED ON THE ESTIMATED VAD POSTERIORS FOR THE DIFFERENT TEST SETS.

	MAD	RMSE	R	VAD F1
Libre	0.26	0.37	0.86	0.89
P23-EN	0.20	0.28	0.84	0.93
P23-FR	0.19	0.26	0.78	0.89
P23-JP	0.18	0.25	0.84	0.91
Mean	0.21	0.29	0.83	0.90

V. Conclusions

We presented a short-time, non-intrusive POLQA estimator of speech quality by exploiting a recurrent neural network and

jointly training it for voice activity estimation. The proposed method uses narrowband MFCC features in combination with features extracted from a novel compressed representation of the modulation spectrum, that help reduce the RMSE of the system by 31.0% relative to the MFCC only system. The use of an RNN topology with these features allows the proposed method to reliably estimate POLQA and VAD with 300 ms of context. The average RMSE across 4 test sets is 0.29 for POLQA and the VAD's F1 accuracy is 0.90. We show that the proposed method is able to generalize well across unseen languages and that the VAD performance is similar to the ground truth. The results demonstrate that in terms of relative RMSE, with a 300 ms context, the proposed method outperforms the baseline method by 38.3%.

REFERENCES

[1] A. E. Coleman, N. Gleiss, and P. Usai, "A subjective testing methodology for evaluating medium rate codecs for digital mobile radio applications," *Speech Communication*, vol. 7, no. 2, pp. 151–166, Jul. 1988.

[2] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Communications Magazine*, vol. 26, no. 10, pp. 36–44, Oct. 1988.

[3] L. F. Gallardo, S. Moller, and J. Beerends, "Predicting Automatic Speech Recognition Performance Over Communication Channels from Instrumental Speech Quality and Intelligibility Scores," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, Aug. 2017, pp. 2939–2943.

[4] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*, ser. Signals and Communication Technology. Berlin Heidelberg: Springer-Verlag, 2005. [Online]. Available: //www.springer.com/gp/book/9783540240952

[5] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," P. A. Naylor and N. D. Gaubitch, Eds. PUB-SV, 2010.

[6] A. Rosenberg and B. Ramabhadran, "Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, Aug. 2017, pp. 3976–3980.

[7] ITU_T_P862_TR, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," INST_ITU_T, Recommendation, Feb. 2001.

[8] "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," INST_ITU_T, Standard, Jan. 2011.

[9] ITU_T_P10, "Amendment 2: Vocabulary for performance and quality of service amendment." INST_ITU_T, Recommendation, 2009.

[10] A. Raja, R. M. A. Azad, C. Flanagan, and C. Ryan, "Real-Time, Non-intrusive Evaluation of VoIP," in *Genetic Programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4445, pp. 217–228.

[11] L. Malfait, J. Berger, and M. Kastner, "P.563 - the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, 2006.

[12] R. K. Dubey and A. Kumar, "Non-intrusive speech quality assessment using several combinations of auditory features," *Int J Speech Technol*, vol. 16, no. 1, pp. 89–101, Mar. 2013.

[13] D. Sharma, L. Meredith, J. Lainez, D. Barreda, and P. A. Naylor, "A non-intrusive PESQ measure," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2014, pp. 975–978.

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

[15] J. Rozhon, M. Voznak, F. Rezac, J. Slachta, and J. Safarik, "A new approach to speech quality assessment based on back-propagation neural networks," *International Journal of Circuits, Systems and Signal Processing*, vol. 10, p. 10, 2016.

[16] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2016, pp. 2315–2319.

[17] E. T. Affonso, R. L. Rosa, and D. Z. Rodriguez, "Speech quality assessment over lossy transmission channels using deep belief networks," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 70–74, Jan. 2018.

[18] H. Yang, K. Byun, H. G. Kang, and Y. Kwak, "Parametric-based non-intrusive speech quality assessment by deep neural network," in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, Oct. 2016, pp. 99–103.

[19] H. Hermansky, J. R. Cohen, and R. M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, 2013.

[20] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[21] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[22] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[23] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[24] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[25] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," *Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[26] M. E. Taffeta and E. F. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput Biol 5(3)*, 2009.

[27] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1593–1602, 1994.

[28] J. Wu and X.-L. Zhang, "An efficient voice activity detection algorithm by combining statistical model and energy detection," *EURASIP J. on Advances in Signal Processing*, vol. 18, 2011.

[29] 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, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[30] D. S. Kim and A. Tarraf, "ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

[31] A.-L. Giraud, C. Lorenzi, J. Ashburner, J. Wable, I. Johnsrude, R. Frackowiak, and A. Kleinschmidt, "Representation of the temporal envelope of sounds in the human brain," *Journal of Neurophysiology*, vol. 84, no. 3, pp. 1588–1598, 2000.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA timit acoustic-phonetic continuous speech corpus CD-rom," INST_NIST, NIST Interagency/Internal Report (NISTIR), Feb. 1993.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[34] D. Sharma, A. Nour-Eldin, P. Harding, S. Karimian-Azari, and P. A. Naylor, "Robust feature extraction from ad–hoc microphones for meeting diarization," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[35] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4 2015.

[37] P. P. Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP J. on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[38] *ITU-T coded-speech database*, International Telecommunications Union (ITU-T) Supplement P.Sup23, Feb. 1998.