Interspeech 2022
18-22 September 2022, Incheon, Korea

Transformer Networks for Non-Intrusive Speech Quality Prediction

MK Jayesh[1], *Mukesh Sharma*[1], *Praneeth Vonteddu*[2], *M. A. B. Shaik*[1], *Sriram Ganapathy*[2]

[1]Samsung Research & Development Institute India, Bangalore
[2]LEAP lab, Electrical Engineering, Indian Institute of Science, Bangalore
{jayesh.mk,s.chandrakan,m.shaik}@samsung.com, {praneethl,sriramg}@iisc.ac.in

Abstract

This paper presents the details of our speech quality prediction system submitted to the Conferencing Speech-2022 challenge. The challenge involved the task of non-intrusive speech quality assessment intended for online conferencing applications. We propose two approaches for speech quality prediction in this work. The first approach uses a combination of deep convolutional neural network (CNN) and LSTM neural network with Kullback-Leibler (KL) loss function and cross entropy (CE) loss function for estimating the mean opinion scores (MOS). Our second approach uses transformer based encoder network before applying attention pooling. We observe that our proposed second method gives significant improvements compared to our first method as well as on the baselines provided by the challenge organizers with respect to Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC) along with reductions in root mean square error (RMSE). The model is also seen to generalize for unseen data resources on the evaluation dataset.

Index Terms: Transformer, mean opinion score (MOS), speech quality estimation.

1. Introduction

The automatic assessment of speech quality is crucial for various applications like the design of communication systems as well as in the development of speech enhancement and synthesis systems. The subjective evaluation of speech quality is considered the gold standard for quality assessment [1]. The mean opinion score (MOS), the estimated quality is the average of users' judgment, is usually given in a scale ranging from 1 to 5 [2]. While the assessment is reliable, the process of quality measurement is laborious, time-consuming, and expensive [3]. In addition, these quality measures do not allow the joint optimization with other neural approaches for speech enhancement and synthesis.

Several objective instrumental quality measures have been proposed in the past. The most commonly used method is the Perceptual Evaluation of Speech Quality (PESQ) metric [4]. Most of these metrics do not perform well for the wide range of speech distortions like telephone channel artifacts, noise, reverberation, multi-talker overlaps etc [5].

Early attempts have explored models based on spectral features [6], Gaussian mixture models [7] and hierarchical Bayesian methods [8]. In the recent past, propelled by the advancements in deep learning approaches, the non-intrusive quality estimation has been investigated using various models like auto encoders [9], deep neural networks [10] with multi-task learning [11], bidirectional long short-term memory networks (BLSTM) [12] and attention based neural networks [13]. Further, recent works have also explored semi-supervised methods for quality assessment [14]. In spite of all these efforts, the non-intrusive quality assessment continues to be a challenging problem partly due to the wide variety of acoustic conditions that impact speech like noise, reverberation, compression, packet-loss as well as the range of parameters in which speech quality assessment can be performed.

An open-call for developing non-intrusive speech quality assessment approaches has been made recently. This challenge, named Conferencing Speech 2022 challenge [15], aims to benchmark multiple sites on a common training and testing data platform. The datasets consist of a combination of data from four voice datasets along with MOS labels. These are the Tencent Corpus (Chinese language speech with noise and reverberation), NISQA corpus (English and German speech with codec and packet-loss distortions), IU-Bloomington Corpus (English multi-party conversational and read speech in noisy and reverberant environments), and PSTN Corpus (English audio books with additive noise and telephone channel artifacts). Thus, the dataset represents a wide-range of conditions like background noise, presence of speech enhancement system, reverberation, codecs, packet-loss and other possible online conference related voice impairment scenarios. Further, the audio data is also provided with the crowd-sourced MOS ratings at utterance level. A baseline system using convolutional neural networks and self-attention is also given by the organizers [16]. In this paper, we describe our efforts in developing a non-intrusive speech quality assessment model in response to the open call for participation.

The proposed approach to developing speech quality assessment (SQA) model extends the prior work by Mittag et. al. [16] in multiple ways. We explore the use of transformer based model architecture that consists of self-attention layers instead of convolutional or recurrent networks. We also investigate multiple loss functions that are based on mean square error (MSE), classification losses as well as KL based loss. We also develop the proposed SQA model targeting specific audio sampling rates of 8 kHz, 16 kHz and 48 kHz to allow domain and language specific modeling. Using the proposed approach for SQA, we improve the baseline system by a considerable margin (average relative improvements of 60 % in the Pearson correlation coefficient (PCC) metric and about 39% in the root mean square error (RMSE) metric). Further, we also report the results obtained on the unseen evaluation set.

2. Data and Experimental Setup

The dataset used in the challenge consists of data derived from multiple resources. A brief description of the databases used in the challenge is given below. Further, the training, development and evaluation splits are also discussed here.

This work was performed with project grant from the Samsung Research India, Bangalore.

10.21437/Interspeech.2022-10020

2.1. Data resources

2.1.1. Tencent corpus

The Tencent corpus consists of speech conditions in reverberant environments and in an-echoic environments. In the non-reverberant environments, there are 10k Chinese speech recordings with simulated distortions reflecting online meeting scenarios. In the reverberant environments, simulated distortions and live recording speech recordings are considered totalling about 4k audio recordings. The reverberation also included a mix of simulated (28%) and natural reverberation (72%) with reverberation time ranging from 0.4-0.7s. Each audio file was rated by more than 20 subjects and these scores are also present along with the raw audio data.

2.1.2. NISQA corpus

The NISQA Corpus includes 14k speech utterances with simulated noise (codec, packet-loss, and background noise) and conference (e.g., mobile phone, Zoom, Skype, WhatsApp) conditions. The subjective ratings measured the overall quality along the quality dimensions of noisiness, coloration, discontinuity, and loudness. Each audio file has, on average, 5 ratings from subjects collected in a crowd-sourced manner [16].

2.1.3. PSTN corpus

The PSTN corpus [17] is derived from the Librivox recordings of public domain audiobooks. For the challenge, 441 hours from 2150 speakers are chosen and segments of 10s are formed. Further, artificial noise is added to the clean files with a SNR between 0 - 40 dB. In total, 58, 709 speech recordings are made available for training, of which 40k files are noisy. Each training file is rated by 5 participants, while the test set files are rated by 30 participants.

The IU Bloomington data shared by the challenge team is not used for training as it used a different ITU recommendation for subjective testing from the other corpora mentioned above. IU Bloomington corpus adopted ITU-R BS.1534 for subjective testing while others used ITU-T P.808.

2.1.4. Data partitions

The training data consists of the entire NISQA corpus, 80% of Tencent Corpus and 95% of PSTN Corpus. The remaining portions of the Tencent and PSTN corpus are used for constructing the development data [15]. The development data is used for model hyperparameter choices as well as the choice of loss functions.

2.1.5. Evaluation data

The final evaluation set of the challenge consists of 5424 recordings. In these recordings, the Tencent Corpus constituted 2898 recordings, the TUB Corpus constituted 434 recordings and MS Corpus constituted 1040 recordings.

2.2. Feature Extraction

As proposed in [16], the input to the model are mel-spec features with 48 mel-warped sub-bands. The window length and hop size are chosen as 20 ms and 10 ms respectively. The mel-spec features with a context are divided into segments with a width of 15 and a height of 48. The hop size between the segments is taken as 3.

2.3. Output non-linearity or post-processing

2.3.1. ReLU Activation

As the models are attempting to predict the MOS that range between 1-5, we explored an output non-linearity that restricts the range to the possible MOS values. Specifically, we explored the transformation as shown below,

$$y \quad = \quad 1 + \text{ReLu}(x) - \text{ReLu}(x - 4) \tag{1}$$

Here, x denotes the linear input at the last layer of the neural model, y represents the final output from model. This output non-linearity allows the mapping of the scores to the range of [1-5]. As we can see from equation (1), $y = 1$ and $y = 5$ for $x < 0$ and $x > 4$ respectively. When $0 \leq x \leq 4$, y is linear between 0 and 5.

2.3.2. Mapping the outputs to MOS range

Another approach is to map the scores from the models to the MOS range of [1-5] using a linear or higher order polynomial mapping. The organizers use a 3rd order mapping for the evaluation.

2.4. Performance metrics

The performance metrics used to compare the predicted MOS and the ground truth values are the Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC) and the root mean square error (RMSE).

3. Model architecture

In this section, we detail the different baseline model architectures as well as the architectures used for the proposed model. All the models, both the baseline as well as the proposed, follow the same framework of audio processing with three major components: 1) Frame-wise modeling, 2) Time-dependency modeling with recurrent or self-attention/transformer layers and, 3) Final pooling layer. The frame-wise modeling is achieved by a set of front-end CNN layers. The model consists of 6 convolution layers and 3 pooling layers. The feature vector produced at the output dense layer of the frame-wise model has a dimension of 384.

3.1. Baseline systems

3.1.1. Baseline-1 (DNN-LSTM-AvgPool)

The Baseline-1 model [16] consists of a DNN with a depth of four layers and 2048 hidden units each. This is followed by a single BLSTM layer with 128 hidden units in each direction and an average pooling (AvgPool) is applied. The model is trained for 50 epochs with MSE loss function.

3.1.2. Baseline-2 (CNN-SA-AttPool)

The Baseline-2 model [16] consists of a CNN with 6 convolutional layers (with number of kernels in each layer chosen as $[16, 32, 64, 64, 64, 64]$ filters) and 3 max-pooling layers. The model contains dropout layers with a dropout rate of 0.2 after each max-pool layer. At the final convolutional layer, the output is flattened into a vector. This is followed by a self attention (SA) block. The attention mechanism adopted is the single head self-attention with the attention width set to 64. The self-attention layer is followed by a feed forward sub-layer. Finally, attention pooling is applied on dimensions of the output to get a

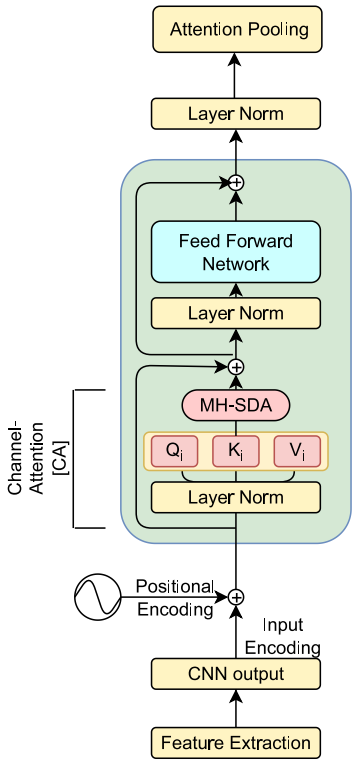

Figure 1: *Model-3 architecture comprising of CNN feature extractor, transformer encoder and attention pooling layer.*

single predicted MOS. The model is trained for 80 epochs with MSE loss function.

3.2. Proposed methods

The proposed models are detailed in section. We refer to the the proposed models to model-1 model-2 and model-3 in this paper. model-1 and model-2 are CLSTM based and derived from the Baseline-1. Model-3 is transformer based.

3.2.1. CLSTM Approach

Model-1 is derived by replacing the DNN output layers of Baseline-1 with CNN network. We also investigated the use of 2 self attention layers of dimension 64 at the output of CLSTM layers of Model-1. In Model-2, the output of the self-attention layers are processed with a feed forward layer having a single output neuron to generate the MOS. We used the ReLU activation function at the output of Model-1 and Model-2.

3.2.2. Transformer Approach (Model-3)

In this approach, we have retained frame-wise model and attention-pooling of the Baseline-2 system. The self attention network is replaced with a transformer encoder network having multi-head attention and positional encoding. We used the basic transformer encoder [18] with different choices of model parameters. This model is shown in Figure 1.

The model feeds the sum of the positional encoding and frame embedding from the CNN layer to the input layer of the encoder. The positional encoding incorporates the information regarding the relative position of the frames in the input. The model uses sinusoidal positional encoding.

The transformer encoder is composed of a stack of N identical layers. Each layer has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Self-attention mechanism of the encoder helps in modeling time-dependency at different time instances. The model also has a residual connection and layer normalization between the two sub-layers. We observed that, for the current work, increasing encoder layers to 4 (i.e, $N = 4$) leads to over-fitting of the model, and the performance degrades. Thus, we fixed the number of layers of the encoder as 3. Further, to avoid domain mis-match seen in the training data (NISQA, PSTN and Tencent corpora), we have used separate transformer based models for each of the different sampling rates of the audio data. For this model configurations, the scores are linear without any non-linearity in the output layer. The model is trained for 485 epochs.

3.3. Model combination

We also investigated combining multiple models. We used a simple score averaging of the model outputs to perform the final score generation.

4. Loss Functions

4.1. MSE loss

The basic loss function used in most of the models developed is the mean squared error (MSE) between the model prediction of the MOS and the ground truth MOS available at the recording level. In this cost metric, the mean score of the subjective ratings is only used for the reference.

4.2. KL divergence loss

Given all the subjective scores for each recording, we explore the possibility of assuming the score values as random variables that are Gaussian distributed. The sample mean and sample variance of the user ratings are chosen. In this scenario, we modify the model architecture to output the mean and (logarithm of) the standard deviation of the opinion scores. The loss function is the Kullbeck-Leibler (KL) divergence loss between the predicted distribution of the scores and the reference distribution. The KL loss between two Gaussian distributions p and q with means μ_1,μ_2 and standard deviations σ_1,σ_2 respectively can be calculated as

$$KL(p\|q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \qquad (2)$$

Here, μ_1 and σ_1^2 denote the reference sample mean and sample variance of user ratings, μ_2, σ_2^2 denote the model's prediction. Once the model is trained, we use the mean predictions in the development and evaluation datasets.

4.3. Cross entropy loss

We converted the scale of MOS in the range of [1-5] to a setting of 5 classes. In order to accommodate for the fractional value of MOS, we convert the ground-truth MOS to a discrete probability distribution. For example, a MOS of 3.25 can be written as $3 \times 0.75 + 4 \times 0.25$ which corresponds to a probability distribution of $[0, 0, 0.75, 0.25, 0]$ over the space of $[1, 2, 3, 4, 5]$ values. For using this loss function, the model's output layer is converted to a softmax output with 5 dimensions. We investigated a cross entropy loss between model's prediction and the ground truth to train the model.

Table 1: *Results for the various models considered in this work on the development data.*

Model	Architecture	Parameters	Loss	PCC	SRCC	RMSE
Baseline-1	DNN-LSTM-AvgPool	-	MSE	0.844	0.844	0.623
Baseline-2	CNN-SA-AttPool	-	MSE	0.914	0.908	0.471
Model-1-MSE	CNN-LSTM-AvgPool	1489717	MSE	0.880	0.875	0.496
Model-1-KL	CNN-LSTM-AvgPool	1490039	KL	0.885	0.875	0.485
Model-1-CE	CNN-LSTM-AvgPool	1490745	CE	0.875	0.870	0.502
Comb. (Model-1-KL, Model-1-CE)	-	-	-	0.889	0.879	0.476
Comb. (Model-1-KL, Baseline-2)	-	-	-	0.911	0.902	0.444
Comb. (Model-1-CE,Baseline-2)	-	-	-	0.909	0.900	0.440
Model-2-MSE	CNN-LSTM-SA-AvgPool	1536441	MSE	0.882	0.876	0.494
Model-3-MSE	CNN-Transformer-AttPool	922376	MSE	**0.962**	**0.961**	**0.287**

Table 2: *Official evaluation results released by Conferencing 2022 challenge organizers on the evaluation data in terms of PCC, RMSE and Metric3 (a 3rd order mapping of the model outputs to the range of [*1-5*]).*

Model	PCC over full eval set	RMSE over full eval set	Metric3 over full eval set	PCC over MS eval set	RMSE over MS eval set	Metric3 over MS eval set	PCC over TUB eval set	RMSE over TUB eval set	Metric3 over TUB eval set	PCC over Tencent eval set	RMSE over Tencent eval set	Metric3 over Tencent eval set
Baseline-1	0.53	0.768	0.497	0.361	0.585	0.293	0.348	1.094	0.649	0.881	0.624	0.55
Model-3-MSE	0.732	0.537	0.362	0.479	0.507	0.282	0.763	0.746	0.454	0.953	0.359	0.35

Table 3: *RMSE results on the subsets of the development data.*

Model	PSTN	Revb. Tencent	No-Revb Tencent
Baseline-1	0.626	0.648	0.480
Baseline-2	0.548	**0.269**	0.363
Model-1-KL	0.550	0.362	0.396
Model-1-CE	0.558	0.342	0.446
Model-3-MSE	**0.298**	0.317	**0.254**

5. Experiments And Results

In this section, we present the results for various model architectures and loss function choices described in Section 3 and 4. The performance metrics are the PCC, SRCC and RMSE described in Section 2.4. Table 1 contains the results averaged over the development dataset for all the models. The first two rows report the baseline system results. The rest of the table consolidates the performance of the proposed models, loss functions and model combinations. The table captures all the model improvisation efforts viz., replacing the time dependency model with transformer-encoder, modifications in the attention network of the baseline systems, introducing new cost functions and model combinations. Among the different cost functions explored on Model-1, Model-1-KL provides the best RMSE results. The model combination of Model-1 with Baseline-2 system also improves over the individual results of either models. Further, the self-attention network in Model-2 improves over the Model-1 results (using MSE loss). The best performance is Table 1 is seen for the Model-3 with the transformer encoder architecture and attention based pooling. We also tried fine-tuning the Model-3 MSE loss with KL loss, but the results did not improve. The performance improvements of Model-3-MSE are significant over the Baseline-2 system with relative improvements of 60% in terms of the PCC metric and 39% in terms of the RMSE metric.

The comparison of the proposed approach (Model-1 and Model-3) with the baseline model on the individual dataset splits in the validation data are reported in Table 3. As seen here, the Model-3-MSE improves over the baseline system on the PSTN and Tencent corpus (without reverberation).

The official evaluation data results shared by the Conferencing challenge 2022 organizers are shown in Table 2. The PCC, RMSE and Metric3 (RMSE after 3rd order mapping) are calculated on the evaluation dataset (Section 2.1.5). We submitted the MOS predicted for our best performing model (proposed Model-3-MSE). It is evident from the table that our model outperforms the baseline system by a significant margin on all the data sets.

The significant performance gains in Model-3-MSE are contributed by transformer-encoder approach which allows the modeling of the non-linear time-frequency dependencies of the input data with the MOS. Further, the improvements can also be attributed to the approach of training separate models for data with different sampling frequencies. The Model-3-MSE system is trained separately for 3 different sampling frequencies 8 kz, 16 kz and 48 kz. During testing, the input audio is re-sampled to the closest among the three choices (8, 16, and 48kHz). The Baseline-2 system used a single network where all the audio data was re-sampled to 48kHz. We hypothesize that this re-sampling contracts the spectrum of low sampling frequency signals and affects the model learning. This problem is mitigated by the proposed approach of using separate models.

6. Conclusions

In this work, we have described the system development approaches of the SRIB-LEAP team towards the Conference 2022 challenge of developing a model for non-intrusive speech quality assessment. The proposed approach explored the use of novel architectures (based on self-attention, LSTM and transformer models) as well as various loss functions. The proposed model based on transformer architecture gives significant improvements over the baseline system. Further, we have observed that the model is also able to perform well on unseen datasets that were tested as part of the evaluation dataset. In future, more detailed evaluation with combinations of model architectures and loss functions will be explored, as the timeline for participation in the challenge did not allow for a detailed experimentation on these fronts.

7. Acknowledgements

We thank the organizers of the ConferencingSpeech 2022 for providing the data, baseline system results, and evaluating the final system submission and the baseline system.

8. References

[1] S. X. C. SQEG, "Subjective performance assessment of telephone-band and wide band digital codecs," *Draft Recommendation*, p. 83, 1992.

[2] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[3] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech," in *IEEE international workshop on acoustic signal enhancement (IWAENC)*, 2016, pp. 1–5.

[4] K. P. A. Ksentini, C. Viho, and J. Bonnin, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *University of Rennes*, 2009.

[5] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[6] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, 2006.

[7] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using gaussian mixture models," *IEEE Signal Processing Letters*, vol. 13, no. 2, pp. 108–111, 2006.

[8] I. Mossavat, P. N. Petkov, W. B. Kleijn, and O. Amft, "A hierarchical bayesian approach to modeling heterogeneity in speech quality assessment," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 136–146, 2011.

[9] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 2315–2319.

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

[11] X. Dong and D. S. Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 100–104.

[12] 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," *arXiv preprint arXiv:1808.05344*, 2018.

[13] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 911–915.

[14] J. Serrà, J. Pons, and S. Pascual, "Sesqa: semi-supervised learning for speech quality assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 381–385.

[15] G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, W. Wardah, G. Mittag, R. Cutler, Z. Zhang, D. S. Williamson, F. Chen, F. Yang, and S. Shang, "Conferencingspeech 2022 challenge: Non-intrusive objective speech quality assessment (nisqa) challenge for online conferencing applications," 2022.

[16] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *Interspeech 2021*, Aug 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-299

[17] G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande, and R. Aichner, "DNN No-Reference PSTN Speech Quality Prediction," in *Proc. Interspeech 2020*, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.