ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414182

# AN END-TO-END NON-INTRUSIVE MODEL FOR SUBJECTIVE AND OBJECTIVE REAL-WORLD SPEECH ASSESSMENT USING A MULTI-TASK FRAMEWORK

*Zhuohuang Zhang*[1,2]*, Piyush Vyas*[1]*, Xuan Dong*[1]*, Donald S. Williamson*[1]

[1] Department of Computer Science, Indiana University, USA
[2] Department of Speech, Language and Hearing Sciences, Indiana University, USA
{zhuozhan, piyush, xuandong}@iu.edu, williads@indiana.edu

## ABSTRACT

Speech assessment is crucial for many applications, but current intrusive methods cannot be used in real environments. Data-driven approaches have been proposed, but they use simulated speech materials or only estimate objective scores. In this paper, we propose a novel multi-task non-intrusive approach that is capable of simultaneously estimating both subjective and objective scores of real-world speech, to help facilitate learning. This approach enhances our prior work, which estimated subjective mean-opinion scores, where our approach now operates directly on the time-domain signal in an end-to-end fashion. The proposed system is compared against several state-of-the-art systems. The experimental results show that our multi-task and end-to-end framework leads to higher correlation performance and lower prediction errors, according to multiple evaluation measures.

***Index Terms***— Speech assessment, non-intrusive metric, subjective evaluation, neural networks

## 1. INTRODUCTION

Speech assessment is important for evaluating and improving the performance of many applications, such as speech separation [1, 2], dereverberation [3, 4], and text-to-speech (TTS) translation [5]. Subjective ratings provide the most reliable and accurate form of assessment, however, conducting listening studies is both time-consuming and costly. Hence, objective metrics are used, since they are easy to compute and allow quick assessment of large-scale datasets.

Objective evaluation metrics can be divided into two categories, intrusive and non-intrusive. Commonly used metrics such as the perceptual evaluation of speech quality (PESQ) [6], perceptual objective listening quality assessment (POLQA) [7], extended short-time objective intelligibility (eSTOI) [8] and signal-to-distortion ratio (SDR) are all intrusive approaches as they require a clean reference during assessment. The clean reference, however, is not always accessible in real-world environments which limits the practicability of intrusive metrics. In contrast, non-intrusive metrics such as ITU-T P.563 [9], speech-to-reverberation modulation energy ratio (SRMR) [10], and ANIQUE [11] perform assessment using only the corrupted speech. Although these approaches enable real-world speech assessment, they do not always correlate well with subjective ratings [12, 13].

Data-driven non-intrusive measures have been recently proposed. Quality-Net [14] performs frame-level speech quality assessment by leveraging the temporal properties of a bidirectional LSTM (Bi-LSTM), using PESQ scores as training targets. NISQA [13] enables super wide-band speech assessment with a convolution and LSTM based network that estimates the POLQA score of a given stimulus. Although these metrics have demonstrated good correlations with objective scores, only estimating objective scores is a limitation as objective measures only serve as approximations of human assessment [15, 16], indicating that further subjective tests are still needed. Alternatively, some systems predict mean opinion scores (MOS) collected from human listeners [17, 18]. However, like the earlier studies, noisy speech materials are manually created which cannot fully capture all the complex details that exist in real-world environments. Thus it is not clear whether these approaches generalize well to unseen real-world data. Moreover, most approaches use hand-crafted features, including the magnitude spectrogram, but these may not be optimal representations for speech assessment, since many studies have shown that phase is also important for human-assessed quality [19, 20, 21] and these features do not allow the network to learn a more optimal representation.

We develop a novel non-intrusive assessment approach, where an encoder first uses convolution and pyramid Bi-LSTM (pBi-LSTM) layers to extract features locally and temporally at different resolutions, directly from the time-domain signal. Then an attention mechanism is applied in a decoder for multi-task learning. The proposed system assesses speech from multiple perspectives, including subjective and objective speech quality (i.e., human MOS and PESQ), objective intelligibility and signal distortions (i.e.,

This work is supported by an NSF grant (IIS-1755844) and is also supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

eSTOI and SDR). Our prior work [22] successfully estimates the subjective quality of real-world speech, but it does not assess other properties, such as signal distortion or intelligibility which limits its applicability. A multi-task approach is developed in [23], but it only estimates objective scores using simulated data. There are three main contributions in this work. Firstly, to the best of our knowledge, this is the first non-intrusive speech assessment system that estimates both subjective and objective ratings for real-world recorded speech. Secondly, an end-to-end model is developed by encoding the time-domain speech with a convolution layer rather than the conventional short-time Fourier transform (STFT) which may not be optimal. Lastly, although not shown here, this approach enables direct comparisons between real-world and laboratory experiments, which can help solve the generalization problem.

## 2. REAL-WORLD SPEECH DATA

### 2.1. Speech Data and Crowdsource Labelling

We use the VOiCES [24] and the COSINE [25] corpora as the source for the real-world speech materials. VOiCES was recorded in acoustically challenging and reverberant environments using twelve mics strategically placed around different rooms. Recordings from two of the mics are used as reverberant stimuli and the foreground speech is used as the reference signal. The approximated speech-to-reverberation ratios (SRRs) of these signals range from $-4.9$ to $4.3$ dB. COSINE, on the other hand, was recorded in a multi-party conversational setting both indoor and outdoor to represent various background noises with seven mics: a 4-mic array placed on the speaker's chest, a close-talking mic, a throat mic and a shoulder mic. Since the recording from the close-talking mic captures high quality speech, we use it as the clean reference. Recordings from the shoulder mic and the 4-mic array are used as noisy signals. The speech-to-noise ratios (SNRs) for COSINE are approximately between $-10.1$ to $11.4$ dB.

We crowdsourced our listening tests on Amazon Mechanical Turk by publishing 700 human intelligence tasks (HITs), each of which was completed by 5 workers. In total 3,500 workers participated (1,455 females and 2,045 males), aged from 18 to 65 years old. All participants are native English speakers and self-reported to have normal hearing. Each HIT contains 15 trials of evaluations that follow ITU-R BS.1534 [26]. Each trial has multiple stimuli from varying conditions including a hidden clean reference, an anchor (low-pass clean reference) and multiple real-world noisy or reverberant signals. Users provide quality ratings (between 0 to 100) for all stimuli. A total of 180K responses are collected for 36K signals (18K signals per dataset) with a total duration of approximately 45 hours. This study has been approved by our Institutional Review Board. This dataset will be open-sourced in the future and more information is provided in [22].

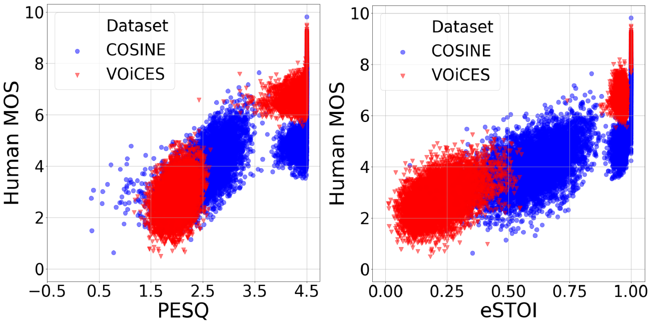

**Fig. 1**. Correlations between subjective and objective ratings on two real-world corpora (COSINE - blue, VOiCES - red).

### 2.2. Data Pre-processing

To remove potential outliers in the collected responses, we first calculated the Z-score for each stimulus across conditions and those with absolute Z-scores above 2.5 are filtered out [27]. Min-max normalization is then performed to normalize the range of ratings between 0 to 10. We further adopted two robust non-parametric techniques to remove outliers with large deviations [28, 29]. The MOS is then computed as the average scaled ratings for each stimuli. After these steps, the clean reference, anchor and noisy speech for COSINE and VOiCES have an average MOS of 7.38 (S.D. = 0.68) and 8.46 (S.D. = 0.26), 4.89 (S.D. = 0.34) and 6.74 (S.D. = 0.39), 4.04 (S.D. = 0.85) and 2.74 (S.D. = 0.65), respectively. The correlations between the MOS and two objective ratings are shown in Fig. 1. The upward trend between subjective and objectives scores indicate that there are some correlations between the two, and that jointly estimating them may be beneficial. The Pearson's correlation coefficients (PCC) between the scores for COSINE and VOiCES are: PESQ (PCC=0.69 and 0.96), and eSTOI (PCC=0.70 and 0.96).

## 3. DATA-DRIVEN MULTI-TASK MODEL

To estimate the subjective and objective scores, our model adopts a hard parameter sharing scheme of multi-task learning [30], where the shared-encoding layers are forced to learn generalized features from the input speech signal. It is reported that the way multi-task learning assesses the speech from different perspectives can account for the heterogeneity of quality ratings [31]. Incorporating this hard parameter sharing scheme of multi-task learning can further help to improve the generalization performance of our model.

### 3.1. Network Architecture

Our proposed model consists of two types of layers, the shared-encoder layers and the task-specific decoding layers that are based on attention mechanisms (Fig. 2). The time-domain speech is first passed through a 1D convolution layer to extract features, then the shared information for multi-task learning is captured using convolution and pBi-LSTM

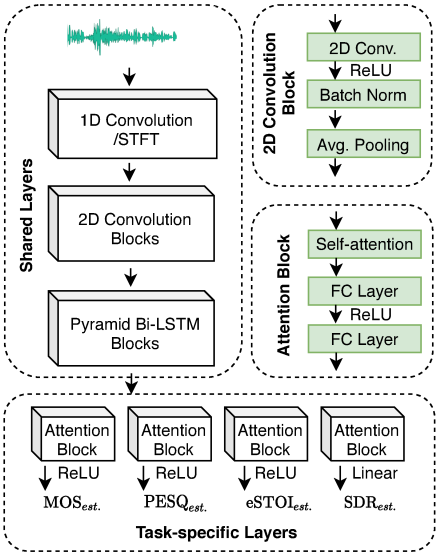

**Fig. 2**. Network architecture of our proposed non-intrusive metric. The pBi-LSTM block is illustrated in Fig. 3.

layers. The 2D convolution operation is good at capturing local dependencies of the encoded speech signal [32], and the pBi-LSTM layers can further explore the encoded feature at different temporal resolutions [33]. Then the decoder focuses on different aspects of the learned latent feature and makes task specific estimates of the different metrics.

Each convolution block consists of a 2D convolution layer with ReLU activations followed by batch normalization. Another average pooling layer is then applied to reduce the feature space of the encoded feature. Every pBi-LSTM block reduces the temporal size of the latent feature by half each time (see Fig. 3), where each output hidden state is merged with its one nearby hidden state. This is formulated as

$$h^{t}_{\text{pBi-LSTM}} = h^{2t-1}_{\text{Bi-LSTM}} + h^{2t}_{\text{Bi-LSTM}}, \tag{1}$$

where $h^{t}_{\text{pBi-LSTM}}$ represents the output hidden state of a pBi-LSTM layer at time $t$, and $h^{2t}_{\text{Bi-LSTM}}$ is the hidden state output of a conventional Bi-LSTM layer at time $2t$. The output hidden states $h$ of the pBi-LSTM blocks has a temporal dimension of size $T/M$, where $T$ denotes the original temporal dimension of the latent feature before feeding into the pBi-LSTM blocks and $M = 2^L$ is a temporal reduction number and $L$ is the number of pBi-LSTM blocks. Then the attention block helps the network focus on task-specific goals to improve performance. We use self-attention mechanism [34], where it takes the same input (i.e., the encoded hidden states $h$ of the pBi-LSTM block) to derive the query $Q_h$, key $K_h$ and value $V_h$ pairs. This process is formulated as

$$f_{\text{self-attention}}(Q_h, K_h, V_h) = softmax\{\frac{Q_h K_h^T}{\sqrt{d_h}}\}V_h, \tag{2}$$

where $d_h$ is the feature dimension of the encoded hidden states. $Q_h = hW^Q$ is the query and $W^Q$ is the corresponding projection weight. Similarly, $K_h$ and $V_h$ can be derived with their corresponding projection weights $W^K$ and $W^V$. We use ReLU as the output activation function for all the targets except for SDR, where a linear activation is used.

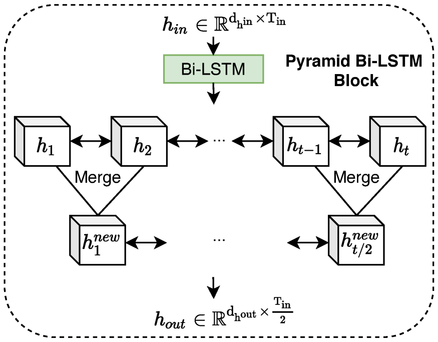
$h_{in} \in \mathbb{R}^{\mathrm{d_{h^{in}}} \times \mathrm{T_{in}}}$
Bi-LSTM
Pyramid Bi-LSTM Block
$h_1$
$h_2$
...
$h_{t-1}$
$h_t$
Merge
Merge
$h_1^{new}$
...
$h_{t/2}^{new}$
$h_{out} \in \mathbb{R}^{\mathrm{d_{h^{out}}} \times \frac{\mathrm{T_{in}}}{2}}$

**Fig. 3**. Illustration for one block of pBi-LSTM.

The model is trained in an end-to-end manner by minimizing the total mean squared error (MSE) of all training targets:

$$\mathcal{L}_{\text{Model}} = \sum_{k=1}^{K} \alpha_k \mathcal{L}^{k}_{\text{MSE}}, \tag{3}$$

where $K$ represents the total number of training targets, $\alpha_k$ denotes the corresponding weight for each individual MSE loss $\mathcal{L}^{k}_{\text{MSE}}$ calculated for different targets.

## 4. EXPERIMENTAL DESIGN

### 4.1. Experimental Setup

All signals are resampled at 16 kHz. Both the VOiCES and COSINE datasets are randomly split into training (80%), validation (10%) and testing (10%) subsets. Audios are truncated to 4 s and a 20 sample kernel with 10 sample stride are used for 1D convolution, where 257 output channels are produced. We use four 2D convolution blocks for our model, and a kernel size of 3×3 in each layer. The output channels are set to 16, 32, 64 and 128. We use 3 blocks of the pBi-LSTM (with 128, 64 and 32 units in each direction), which results in a temporal reduction factor of $M = 8$. The training targets are human MOS (ranging from 0-10), PESQ (-0.5 to 4.5), eSTOI (0 to 1) and SDR (-25 dB to 36 dB). We determine the weights for each loss term in Eq. (3) empirically, where $\alpha_1 = 10$, $\alpha_2 = 1$, $\alpha_3 = 12$, $\alpha_4 = 0.1$ for MOS, PESQ, eSTOI and SDR, respectively. The models are trained with 100 epochs using Adam optimizer and all models are trained and evaluated separately on COSINE and VOiCES datasets.

We include 5 non-intrusive data-driven models as comparison approaches, including a multi-task model for objective score estimation (AMSA) [23], a deep neural network (DNN) model [18], Quality-Net [14], NISQA [13] and pBi-LSTM+Att [22]. Note that all these data-driven models except AMSA are separately trained for each target since they

**Table 1**. Average performance between comparison and our proposed models. '-' indicates that the model is not capable of estimating such scores. The best results are in **bold**.

| | MOS | | | PESQ | | | eSTOI | | | SDR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Systems | MAE | PCC | SRCC | MAE | PCC | SRCC | MAE | PCC | SRCC | MAE | PCC | SRCC |
| AMSA [23] | - | - | - | 0.30 | 0.94 | 0.79 | 0.11 | 0.90 | 0.78 | 5.20 | 0.94 | 0.83 |
| DNN [18] | 0.49 | 0.94 | 0.88 | 0.19 | 0.96 | 0.83 | 0.05 | 0.96 | 0.86 | 3.50 | 0.98 | 0.88 |
| Quality-Net [14] | 0.48 | 0.93 | 0.87 | 0.15 | 0.97 | 0.81 | 0.06 | 0.95 | 0.80 | 2.72 | 0.96 | 0.88 |
| NISQA [13] | 0.50 | **0.96** | **0.90** | 0.18 | **0.98** | 0.88 | 0.06 | 0.96 | **0.88** | 2.20 | 0.98 | **0.93** |
| pBi-LSTM+Att [22] | 0.44 | 0.94 | 0.88 | 0.17 | 0.95 | 0.78 | 0.05 | 0.95 | 0.74 | 3.58 | 0.94 | 0.83 |
| **Prop. System (STFT)** | 0.42 | 0.95 | 0.88 | 0.17 | 0.95 | 0.80 | **0.04** | 0.94 | 0.85 | 2.69 | 0.97 | 0.89 |
| **Prop. System (1D-Conv)** | **0.40** | **0.96** | **0.90** | **0.12** | **0.98** | **0.89** | **0.04** | **0.97** | **0.88** | **1.87** | **0.99** | **0.93** |

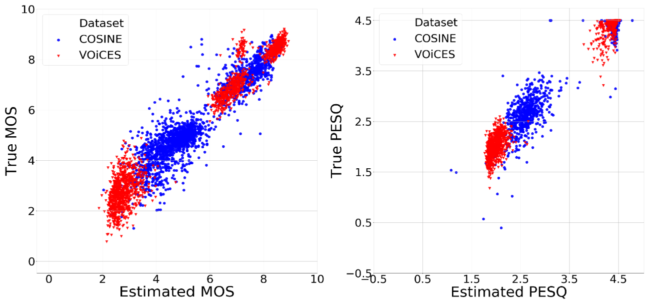

**Fig. 4**. Correlation between estimated scores and groundtruth on COSINE and VOiCES datasets (left: MOS, right: PESQ).

are not designed for multi-task estimation. We use the mean absolute error (MAE), PCC and Spearman's rank correlation coefficient (SRCC) to assess performance.

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

The average performance across both datasets is shown in Table 1. Our proposed approach (i.e. with 1D-Conv) achieves the best performance according to all evaluation metrics. Our system obtains a PCC of 0.96 vs. 0.93 in MOS, when compared to the Quality-Net. The SRCC is also higher (0.90 vs. 0.87). It suggests that our proposed system can better correlate with human ratings on real-world datasets. For PESQ estimation, the proposed approach also achieves the best performance (e.g., PCC 0.98 vs 0.95 in pBi-LSTM+Att). Although Quality-Net achieves similar performance on PCC in PESQ, its SRCC is lower than our approach (0.81 vs. 0.89). For objective speech intelligibility (i.e., eSTOI), our approach outperforms Quality-Net in terms of the monotonicity by 10% (i.e., SRCC, 0.88 vs. 0.80). Nonetheless, our model can better capture the signal distortion effects in the real-world speech signals, for SDR target, the PCC is 0.99 with an SRCC of 0.93. When compared to NISQA, although similar correlation performance is achieved for some targets, our model performs better according to MAE (MOS: 0.40 vs. 0.50, PESQ: 0.12 vs. 0.18, eSTOI: 0.04 vs. 0.06 and SDR: 1.87 vs. 2.20).

Note that the comparison approaches: DNN, Quality-Net, NISQA and pBi-LSTM+Att are all trained and evaluated for a single target each time, while our approach assesses the speech from different perspectives at the same time. Therefore, we further compare our approach with our prior work (i.e., AMSA [23]) that is capable of estimating multiple objective targets. Results still demonstrate the superiority of our system in all these objective targets, where joint subjective and objective assessment improves performance. For instance, on PESQ, our model achieves a PCC of 0.98 vs 0.94 in AMSA. For eSTOI, our model has a PCC of 0.97 compared to 0.90 in AMSA and PCC of 0.99 vs. 0.94 in SDR. The difference here suggest that our model is not only more capable in assessing multiple objective targets at the same time but also generalizes well on unseen human data.

Among our proposed models with different speech encoders, we find that the learnable 1D convolution layer leads to slight improvements for all targets in nearly all criteria (e.g., PCC on MOS: 0.96 vs. 0.95, PCC on PESQ: 0.98 vs. 0.95, PCC on eSTOI: 0.97 vs. 0.94 and PCC on SDR: 0.99 vs. 0.97). This suggests that the 1D-Conv can better encode the input speech than conventional STFT under mutli-task learning scenarios. We also illustrate the correlation results of our proposed model on the estimated MOS and PESQ scores in Fig. 4. It is easy to observe that our model has most of its estimations fall on the diagonals which reflects the high correlation results represented in Table 1.

## 6. CONCLUSIONS

We proposed a novel multi-task data-driven non-intrusive speech assessment model that is capable of analyzing the speech quality from subjective and different objective perspectives. The experimental results demonstrate that our proposed model achieves higher correlations, lower estimation errors when compared to the other state-of-the-art systems. The results also suggest a better encoding capability with a 1D convolution layer instead of conventional STFT. Our future work will move on to subjective intelligibility estimation.

## 7. REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer*

*Speech & Language*, vol. 24, pp. 1–15, 2010.

[2] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," *arXiv preprint arXiv:2008.06994*, 2020.

[3] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *TASLP*, vol. 25, pp. 1492–1501, 2017.

[4] Z.-Q. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *TASLP*, vol. 28, pp. 941–950, 2020.

[5] R. Liu et al., "Teacher-student training for robust tacotron-based tts," in *ICASSP*, 2020, pp. 6274–6278.

[6] A. W. Rix et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752.

[7] J. G. Beerends et al., "Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *J. Aud. Eng. Soc.*, vol. 61, pp. 366–384, 2013.

[8] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *TASLP*, vol. 24, pp. 2009–2022, 2016.

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

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

[11] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *TSAP*, vol. 13, pp. 821–831, 2005.

[12] A. H. Andersen et al., "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, 2017, pp. 5085–5089.

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

[14] S.-W. Fu et al., "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on blstm," *INTERSPEECH*, 2018.

[15] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *TASLP*, vol. 19, pp. 2046–2057, 2011.

[16] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *EUSIPCO*, 2016, pp. 1758–1762.

[17] B. Patton et al., "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *End-to-end Learning for Speech and Audio Processing Workshop NIPS*, 2016.

[18] A. R. Avila et al., "Non-intrusive speech quality assessment using neural networks," in *ICASSP*, 2019, pp. 631–635.

[19] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, pp. 465–494, 2011.

[20] A. Gaich and P. Mowlaee, "On speech quality estimation of phase-aware single-channel speech enhancement," in *ICASSP*, 2015, pp. 216–220.

[21] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," *INTERSPEECH*, 2020.

[22] X. Dong and D. S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," *INTERSPEECH*, 2020.

[23] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP*, 2020, pp. 911–915.

[24] Colleen Richey et al., "Voices obscured in complex environmental settings (VOICES) corpus," 2018.

[25] A. Stupakov et al., "COSINE-a corpus of multi-party conversational speech in noisy environments," in *ICASSP*, 2009, pp. 4153–4156.

[26] ITU-R, "BS. 1534 method for the subjective assessment of intermediate quality level of audio systems," *ITU Recommendation*, 2014.

[27] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[28] M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, 1996, vol. 96, pp. 226–231.

[29] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *ICDM*, 2008, pp. 413–422.

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

[31] I. Mossavat, P. N. Petkov, W. B. Kleijn, and O. Amft, "A hierarchical bayesian approach to modeling heterogeneity in speech quality assessment," *TASLP*, vol. 20, pp. 136–146, 2011.

[32] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, 2018, pp. 3229–3233.

[33] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[34] A. Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.