AN ATTENTION ENHANCED MULTI-TASK MODEL FOR OBJECTIVE SPEECH ASSESSMENT IN REAL-WORLD ENVIRONMENTS

Xuan Dong and Donald S. Williamson

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

ABSTRACT

Computational objective metrics that use reference signals have been shown to be effective forms of speech assessment in simulated environments, since they are correlated with subjective listening studies. Recent efforts have been dedicated towards effective forms of reference-less assessment to make real-world assessment more practical, but these approaches predict a limited number of assessment measures and they have not been evaluated in real-world conditions. In this work, we present a novel reference-less based framework called the attention enhanced multi-task speech assessment (AMSA) model, which provides reliable estimates of multiple objective quality and intelligibility measures in simulated and real-world environments. The multi-task learning (MTL) architecture effectively generates discriminative features that assist in improving our model's robustness. An attention mechanism is employed to identify key features within the feature space, and it noticeably reduces the estimation errors. A classification-aided module is also included to further suppress prediction outliers. Our model achieves the state-of-the-art performance in simulated and real-world data environments, where the results are strongly correlated with the corresponding reference-based objective scores.

Index Terms— speech quality and intelligibility, objective metrics, multi-task learning, attention networks, neural networks

1. INTRODUCTION

Speech quality and intelligibility are key factors when assessing a listening environment, communication channel, or speech enhancement algorithm. Subjective listening studies are the most accurate forms of assessing speech quality and intelligibility, but this form of assessment is general costly and time-consuming to perform when large-scale assessment is needed [1]. Thus, computational objective measures are often used, since they provide large-scale assessment in a short period of time.

Objective metrics can be divided into two categories. Intrusive (or reference-based) metrics assess the quality and intelligibility of a distorted speech signal by comparing it to its clean undistorted version. Hence, a clean reference signal is required. Commonly-used intrusive metrics include the perceptual evaluation of speech quality (PESQ) [2], short-time objective intelligibility (STOI) [3], signal-to-distortion ratio (SDR) [4], perceptual objective listening quality assessment (POLQA) [5], hearing aid speech quality index (HASQI) [6], and the speech transmission index (STI) [7]. A fundamental limitation of intrusive metrics is that the reference signal is usually not available in real-world environments or it may be difficult to obtain. Non-intrusive (or reference-less) metrics, on the other hand, assess speech based on the distorted signal only, which means that real-world assessment is possible. Example non-intrusive metrics include the ITU-T standard P.563 [8], ANIQUE [9], and the speech-to-reverberation modulation energy ratio (SRMR) [10], to name a few. Although reference-less based approaches enable real-world testing, these metrics have been shown to be less correlated to subjective ratings as compared to their reference-based counterparts [11, 12]. Hence, an active area of research involves developing non-intrusive metrics that can assess speech in real-life scenarios and that are strongly correlated with human assessment.

Many data-driven assessment approaches have been developed recently [13, 14, 15, 16, 17], where the goal is to predict subjective or objective scores. In [16], a full convolutional network is used to estimate STI. AutoMOS [14] is a long short-term memory (LSTM) model that assesses the naturalness of synthesized speech. A frame-level speech quality evaluation model named Quality-Net that consists of one bidirectional long short-term memory (BLSTM) layer and two fully connected (FC) layers is proposed in [18], where the authors predict PESQ. A similar approach is proposed in [19], where the authors estimate POLQA at the frame level with a convolutional neural network (CNN). Recently, Mel-frequency features and a deep neural network (DNN) are used to predict the subjective mean opinion score (MOS) of degraded acoustic signals [20]. Similarly, the authors in [21] utilize a CNN to predict subjective intelligibility. Although neural network-based non-intrusive speech assessment has achieved considerable success, several significant issues remain unsolved and require further development, including: 1) generalization performance in unseen and realistic conditions (e.g., noisy and/or reverberant); 2) assessment approaches provide singular assessments in terms of quality or intelligibility, but no approach provides both of these assessments. A unified model that leverages different aspects of speech assessment may be more robust.

In this paper, we propose an attention enhanced multi-task speech assessment (AMSA) model to estimate objective speech quality and intelligibility scores. Our model takes a speech signal as input and generates the corresponding estimates of PESQ, extended STOI (ESTOI) [22], HASQI, and SDR in a single model. These metrics reliably describe different attributes of speech. We use multiple CNN layers to extract discriminative features and reduce unwanted variations. A BLSTM layer is then used to further model the global temporal structure. An attention network [23] is

This research was supported by a NSF grant (IIS-1755844).

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

then used to adaptively measure the importance of different components within the feature space. From our prior work [24], we found that jointly predicting the class and true objective score can further reduce estimation outliers, so this structure is adopted here. Note that we use objective scores as the training label as they are readily available, unlike large-scale human assessment scores. We do this as a proof-of-concept of our proposed algorithm, where this will serve as preliminary work towards real-world human-level assessment.

2. MULTI-TASKS SPEECH ASSESSMENT MODEL

Multi-task learning (MTL) [25] has been beneficial to many speech applications [26, 27, 28, 29]. In [30], the authors develop a linear hierarchical Bayes (HB) predictor to fit subjective rating data, and the results show that MTL offers a natural way to account for the heterogeneity of quality ratings. By applying a MTL paradigm in our model, we aim to leverage the useful information contained in multiple related tasks to help improve the generalization performance.

Our proposed model utilizes the hard parameter sharing approach of MTL [31], by sharing the convolutional and BLSTM layers between all tasks. Objective-specific attention and fully connected layers (see Figure 1) are then used to predict objective scores. Convolutions allow the model to detect intrinsic low-level local patterns and generalize across frequency. The subsequent recurrent BLSTM network captures the temporal structure. Attention layers allow the model to focus on the key features to improve the performance for a specific objective. The main tasks use the learned latent representation from previous shared layers to output four estimates of speech quality or intelligibility. Adding classification as an auxiliary task allows the model to suppress unwanted prediction outliers.

2.1. Shared layers

The shared layers are constructed using four convolution blocks (ConvBlock) and a BLSTM layer. A ConvBlock consists of a 2-D convolutional layer with a kernel size of 3×3 and ReLU activations, batch normalization, and a 2×2 average pooling layer. Although max-pooling layers make the output of convolution networks transitionally invariant, they may also cause the network to lose information about the detailed T-F structures. Therefore, we adopt an average pooling layer here. The number of output filters for the four ConvBlocks are 16, 32, 64, and 128, respectively. The output features from the last ConvBlock are then flattened into a vector, and they are provided as input to the BLSTM layer which has 128 hidden units in each direction. We determined the above parameters empirically.

2.2. Attention block

For attention, we use the self-attention mechanism similar to [32], as depicted in Figure 1, since this allows the network to key in on task-specific information that may improve prediction performance. Note that every task-specific subnet (e.g., predicted objective metric) has its own attention block (AttnBlock). The attention block takes as input the hidden states of the BLSTM, $\mathbf{h} = \{h_1, h_2, \ldots, h_L\}$. In self attention, all of the keys (h_t^K), values (h_t^V) and queries (h_t^Q) at time step t come from the same output of the previous layer (i.e., BLSTM). Thus, $h_t^K = h_t^V = h_t^Q$. We pass a dot-product attention model [33], $f_{attn}(h_t^K, h_i^Q) = h_t^{K\top} h_i^Q$, to a softmax function to compute the attention weight:

$$\alpha_{i,t} = \frac{e^{f_{attn}(h_t^K, h_i^Q)}}{\sum_{t'} e^{f_{attn}(h_{t'}^K, h_i^Q)}}. \tag{1}$$

Once the weights are obtained, the context vector c_i is computed as: $c_i = \sum_{t=1}^{L} \alpha_{i,t} h_t^V$, which is a dynamic representation of the relevant part of the feature sequence at every output step i. It will be used by subsequent regression and classification tasks.

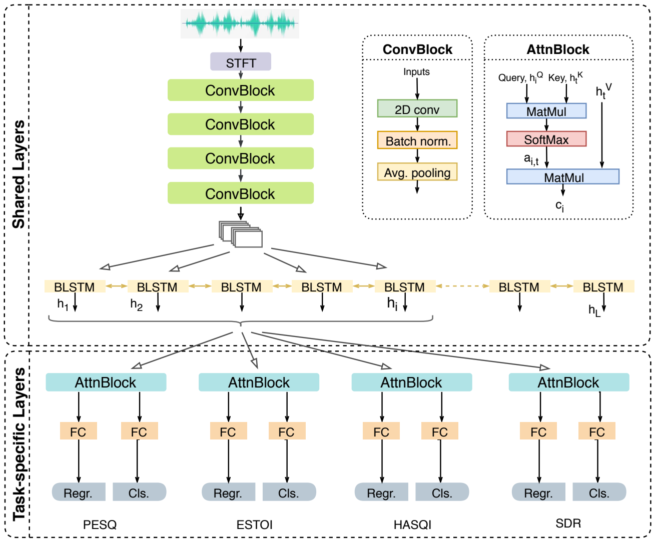

Fig. 1. The proposed attention enhanced MTL model.

2.3. Classification-aided module

The motivation for this module is that the regression task only generally minimizes the mean-square error (MSE), but this may result in prediction outliers [24]. This occurs when large estimation errors are "averaged out" by a large denominator (i.e., the number of test samples) when reporting MSE performance, which causes the model to overfit. Including the classification-aided module punishes samples with large estimation errors, and it leads to more robust performance.

For the k-th task, two training targets are simultaneously predicted. One is the raw objective score $score_{k,s}$ of the speech signal s, and the other is the corresponding categorical class $class_{k,s}$. Define $L_{k,thres}$ and $H_{k,thres}$ as the minimum and maximum values, respectively, of the k-th objective score. N_k is the number of classes. The classification label of the k-th objective score of a given signal s is calculated as

$$\text{class}_{k,s} = \min(\max\left(1, \text{ceil}\left(\frac{score_{k,s} - L_{k,thres}}{(H_{k,thres} - L_{k,thres})/N_k}\right)\right) N_k). \tag{2}$$

For the classification task, we use the cross-entropy loss as the objective:

$$\mathcal{L}_{k,cls} = -\sum_{x=1}^{N_k} \mathbf{1}(s,x) \log(\mathcal{P}(x|s)), \tag{3}$$

where $\mathbf{1}(\cdot)$ is the binary indicator that identifies if the predicted class label x matches the correct class label for signal s. $\mathcal{P}(x|s)$ is the predicted probability that s is of class x.

Table 1. Comparison results between several baselines and the proposed model on various test conditions of TIMIT data.

		PESQ			ESTOI			HASQI			SDR		
		MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC
Noisy TIMIT	MTL-Dense	0.15	0.16	0.92	0.05	0.04	0.94	0.04	0.03	0.91	1.26	1.14	0.93
	MTL-Attn	0.10	**0.09**	0.96	0.03	**0.02**	0.98	0.02	0.02	0.96	0.71	0.72	**0.99**
	Solo-AMSA	0.11	0.11	0.96	0.03	0.03	0.97	0.02	0.02	0.96	1.03	1.07	0.98
	AMSA	**0.09**	**0.09**	**0.98**	**0.03**	**0.02**	**0.99**	**0.01**	**0.01**	**0.98**	**0.67**	**0.67**	**0.99**
Reverberant TIMIT	MTL-Dense	0.18	0.20	0.86	0.05	0.06	0.90	0.07	0.04	0.81	1.52	1.74	0.84
	MTL-Attn	**0.12**	0.16	0.88	**0.03**	**0.04**	**0.93**	**0.03**	**0.03**	0.81	1.03	1.32	0.86
	Solo-AMSA	0.14	0.17	0.87	0.04	**0.04**	0.92	0.04	0.04	**0.83**	1.33	1.41	0.85
	AMSA	0.13	**0.15**	**0.90**	**0.03**	**0.04**	0.92	**0.03**	**0.03**	0.82	**0.78**	**0.85**	**0.90**
Reverberant noisy TIMIT	MTL-Dense	0.16	0.18	0.86	0.03	0.05	0.90	0.06	0.05	0.83	1.38	1.23	0.91
	MTL-Attn	**0.12**	**0.14**	**0.90**	0.02	0.03	0.96	**0.03**	0.03	0.86	0.52	0.64	0.92
	Solo-AMSA	0.13	0.15	0.88	0.04	0.03	0.97	0.04	0.04	**0.87**	0.81	0.79	0.92
	AMSA	**0.12**	**0.14**	0.89	**0.01**	**0.02**	**0.98**	**0.03**	**0.02**	0.86	**0.36**	**0.42**	**0.94**

2.4. Objective function

We train the entire network end-to-end with an unified loss function. The reason is that when training task-specific models simultaneously, the different but related models can interact at a high level, such that they regularize each other and gain statistical strength. The mean squared loss (regression loss denoted as $\mathcal{L}_{k,regr}$) together with the classification loss $\mathcal{L}_{k,cls}$ of the k-th task are utilized to update the weights of the shared network:

$$\mathcal{L}_{total} = \sum_{k=1}^{K} \beta_k (\mathcal{L}_{k,regr} + \lambda_k * \mathcal{L}_{k,cls}), \tag{4}$$

where β_k denotes the task weight of the k-th objective score prediction task, and K is the total number of objective tasks. λ_k denotes the loss weight of the k-th auxiliary task, which is a tunable parameter that balances the regression and classification terms.

3. EXPERIMENTS

3.1. Experiment setup

The TIMIT speech corpus is used to evaluate performance. Three evaluation datasets are created: 1) a noisy dataset where 1,000 utterances are corrupted by 12 noise types at one of 10 SNR levels (-15 db to 30 dB with 5 dB step); 2) a reverberant dataset where we create 120 artificial room impulse responses (RIR) using the image-source method [34]. These RIRs are generated in 3 room sizes using T60s from 0.05 to 0.4 with 0.05 increments. Each RIR is convolved with 10 clean utterances, and generates 12,000 reverberant speech signals. The third set consists of reverberant-noisy speech. For this set, 500 utterances are mixed with 12 noise types using one of 5 SNRs (0 dB to 12 dB with 3 dB step). We then convolve the resulting noisy signals with another 60 RIRs, resulting in 6,000 reverberant noisy speech signals. Finally, we split the 30,000 signals into training, validation, and testing sets of 20,000, 5,000, and 5,000, respectively.

The sampling rate of the speech signal is 16 kHz. We use the short-time Fourier transform (STFT) to extract the spectrogram from each utterance (512 point FFT). A Hanning window with 512 points and an overlap of 128 are used. Mean and variance normalization is applied to the input feature vector.

Our model takes a 6-second clip of speech as the input, and it outputs estimates of PESQ, ESTOI, HASQI, and SDR. PESQ covers a scale from 1 to 5 under P.862.1. The range of ESTOI and HASQI are from 0 to 1. The SDR range of our test data is from -22 to 35 dB. Since the range of each objective score is different, we want the prediction cost of each score to contribute equally within the model. Therefore, we set the task weight of each subnet to roughly be an inverse proportion of the square of the corresponding score range, that is, $\beta_1 = 1$, $\beta_2 = 12$, $\beta_3 = 12$, $\beta_4 = 0.1$ for PESQ, ESTOI, HASQI, and SDR, respectively. Also, $\lambda_k = 0.2$ and $N_k = 20$ for $k = \{1, 2, 3, 4\}$. These values are determined empirically.

We report three established measures for evaluating the performance of non-intrusive methods: epsilon insensitive root mean squared error (RMSE*) [35], mean absolute error (MAE), and Pearson correlation coefficient (PCC). RMSE* considers the confidence interval (CI) when calculating prediction errors, where the value of ϵ defines a margin of tolerance where no penalty is given to errors. Thus, RMSE* can assess statistical significance. We use 95% CI as recommended in [11, 19].

3.2. Experimental results

We first analyze the roles of multi-task training, attention, and classification in the proposed model, and set up three baselines: MTL-Dense replaces the entire task-specific layer of the proposed model with 2 dense layers (e.g., no attention or classification modules); MTL-Attn does not include the classification module (i.e., only do regression using attention output). Solo-AMSA adopts the same architecture of the shared layers, but it only predicts one score (e.g., PESQ only or ESTOI only), where both regression and classification modules are used. This is similar to our prior approach [24].

The results are shown in Table 1. In general, the performance of the different architectures is similar for the four objective scores. We can clearly see that the MTL-Attn significantly outperforms the MTL-Dense consistently on different testing sets across all metrics. For instance, the RMSE* of PESQ is reduced around 0.04 to 0.07 and that of SDR is halved. It indicates that introducing an attention mechanism to enhance the key features in multi-task model is beneficial for prediction accuracy. We also notice that the classification-

Table 2. Performance of several state-of-the-art methods and the proposed model on TIMIT data. The best results are **bold**.

	PESQ			ESTOI			HASQI			SDR		
	MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC
AutoMOS [14]	0.35	0.30	0.84	0.14	0.10	0.83	0.12	0.12	0.83	2.71	2.56	0.87
CNN [21]	0.29	0.27	0.86	0.07	0.06	0.93	0.08	0.06	0.90	2.13	1.97	0.91
DNN [20]	0.19	0.18	0.90	0.11	0.08	0.86	0.06	0.07	0.88	1.90	1.84	0.91
Quality-Net [18]	0.16	0.17	0.91	0.05	0.04	0.96	0.04	0.04	**0.91**	1.52	1.48	0.92
NISQA [19]	0.19	0.17	0.90	0.06	0.06	0.94	0.05	0.04	**0.91**	1.24	1.27	0.92
Solo-AMSA	0.14	0.13	0.92	0.03	**0.03**	0.95	0.03	0.03	0.90	1.03	1.08	0.93
AMSA	**0.11**	**0.10**	**0.94**	**0.02**	**0.03**	**0.97**	**0.02**	**0.02**	**0.91**	**0.62**	**0.65**	**0.95**

Table 3. Generalization results of three approaches on real-world corpora (COSINE and VOiCES).

	PESQ			ESTOI			HASQI			SDR		
	MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC	MAE	RMSE*	PCC
Quality-Net [18]	0.56	0.63	0.69	0.17	0.19	0.56	0.1	0.12	0.71	4.37	5.69	0.67
NISQA [19]	0.34	0.38	0.77	0.14	0.18	0.63	0.06	0.08	0.75	4.13	4.55	0.71
AMSA	**0.25**	**0.29**	**0.84**	**0.06**	**0.05**	**0.81**	**0.05**	**0.05**	**0.79**	**2.63**	**2.30**	**0.81**

aided MTL model (i.e., AMSA) has slightly lower prediction errors than MTL-Attn and has an obvious improvement in correlation (i.e., 0.02 gain on average). Noticeable performance improvements from Solo-AMSA to AMSA verifies the effectiveness of MTL. One reason to learn common feature representations instead of using the solo model, is that the representation from a single model may not have enough expressive power for mismatched testing conditions. With data from all tasks, a more powerful representation can be generated that leads to improved performance.

We compare our approach with several state-of-the-art methods: AutoMOS [14] consists of a stack of LSTMs, [21] is a CNN-based approach, Quality-Net [18] uses BLSTM and FC layers, [20] uses a DNN, and NISQA [19] uses a combination of CNN and LSTM. Our model is trained jointly on four score prediction tasks, and other comparison approaches will be trained separately on every single task with the same set of hyper-parameters. As can be seen from the results in Table 2, our method outperforms AutoMOS, CNN, and DNN-based approaches in every score prediction task with a good margin. Compared to Quality-Net and NISQA, AMSA still obtains noticeable performance gains. Specifically, MAE, RMSE* and PCC of PESQ are improved by 0.08, 0.07, and 0.04 compared to NISQA while 0.05, 0.07, and 0.03 to Quality-Net. Meanwhile, AMSA outperforms NISQA by reducing RMSE* of ESTOI (i.e., 0.03 absolute value) by 50%. For HASQI, the PCC of Quality-Net and NISQA are comparable with AMSA, but AMSA outperforms them in terms of MAE and RMSE*. In fact, more SDR improvement is observed, where AMSA gets 0.65 RMSE* that is much lower than the 1.27 of NISQA and 1.28 of Quality-Net. Note that Solo-AMSA obtains performance improvements as well compared to other approaches, which indicates the important roles of the attention mechanism and classification-aided module in accurate speech assessment.

To test the generalization ability of our model in real-world environments, we also consider two real-world datasets, namely COnversational Speech In Noisy Environments (COSINE) [36] and Voices Obscured in Complex Environmental Settings (VOiCES) [37] corpora. COSINE is a set of multi-party conversations recorded in real world environments with background noise and interfering speakers. The recordings from the close-talking microphone and the body microphones (e.g., shoulder or chest) are used as the clean reference and distorted speech respectively when calculating the ground-truth objective scores. VOiCES was recorded by playing clean audio in rooms of different sizes, each having distinct room acoustic profiles. Background noise was played concurrently. We assess 1,500 distorted signals from each of the above corpora, and compare AMSA with the best two comparison approaches of Table 2 (i.e., Quality-Net and NISQA). We summarize the results of these experiments in Table 3. Not surprisingly, the performance of AMSA is declined compared to the results on the simulated TIMIT dataset, but its drop is much smaller than that of Quality-Net and NISQA. When considering these challenging test conditions where the speech corpus and distortions are totally unseen, it is promising to learn that the MAE and RMSE* are still within 6% of the actual range of each metric. Moreover, the PCC is around 0.8 on average which indicates the estimated scores follow the trend of the true scores well.

4. CONCLUSION

In this paper, we propose an attention enhanced multi-task model for speech assessment, aiming to use a single model to predict a number of objective speech quality and intelligibility metrics simultaneously. In particular, applying multi-task learning improves feature learning due to the underlying commonality among the tasks. Different from existing non-intrusive approaches, we incorporate a self-attention layer to detect the patterns within each feature map that are relevant to the current prediction. This operation significantly reduces the estimation error and improves the generalization ability in real-world acoustic environments. We also conclude that jointly training a classification-aided regression module is promising for speech assessment.

5. REFERENCES

[1] R. Streijl, S. Winkler, and D. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, 2016.

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

[5] J. Beerends, C. Schmidmer, J. Berger, et al., "Perceptual objective listening quality assessment (POLQA)," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384, 2013.

[6] J. Kates and K. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, 2014.

[7] International Electrotechnical Commission, "Objective rating of speech intelligibility by speech transmission index," 2011.

[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] D. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE TSAP*, vol. 13, no. 5, 2005.

[10] T. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE TASLP*, vol. 18, no. 7, pp. 1766–1774, 2010.

[11] T. Falk, V. Parsa, J. Santos, et al., "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal proc. mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[12] A. Andersen, J. de Haan, Z. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. ICASSP*. IEEE, 2017.

[13] D. Sharma, Y. Wang, P. Naylor, et al., "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, 2016.

[14] B. Patton, Y. Agiomyrgiannakis, M. Terry, et al., "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *Workshop NIPS*, 2016.

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

[16] 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.

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

[18] S. Fu, Y. Tsao, H. Hwang, et al., "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on blstm," *Interspeech*, 2018.

[19] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. ICASSP*. IEEE, 2019, pp. 7125–7129.

[20] A. Avila, H. Gamper, C. Reddy, R. Cutler, et al., "Non-intrusive speech quality assessment using neural networks," in *Proc. ICASSP*. IEEE, 2019, pp. 631–635.

[21] A. Andersen, J. Haan, Z. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE TASLP*, vol. 26, pp. 1925–1939, 2018.

[22] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE TASLP*, vol. 24, no. 11, pp. 2009–2022, 2016.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[24] X. Dong and D. S Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *Proc. WASPAA*, 2019.

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

[26] M. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 6965–6969.

[27] D. Chen and B. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE TASLP*, vol. 23, no. 7, pp. 1172–1183, 2015.

[28] S. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement.," in *Interspeech*, 2016, pp. 3768–3772.

[29] R. Fakoor, X. He, I. Tashev, et al., "Constrained convolutional-recurrent networks to improve speech quality with low impact on recognition accuracy," in *Proc. ICASSP*. IEEE, 2018.

[30] I. Mossavat, P. Petkov, B. Kleijn, et al., "A hierarchical bayesian approach to modeling heterogeneity in speech quality assessment," *IEEE TASLP*, vol. 20, no. 1, pp. 136–146, 2011.

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

[32] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[33] M. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[34] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, 1979.

[35] ITU-T, "P.1401, methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2012.

[36] A. Stupakov, E. Hanusa, J. Bilmes, et al., "COSINE-a corpus of multi-party conversational speech in noisy environments," in *Proc. ICASSP*. IEEE, 2009, pp. 4153–4156.

[37] C. Richey, M. Barrios, Z. Armstrong, et al., "Voices obscured in complex environmental settings (VOICES) corpus," *arXiv preprint arXiv:1804.05053*, 2018.