A CLASSIFICATION-AIDED FRAMEWORK FOR NON-INTRUSIVE SPEECH QUALITY ASSESSMENT

Xuan Dong and Donald S. Williamson

Department of Computer Science, Indiana University, USA
{xuandong, williads}@indiana.edu

ABSTRACT

Objective metrics, such as the perceptual evaluation of speech quality (PESQ) have become standard measures for evaluating speech. These metrics enable efficient and costless evaluations, where ratings are often computed by comparing a degraded speech signal to its underlying clean reference signal. Reference-based metrics, however, cannot be used to evaluate real-world signals that have inaccessible references. This project develops a nonintrusive framework for evaluating the perceptual quality of noisy and enhanced speech. We propose an utterance-level classification-aided nonintrusive (UCAN) assessment approach that combines the task of quality score classification with the regression task of quality score estimation. Our approach uses a categorical quality ranking as an auxiliary constraint to assist with quality score estimation, where we jointly train a multi-layered convolutional neural network in a multi-task manner. This approach is evaluated using the TIMIT speech corpus and several noises under a wide range of signal-to-noise ratios. The results show that the proposed system significantly improves quality score estimation as compared to several state-of-the-art approaches.

Index Terms— speech quality assessment, objective metrics, convolutional neural networks, multi-task learning

1. INTRODUCTION

The performance of speech enhancement algorithms is often evaluated with objective metrics, since objective metrics provide important information about speech quality and intelligibility in a short-period of time [1]. Objective metrics can be divided into two major categories: intrusive and nonintrusive. Intrusive metrics require the clean speech (or reference) signal during the evaluation process, where these metrics compare a time-frequency (T-F) representation of the enhanced or noisy speech signal to the clean speech signal. Differences between the two signals result in quality and intelligibility scores, where the scores improve with increasing spectral-temporal similarity. Commonly-used intrusive metrics include the perceptual evaluation of speech quality (PESQ) [2], short-time objective intelligibility (STOI) [3], perceptual objective listening quality assessment (POLQA) [4], hearing aid speech quality index (HASQI) [5], and metrics from the blind source separation (BSS) toolkit, signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) [6]. These metrics use signal-processing techniques during the comparison process. Although these metrics have been shown to have correlations with human evaluations [3, 7], the need for a reference signal is a major limitation, since this does not allow evaluation of real-world signals that have inaccessible references.

Nonintrusive metrics, on the other hand, perform evaluations directly on the signal of interest (e.g. noisy or enhanced), without the need for a reference signal [8, 9, 10]. These metrics rely on properties of signals or environmental factors to determine quality and intelligibility scores. Current nonintrusive metrics have many limitations, including: 1) they perform worse than intrusive measures in terms of correlations to human listening evaluations [11, 12]; 2) they have not been thoroughly evaluated in realistic environments that contain many speakers or different types of acoustical noise [13]; and 3) they are only intended for specific-signal types, e.g. over telecommunication networks [14] or for hearing aid applications [15]. As a result, nonintrusive metrics are not often used for assessment. Listening studies involving human participants offer the most accurate way to assess speech, where participants provide a quality rating or try to identify the words in each signal [1, 16]. These studies, however, can be costly and time consuming.

Data-driven approaches have been proposed recently for speech evaluation. These approaches use machine learning techniques, such as hidden markov models (HMM) [17], or classification and regression trees (CART) [18]. More recent approaches use deep learning (autoencoders or deep neural networks (DNNs)) as a means of evaluating speech quality and naturalness [19, 20, 21, 22, 23]. In [24], a full convolutional network is used to estimate the speech transmission index (STI). The authors in [25] utilize a single convolutional layer to predict subjective intelligibility ratings from four listening tests. A frame-level speech quality evaluation model which consists of one bidirectional long short-term memory (BLSTM) layer and two fully connected layers is proposed in [26]. It predicts the PESQ score of a single time frame, and then calculates an utterance-level prediction by averaging frame-level outputs. Recently, [27] uses a DNN-based voice activity detection (VAD) to predict the mean opinion score (MOS) of degraded acoustic signals. The use of machine learning for objective speech evaluation is promising since it enables quick reference-less evaluation, and it allows the metric to learn from data without prior assumptions.

Inspired by the latter deep-learning based metrics, we propose a convolutional neural network (CNN) framework for assessing the perceptual quality of speech. More specifically, we jointly train a CNN to predict the categorical objective ranking and true PESQ score, where PESQ scores are grouped into categorical classes based on pre-defined ranges. Hence, we propose to treat objective speech evaluation as the combination of a classification and a regression task. The two tasks share the same feature extraction layers while each task also has independent modules to achieve specific goals. Learning tasks in parallel while using a shared representation has been shown to be helpful for other multi-task learning problems [28, 29].

This research was supported in part by a NSF Grant (IIS-1755844).

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

Existing approaches do not always perform well in varying environments, and this can occur because a regression-only network cannot develop adequate representations in each environment. The categorical classification task imposes additional restrictions on the model across all environments. This can result in more effective learning, which is evidenced by a reduction in estimation errors as compared to training a single regression model.

Additionally, prior approaches often make frame-level quality predictions, where each frame of the signal is given the utterance-level quality score as a label. This is a major shortcoming, as frame-level scores ($\sim$ over millisecond length windows) are not the same as utterance-level quality scores ($\sim$ over 4-5 seconds), as noisy speech varies much over this time period. Our approach overcomes this drawback as a single quality prediction is made for the utterance.

The rest of this paper is organized as follows. Section 2 describes details of our proposed approach. The experimental setup and results are presented in Sections 3. Section 4 concludes the discussion of the proposed approach.

2. SYSTEM DESCRIPTION

Our utterance-level classification-aided nonintrusive (UCAN) assessment approach uses a multi-layered CNN to predict both the categorical quality rankings of noisy speech and the corresponding objective quality scores. CNNs utilize convolutional and pooling layers to map input features into higher representations that are less sensitive to minor input variations. They offer benefits over traditional feed-forward networks (e.g. DNNs), since CNNs focus on local spectral-temporal connections, by using spatial filters that look at neighboring regions around each T-F unit. Details of the specific architectural components are given below.

2.1. Network architecture

The proposed framework consists of three modules: shared feature extractor, quality-score classification, and absolute quality score prediction. The specific architecture is illustrated in Fig. 1. The CNN feature extractor consists of six convolutional layers (Conv). The first two Conv layers each use 16 kernel filters, the middle two Conv layers use 32 kernel filters, and the last two Conv layers utilize 64 kernel filters. Every kernel filter has a kernel size of 3×3. The leaky version of the rectified linear (LeakyReLU) activation function with a negative slope coefficient $\alpha = 0.1$ is applied to perform nonlinear mapping. Batch normalization (BN) is always performed before the LeakyReLU nonlinearity. A max pooling layer with a 2×2 pooling size follows to subsample every other convolved intermediate output.

Next, the output features of the shared convolutional layers are applied to two separate tasks. In the right branch, which is used for quality-score classification, the features are flattened into a 1-dimensional vector and given as inputs to a two-dense layer subnetwork, which consists of 64 and 32 hidden units, respectively. The outputs of the last dense layer are given to a softmax layer, which produces a distribution over the class labels. Standard back-propagation with Adam optimization is used to minimize the cross entropy (denote as $\mathcal{L}_{cls}$) between the subnetwork outputs and the training labels. In the left branch, which is used for quality-score estimation, the shared features are sequentially processed by a Conv layer with 128 3×3 kernel filters and a 2×2 average pooling layer. Similarly, the convolutional outputs are flatten into a 1D vector and

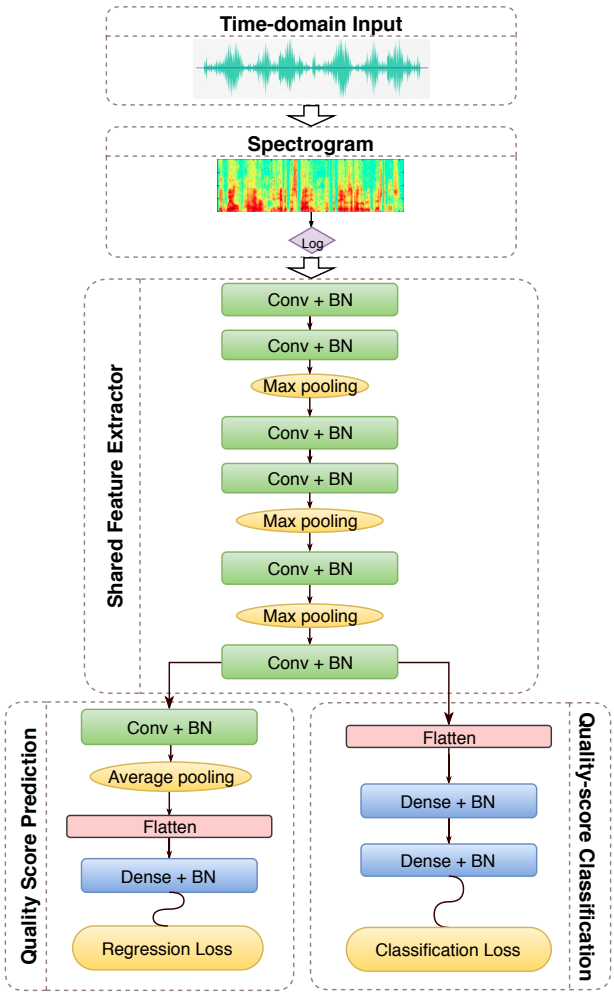

Figure 1: Architecture of the proposed framework with shared convolutional and task-specific fully connected layers.

then fed to a dense layer with 32 hidden units. The last layer applies a linear activation function and outputs the estimated quality score of the inputted speech signal. The mean squared loss (regression loss denoted as $\mathcal{L}_{regr}$) that stems from the left subnet together with the classification loss $\mathcal{L}_{cls}$ are utilized to update the weights of the shared network:

$$\mathcal{L}_{total} = \beta * \mathcal{L}_{cls} + (1 - \beta) * \mathcal{L}_{regr}, \quad (1)$$

where β controls the trade-off between optimizing the network for the classification or regression task. Note that we experimented with different CNN architectures (e.g. number of layers and parameters), but we empirically determined that the proposed architecture performed best.

2.2. Input features

Typically there are two ways to handle input signals of varying length for non-intrusively predicting objective metrics. One approach is to adopt the frame-level magnitude spectrogram (tens of milliseconds) as the input feature, and use the utterance-wise quality score as the training label for each frame. The second approach

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

assigns a single utterance-level score as the label for an input signal. This makes the prediction process difficult, because the input features may differ in size due to differences in signal lengths. This, however, can be addressed by padding or truncating each signal to a fix length. We elect to use the latter approach as utterance-level score prediction is more reliable than frame-level prediction, since frame-level score assignments are often inaccurate.

Our system is performed in the T-F domain using the short-time Fourier transform (STFT). Each signal is first downsampled to a 16 kHz sampling rate. The STFT of each signal is computed using a Hanning window and a 40 ms time frame with 25% overlap between adjacent frames. The fast Fourier transform (FFT) is computed using a 640-point FFT. Most of the speech signals in our experiments have lengths between 3 to 5 seconds. Thus, a temporal length $T = 5$ sec has been chosen as the maximum length of our signals to ensure a fixed-sized CNN architecture. A speech signal is zero-padded if its length is less than T, while the signal is cropped to a length of T otherwise. Finally, the log-spectral magnitude of the STFT with a dimension of 321×166 is applied as the input feature.

2.3. PESQ quality labels

Two training targets are simultaneously applied in our model. One is the quality class of a speech signal, and the other is the corresponding raw PESQ score. PESQ returns scores between -0.5 and 4.5, where higher scores correspond to higher perceptual speech quality [2]. Signals with extremely low or high PESQ scores are infrequently encountered. The observed upper and lower PESQ scores from our experimental dataset are 0.13 and 4.32, respectively. According to this observation, we define three variables for the classification task: the low threshold L_t, the high threshold H_t, and the category bin size B of PESQ scores, which are used to determine how PESQ scores are assigned for the N classes. Denote $\mathcal{S}_{pesq}$ as the raw PESQ score for a particular signal. The PESQ classification label of a given signal is calculated by

$$\text{Class}(\mathcal{S}_{pesq}) = \min(\max\left(1, \text{ceil}\left(\frac{S_{pesq} - L_t}{B}\right)\right), N), \quad (2)$$

where $\text{ceil}(\cdot)$ denotes the ceiling function. Notice that class 1 is assigned if $\mathcal{S}_{pesq}$ is less than L_t, whereas class N is assigned if $\mathcal{S}_{pesq}$ is greater than or equal to H_t. The parameters $N = 20$, $B = 0.2$, $L_t = 0.2$, and $H_t = 4.2$ are used in our experiments.

For each input training signal, a binary vector of length N is constructed that consists of all zeros, except for the label index that has a value of 1. This one-hot vector is supplied to the classification module as the training label. In addition to this, the raw PESQ score that is the regression training target, along with the inputted log-magnitude spectrogram, are used to train the classification-aided framework jointly to predict the categorical ranking and to estimate the quality score in parallel.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental setup

We setup three datasets in our experiments: (1) a seen noisy speech dataset is used for training, validating and testing each approach with the seen types of noise and SNRs; (2) an unseen noisy speech dataset is used for testing the generalization capability of the approach under unseen noise conditions; (3) the enhanced speech dataset is used for testing the prediction capability on speech signals that are degraded by additive noise and then enhanced by a speech separation algorithm.

The seen noisy speech dataset is generated by combining $3,000$ clean speech utterances from the TIMIT database [31] with ten types of noise (babble, factory, fighter jets, vehicle, radio channel, destroyer engine, machine gun, pink, tank and white noise) from the NOISEX-92 database [32]. The random segments of noise and speech are combined using one of 12 SNRs, which range from -25 dB to 30 dB with 5 dB increments. We use a large range of SNRs to ensure balanced coverage of PESQ scores. It results in $30,000$ seen noisy speech utterances, where $25,000$ of them are used for training models, $2,000$ for model selection and hyperparameter optimization, and the other $3,000$ for testing.

The unseen noisy speech dataset is generated by combining 2000 different TIMIT utterances with one of five unseen noises (cafeteria, cockpit, live restaurant, operating room, speech-shaped noise) using one of the above 12 SNRs, which results in $10,000$ unseen noisy speech signals. The enhanced speech dataset contains $2,000$ separated speech signals, which are enhanced by four speech enhancement algorithms: nonnegative matrix factorization (NMF) [33], ideal binary mask (IBM) estimation [34], ideal ratio mask (IRM) estimation [35], and complex ideal ratio masking (cIRM) approach [36]. The enhancement systems are training from 500 clean speech utterances that are combined with the above noises at 5 SNRs (e.g. -6 to 6 dB with 3 dB increment). Then the time-domain enhanced speech signals are restored by each of the above algorithms.

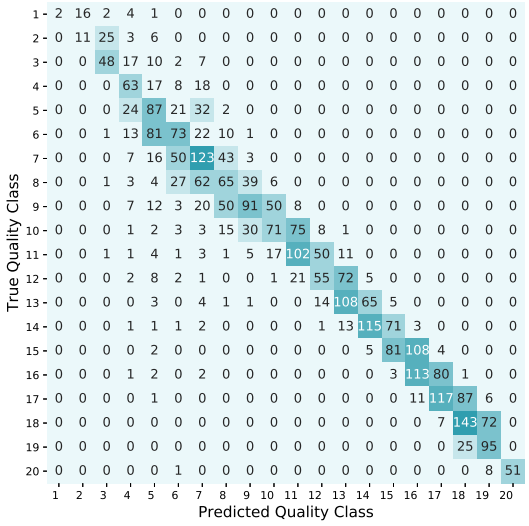

Figure 2: Confusion matrix of the categorical classification task. Class-1 indicates the lowest quality rank while Class-20 is the highest quality rank.

3.2. Experimental results and comparisons

Fig. 2 shows a confusion matrix that illustrates the classification-level accuracy of the proposed approach. Darker boxes indicate that more noisy speech signals are classified into this group by our approach. A series of dark boxes along the diagonal indicates ideal performance. As can be seen from the figure, there is an obvious

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

Table 1: Performance comparison on seen and unseen conditions. Best results of each case are marked in bold.

	Seen noisy speech			Unseen noisy speech			Enhanced speech		
	MSE	MAE	PCC	MSE	MAE	PCC	MSE	MAE	PCC
NISA [18]	0.156	0.309	0.86	0.183	0.346	0.84	0.151	0.232	0.88
DESQ [23]	0.170	0.339	**0.91**	0.246	0.385	**0.90**	0.168	0.248	0.91
CNN [25]	**0.139**	**0.269**	0.89	0.185	0.366	0.86	0.123	0.239	0.90
AutoMOS [30]	0.162	0.327	0.88	0.391	0.526	0.85	0.175	0.269	0.90
Quality-Net [26]	0.149	0.285	0.90	**0.170**	**0.325**	0.89	**0.102**	**0.217**	**0.93**
UCAN ($\beta = 0$)	0.097	0.197	0.94	0.112	0.246	0.92	0.087	0.196	0.94
UCAN ($\beta = 0.2$)	**0.078**	**0.177**	**0.95**	**0.096**	**0.193**	**0.93**	**0.062**	**0.148**	**0.96**

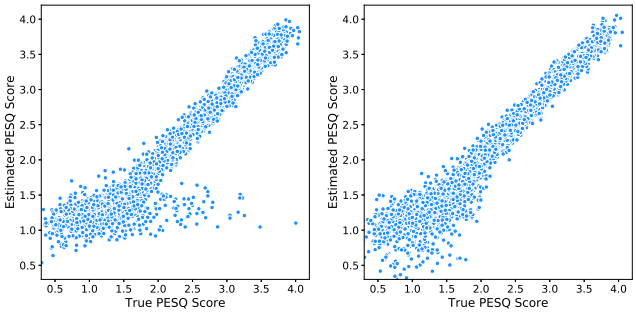

Figure 3: Scatter plots of the true and the estimated PESQ scores on seen noise condition. From left to right: UCAN without ($\beta = 0$) or with ($\beta = 0.2$) classification-aided module.

diagonal, which indicates that the categorical classification module gives rather good prediction performance (i.e. overall accuracy is 53.9%) for the 20-class case. Specially, UCAN can accurately predict in very low and very high noise conditions. Even when it predicts incorrectly, the predicted class label usually falls into the 1-nearest left or right neighbor of the true label with a high probability.

Fig. 3 shows how the classification portion of our UCAN model aids with estimating objective PESQ scores. Our proposed approach restrains most outliers, which is not possible when only a regression-loss function (e.g. MSE) is used. This is evidenced by setting β to 0. Many outliers are classified incorrectly when only the regression loss function is used (see left panel of Fig. 3). This, however, does not occur for our proposed approach (see right panel of Fig. 3). This point is inconspicuous when previous approaches measured performance.

We compare our system with five state-of-the-art nonintrusive methods. Non-intrusive speech assessment (NISA) [18] consists of a combination of short-term and long-term feature extraction followed by a regression tree. Deep machine listening for estimating speech quality (DESQ) [23] is a DNN-based model, which quantifies the degradation of phoneme representations obtained from the DNN as the speech quality prediction. A CNN architecture [25] consists of one convolutional layer and three dense layers and the summation of its outputs is used as an intelligibility estimate. AutoMOS [30] provides utterance-level estimates of MOS and is originally intended for assessing the naturalness of synthesized speech. We used their stacked long short-term memory (LSTM) model to predict the speech quality. Quality-Net [26] is a BLSTM model and its evaluation of utterance-wise quality is based on a frame-level assessment. Three measurements are used to assess how well our approach estimates the true PESQ score: mean absolute error (MAE), mean squared error (MSE), and Pearson correlation coefficient (PCC).

Table 1 shows the prediction performance of different approaches on the seen noisy speech dataset. In general, the proposed framework is significantly superior to other deep learning-based models. When the weight of classification loss $\beta = 0.2$, UCAN achieves the lowest MSE (0.078) and MAE (0.177) and the highest PCC (0.95). Notice that when $\beta = 0$ the proposed system is equivalent to a regression model without the classification constraint. In this situation, the MSE and MAE slightly increase to 0.097 and 0.197, but they are still noticeably lower than other approaches.

In order to evaluate the generalization capability of our model, we further test the proposed approach on two unseen conditions. The MSE and MAE of all approaches rise in general, but performance degradation in these unseen conditions is less than 0.02 for our proposed UCAN approach, which is the smallest performance degradation amongst all approaches. The errors with the enhanced speech case are generally lower than other scenarios as well. This likely occurs because the true PESQ scores of enhanced speech are generally higher, since they contain less noise, which makes for more accurate prediction. The best performance on the enhanced dataset is achieved by UCAN. Its MSE of 0.062, MAE of 0.148, and PCC of 0.96 far exceed other nonintrusive benchmarks. These results show that our proposed UCAN approach, which is trained with seen noise types, can still give the lowest prediction error when tested in unseen environments, indicating that it can generalize well.

4. CONCLUSION

We present an utterance-level classification-aided nonintrusive speech quality assessment approach to predict both the objective quality class and the quality score of noisy and enhanced speech signals. This framework enables real-world testing, since it does not require a reference clean signal. Overall, the performance of UCAN outperforms previous state-of-the-art approaches, and significantly lowers estimation errors, which indicates that jointly training a classification-aided regression module is promising for speech quality assessment.

5. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

[2] R. ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[3] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE TASLP*, vol. 19, pp. 2125–2136, 2011.

[4] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA)," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384, 2013.

[5] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, pp. 99–117, 2014.

[6] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, 2006.

[7] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE TASLP*, vol. 16, pp. 229–238, 2008.

[8] L. Malfait, J. Berger, and M. Kastner, "P.563-the ITU-T standard for single-ended speech quality assessment," *IEEE TASLP*, vol. 14, pp. 1924–1934, 2006.

[9] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE TASLP*, vol. 18, pp. 1766–1774, 2010.

[10] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *Proc. ICASSP*. IEEE, 2017, pp. 386–390.

[11] M. Delcroix, T. Yoshioka, A. Ogawa, *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *REVERB workshop*, 2014.

[12] T. H. Falk, V. Parsa, *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *Signal processing magazine*, vol. 32, pp. 114–124, 2015.

[13] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," 2019.

[14] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. ICASSP*. IEEE, 2019.

[15] H. Salehi, D. Suelzle, P. Folkeard, and V. Parsa, "Learning-based reference-free speech quality measures for hearing aid applications," *IEEE TASLP*, vol. 26, pp. 2277–2288, 2018.

[16] K. Arehart, J. Kates, M. Anderson, and L. Harvey, "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 122, pp. 1150–1164, 2007.

[17] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *Proc. ICASSP*. IEEE, 2016, pp. 624–628.

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

[19] M. H. Soni and H. A. Patil, "Novel subband autoencoder features for non-intrusive quality assessment of noise suppressed speech," in *Proc. INTERSPEECH*, 2016, pp. 3708–3712.

[20] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks." in *INTERSPEECH*, 2016, pp. 342–346.

[21] A. H. Andersen, E. Schoenmaker, and S. van de Par, "Speech intelligibility prediction as a classification problem," in *Proc. MLSP*. IEEE, 2016, pp. 1–6.

[22] X. Dong and D. S. Williamson, "Long-term SNR estimation using noise residuals and a two-stage deep-learning framework," in *Proc. LVA/ICA*. Springer, 2018, pp. 351–360.

[23] J. Ooster, R. Huber, and B. T. Meyer, "Prediction of perceived speech quality using deep machine listening," in *Proc. INTERSPEECH*, 2018.

[24] P. Seetharaman, G. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. ICASSP*, 2018, pp. 591–595.

[25] A. H. Andersen, J. M. Haan, Z. Tan, and J. Jensen, "Non-intrusive speech intelligibility prediction using convolutional neural networks," *IEEE TASLP*, vol. 26, pp. 1925–1939, 2018.

[26] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *Proc. INTERSPEECH*, 2018.

[27] J. Ooster and B. T. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. ICASSP*. IEEE, 2019, pp. 636–640.

[28] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[29] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[30] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *End-to-end Learning for Speech and Audio Processing Workshop NIPS*, 2016.

[31] J. S. Garofolo *et al.*, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993.

[32] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[33] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and spareness criteria," *IEEE TASLP*, vol. 15, pp. 1066–1074, 2007.

[34] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[35] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[36] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE TASLP*, vol. 24, pp. 483–492, 2016.