MOS PREDICTOR FOR SYNTHETIC SPEECH WITH I-VECTOR INPUTS

Miao Liu[1] *Jing Wang*[1] *Shicong Li*[2] *Fei Xiang* [2] *Yue Yao*[3] *Lidong Yang*[3]

[1]Beijing Institute of Technology, Beijing, China [2]Xiaomi Inc., Beijing, China
[3]Inner Mongolia University of Science and Technology, Baotou, China

ABSTRACT

Based on deep learning technology, non-intrusive methods have received increasing attention for synthetic speech quality assessment since it does not need reference signals. Meanwhile, i-vector has been widely used in paralinguistic speech attribute recognition such as speaker and emotion recognition, but few studies have used it to estimate speech quality. In this paper, we propose a neural-network-based model that splices the deep features extracted by convolutional neural network (CNN) and i-vector on the time axis and uses Transformer encoder as time sequence model. To evaluate the proposed method, we improve the previous prediction models and conduct experiments on Voice Conversion Challenge (VCC) 2018 and 2016 dataset. Results show that i-vector contains information very related to the quality of synthetic speech and the proposed models that utilize i-vector and Transformer encoder highly increase the accuracy of MOSNet and MBNet on both utterance-level and system-level results.

Index Terms— speech quality assessment, speech synthesis, i-vector, Transformer encoder

1. INTRODUCTION

Due to the lack of the “right answer”, speech quality assessment has played an important role in speech synthesis areas such as text to speech (TTS) [1, 2] and voice conversion (VC) [3, 4]. Currently subjective test is the most widely used method that generally uses Mean Opinion Score (MOS) [5] to measure speech quality. It is accurate but laborious because it needs large number of listeners to give perceptual ratings.

Some objective measures have been developed in order to approach or even replace subjective evaluation, such as PESQ [6] and POLQA [7]. But they are not applicable in many scenarios because they need reference speech for comparison calculations. So far P.563 [8] is the only published standard in ITU-T to evaluate non-intrusive speech quality. It was applied to measure call quality and its accuracy is far from being optimal compared with full-reference methods.

With the development of deep learning, many researchers have applied this technology to MOS prediction, which greatly improved the accuracy of non-intrusive methods. AutoMOS [9], based on Long Short Term Memory (LSTM), predicted quality score of synthetic speech. Fu et al. proposed Quality-Net [10] based on bidirectional LSTM (BLSTM), which can capture frame-level quality. Mittag and Möller [11] proposed a TTS naturalness prediction model with transfer learning. Due to the great performance of MOSNet [12] on the assessment for VC task, many improvements have been proposed based on it. Choi et al. introduced multi-task learning [13] and cluster-based modeling methods [14] to improve the performance of MOSNet. Leng et al. proposed MBNet [15] with a mean subnet and a bias subnet to better utilize judge scores. Moreover, self-attention mechanism has been used to speech quality prediction in NISQA [16].

Recently, i-vector model has seen wider application in several speech processing areas such as speaker recognition [17], language recognition [18], and speech emotion recognition [19], which indicates that i-vector contains a wealth of non-semantic side information. Avila et al. have tried to use i-vector for enhanced and realistic speech quality assessment [20, 21]. But as shown in [21], the results from the simple deep neural network (DNN) model using i-vector as a feature set are worse than those from DNN using spectral features.

In order to verify whether quality information is included in i-vector of synthetic speech, we analyze the principle of i-vector and design models to predict MOS only using i-vector as input. In this paper, we propose a MOS prediction model with feature fusion and advanced networks to make further use of i-vector. Specifically, we insert i-vector into the beginning and end of extracted features output from CNN. The fusion features were fed to self-attention-based Transformer encoder [22] to achieve the prediction results. When Transformer encoder learns time-dependent information, i-vector is used to provide reliable quality information for each frame of extracted deep features. The experimental results show that our proposed method greatly improves MOS prediction performance of MOSNet and MBNet on VCC 2018 and VCC 2016 dataset.

2. I-VECTOR SPACE MODELING

I-vector was originally proposed to be used for speaker recognition [17]. I-vector space modeling approach maps the high

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

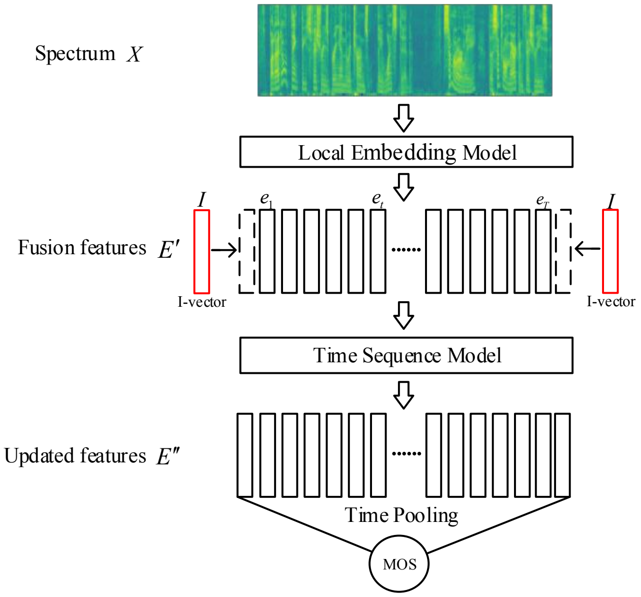

Fig. 1. An overview of speech quality model structure.

dimensional gaussian mixture model (GMM) supervector space to low dimensional total variability space which contains most variation between segments [23]. Given an utterance, a target GMM supervector M is written as follows:

$$M{=}m + Tw \tag{1}$$

where m is the universal background model (UBM) supervector, T is a low dimensional rectangular total variability matrix, and w is termed as i-vector.

The predecessor of i-vector is Joint Factor Analysis (JFA) [24], which uses V and U to represent the channel and speaker space respectively. Dehak et al. [17] simplified JFA and used the T matrix to model the speaker and channel spaces simultaneously, which means that i-vector w contains not only speaker information but also channel information.

In speech synthesis areas such as VC, a synthesis system can be regarded as a unique 'channel'. This 'channel' can not only convert the speaker in speech signals, but also bring some impairments that could be reflected in the channel space. Therefore, we think i-vector contains the speech impairment information and can be used to predict the synthetic quality scores.

3. METHODS

An overview of our proposed method is shown in Fig. 1. The network consists of three components: local embedding model, time sequence model, and time pooling.

First, we extract spectrum X from the speech signal, where x_t is a feature vector at time t, and T is the length of X sequence. Next, we obtain embedded feature E through the local embedding model such as CNN, where e_t is a deep feature vector at time t. Note that the length of X is same as that of E in our framework. Here, we introduce the i-vector I, which is extracted from the speech signal in advance. The i-vector is inserted into the beginning and end of E, which forms $E' = (I, E, I)$. We assume that i-vector is a kind of deep features, which is highly correlated with speech quality. It is reasonable to combine deep features extracted by CNN with i-vector and send them to the time sequence model.

After that, the time sequence model is applied to learn time information from the fusion features. We expect that the time sequence model can capture the quality information provided by i-vector at the beginning and end of the feature sequence to improve the prediction accuracy. In this part, we propose to use the Transformer encoder with position encoding instead of BLSTM commonly used before. Position encoding is necessary to inject the order information corresponding to the position of the sequence, especially for i-vector. Self attention is applied on Transformer encoder to capture the correlation between each feature vector, which helps the individual time steps of the fusion feature sequence to interact with each other.

E'' is obtained through the time sequence model. After adjustment through full connected layers, frame-level quality scores are obtained. Finally, global average pooling aggregates frame-level quality scores into utterance-level scores.

To evaluate the applicability of our proposed method, we use i-vector and Transformer encoder to reform the two structures of MOSNet and MBNet. Especially for MBNet, both MeanNet and BiasNet are modified by the proposed method. More details about reformation can be found in Section 4.

4. EXPERIMENT SETTINGS

4.1. Dataset

We evaluate the proposed method on the MOS dataset from VCC 2018 [25], which is a large-scale and open dataset. This dataset contains 20580 audios submitted by 38 different systems. A total of 267 judges rated for 20580 audios with MOS ranging from 1 to 5 and each utterance has an average of 4 judge scores. The utterance-level MOS (which is also called mean score) is obtained by averaging all the judge scores of the utterance. The system-level MOS is obtained by averaging all utterance-level MOS of the system. We randomly select 13,580, 3,000, and 4,000 audios into training, validation, and test sets, respectively. To prove the robustness of the models, we set the MOS dataset from VCC 2016 [26] as the other test set, which contains 26028 audios from 20 systems.

4.2. Model details

4.2.1. Models using i-vector only

In order to verify whether i-vector contains speech quality information, we design a support vector regression (SVR) model and a DNN model, which only use the i-vector as input for training. For SVR, the radial basis function (rbf) is

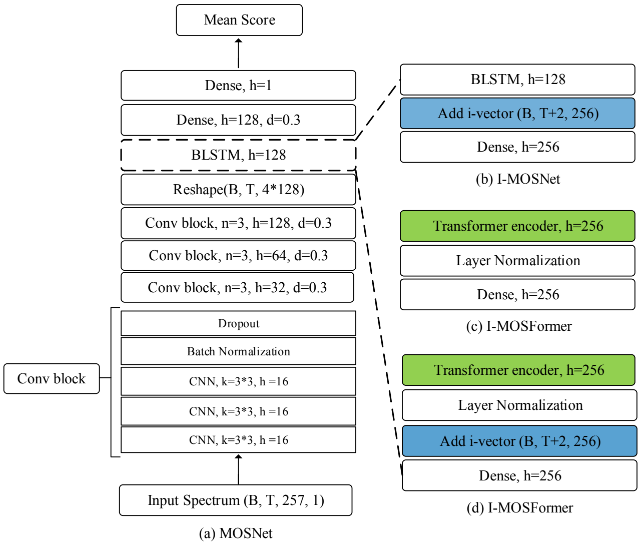

Fig. 2. Architecture of MOSNet and its three variants. The convolution block contains three convolutional layers, batch normalization (BN) [27] and dropout. In this figure, n denotes the number of CNN, k denotes kernel size of CNN, h denotes the output hidden size or channel size, B denotes batch size and T denotes the number of frames. In (b), i-vector passes through a BN layer before being inserted.

applied to learn non-linear relationship between i-vector and quality scores. For DNN, three hidden layers with batch normalization are used and each of them has 800 hidden nodes.

4.2.2. *MOSNet and its variants*

As Fig. 2(a) shows, we choose CNN-BLSTM-based MOSNet as baseline, which is a classic and effective MOS prediction model. The baseline MOSNet includes 4 convolution blocks, 1 BLSTM layer and 2 fully connected layers.

Fig. 2(b)(c)(d) shows the structures of three variants of MOSNet, which only change the dashed box part in Fig. 2(a). In i-MOSNet, a fully connected layer is used to resize extracted features in order to fit the dimension of i-vector. Then, i-vector is inserted into the beginning and end of extracted features. After that, a BLSTM layer and 2 dense layers are added. In MOSFormer, the Transformer encoder is implemented with 2 blocks. The block parameters are set as a single head, 256 attention units, and a feedforward network with 64 hidden units. As Fig. 2(d) shows, i-MOSFormer combine the modification of i-MOSNet and MOSFormer. More details of models can be found in Fig. 2.

4.2.3. *MBNet and its variants*

As far as we know, MBNet [15] is one of the best performing frameworks in synthetic speech assessment, which uses judge information for training. In our work, we reproduce MBNet and modify it similar to the description in Subsection 4.2.2. Fig. 3 shows the structures of MBNet and its three variants.

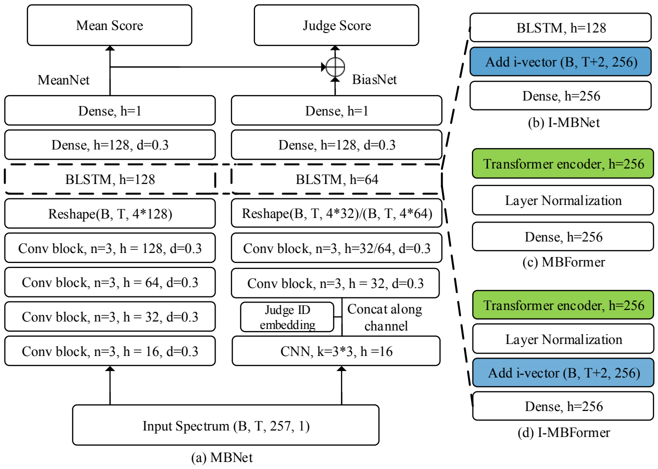
Mean Score
Judge Score
MeanNet
BiasNet
Dense, h=1
Dense, h=128, d=0.3
BLSTM, h=128
Reshape(B, T, 4*128)
Conv block, n=3, h = 128, d=0.3
Conv block, n=3, h = 64, d=0.3
Conv block, n=3, h = 32, d=0.3
Conv block, n=3, h = 16, d=0.3
Dense, h=1
Dense, h=128, d=0.3
BLSTM, h=64
Reshape(B, T, 4*32)/(B, T, 4*64)
Conv block, n=3, h=32/64, d=0.3
Conv block, n=3, h = 32, d=0.3
Judge ID embedding
Concat along channel
CNN, k=3*3, h =16
Input Spectrum (B, T, 257, 1)
(a) MBNet
BLSTM, h=128
Add i-vector (B, T+2, 256)
Dense, h=256
(b) I-MBNet
Transformer encoder, h=256
Layer Normalization
Dense, h=256
(c) MBFormer
Transformer encoder, h=256
Layer Normalization
Add i-vector (B, T+2, 256)
Dense, h=256
(d) I-MBFormer

Fig. 3. Architecture of MBNet and its variants. The meaning of symbols and some details are consistent with Fig. 2.

For the three variants, the output channel size of the last conv block in BiasNet is changed to 64. Moreover, the dashed box parts of both MeanNet and BiasNet are changed to the same structure shown in Fig. 3(b), (c), or (d).

4.3. Implementation setting

In our experiments, all speech samples are down-sampled to 16 kHz. We extract 257-dimension spectrum by short-time Fourier transform (STFT) from each sample for MOS prediction models. A referenced i-vector system is built based on Kaldi toolkit. The acoustic features we use in the system are 60-dimensional mel-frequency cepstral coefficients (MFCC) including first and second derivatives. A 2048-components full covariance GMM-UBM is trained, along with a 256-dimensional i-vector extractor. We used the VCC 2018 training set to train the UBM and i-vector extractor.

We implement the SVR model using scikit-learn toolkit and the neural-network(NN)-based model using PyTorch toolkit. For NN-based model, we use Adam [28] optimizer with 0.001 learning rate and the batch size is set to 64. We use the validation set to select the model with the lowest mean squared error (MSE) during 50 epochs. All experiments are repeated 5 times with different random seeds. We report the average MSE, linear correlation coefficient (LCC) [29], and Spearman's rank correlation coefficient (SRCC) [30] of each model as the final result. We report both the utterance-level result and the system-level result on VCC 2018 dataset. Due to the lack of utterance-level MOS in VCC 2016 dataset, only the system-level performance is reported in the paper.

5. RESULTS AND ANALYSES

5.1. MOSNet v.s. Models using i-vector only

First, we intend to compare the prediction performance of the models using i-vector only and MOSNet. Table 1 shows the

Table 1. Performance of MOSNet, SVR and DNN models on VCC 2018 and VCC 2016 dataset. Bold values indicate the best performance.

Model	VCC 2018						VCC 2016		
	utterance-level			system-level			system-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE
MOSNet [12]	**0.641**	**0.614**	**0.472**	**0.952**	**0.917**	**0.052**	**0.918**	0.872	0.484
SVR	0.582	0.562	0.599	0.794	0.805	0.231	0.868	**0.896**	**0.302**
DNN	0.560	0.536	0.860	0.755	0.732	0.491	0.869	0.863	0.757

Table 2. Performance of MOSNet and its variants on VCC 2018 and VCC 2016 dataset.

Model	VCC 2018						VCC 2016		
	utterance-level			system-level			system-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE
MOSNet [12]	0.641	0.614	0.472	0.952	0.917	0.052	0.918	0.872	0.484
i-MOSNet	**0.656**	**0.629**	**0.451**	0.952	0.905	0.046	0.922	**0.884**	0.575
MOSFormer	0.642	0.614	0.471	0.950	0.909	0.038	0.924	0.851	0.368
i-MOSFormer	0.650	0.623	0.474	**0.957**	**0.932**	**0.035**	**0.937**	0.871	**0.342**

Table 3. Performance of MBNet and its variants on VCC 2018 dataset.

Model	VCC 2018					
	utterance-level			system-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
MBNet [15]	0.667	0.639	0.509	0.964	0.926	0.088
i-MBNet	0.672	0.642	0.485	0.957	0.929	0.089
MBFormer	0.674	0.648	0.458	**0.968**	0.933	0.032
i-MBFormer	**0.680**	**0.656**	**0.449**	0.965	**0.938**	**0.029**

LCC, SRCC and MSE values of these models at the utterance and system levels. The performance of the model using 256-dimensional i-vector is comparable to MOSNet using spectrum with shape as $(257 \times T)$ on the utterance-level results of VCC 2018 dataset, even better on the utterance-level SRCC and MSE of VCC 2016 dataset, which fully proves that the i-vector is a representation with rich information about synthetic speech quality. In other fields such as speech recognition [31], i-vector, regarded as an important feature, is input into the model together with spectral features. Therefore, we consider that a reasonable fusion of spectral features and i-vector is a promising method in quality assessment field.

5.2. MOSNet v.s. Its variants

The results of MOSNet and its variants are given in Table 2. I-MOSNet has better results than MOSNet on the utterance-level metrics of VCC 2018 dataset, which indicates i-vector helps MOSNet to predict MOS more accurately. However, the mediocre performance on the system-level results is not satisfactory. Although MOSFormer has similar performance to MOSNet, i-MOSFormer makes up for the shortcomings of i-MOSNet and has a better performance on both datasets than MOSNet, which illustrates Transformer encoder is suitable for capturing the dependencies between i-vector and deep feature vectors extracted by CNN.

5.3. MBNet v.s. Its variants

Furthermore, we conduct experiments to apply our method to MBNet. Due to space limitation, only the results on VCC 2018 dataset are reported in Table 3. Compared to MBNet, i-MBNet improves the utterance-level results and basically maintained the system-level results, which is same as i-MOSNet. Similarly, i-MBFormer has the best results among four models. I-MBFormer improves utterance-level SRCC from 0.639 to 0.656 and reduces system-level MSE from 0.088 to 0.029, which can verify the generality of our proposed method to a certain extent.

In general, experiment results on VCC 2018 and 2016 dataset show the significant improvement of our methods over MOSNet and MBNet.

6. CONCLUSIONS

In this paper, we introduced i-vector to synthetic speech quality assessment and proposed a novel method that incorporates i-vector and Transformer encoder. The experimental evaluation on the VCC dataset shows that i-vector is a kind of features closely related to speech quality and our proposed method highly improves the performance of both MOSNet and MBNet. According to our analysis, the proposed i-vector-based method is not only suitable for the assessment of speech synthesis, but can also be applied to speech enhancement or real channel scenarios, etc. Furthermore, we can try to use large quality assessment irrelevant datasets such as VoxCeleb or Librispeech for the training of i-vector extractor. We will focus on this in future work.

7. ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation of China (No.62071039 and No.61620106002).

8. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," 2017.

[4] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," 2019.

[5] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," https://www.itu.int/rec/T-REC-P.800-199608-I.

[6] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," https://www.itu.int/rec/T-REC-P.862-200102-I/en.

[7] ITU-T Recommendation P.863, "Perceptual objective listening quality prediction," https://www.itu.int/rec/T-REC-P.863-201803-I/en.

[8] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," https://www.itu.int/itu-t/recommendations/rec.aspx?rec=P.563.

[9] Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin W. Wilson, Rif A. Saurous, and D. Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *CoRR*, vol. abs/1611.09207, 2016.

[10] Szu wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proc. Interspeech 2018*, 2018, pp. 1873–1877.

[11] Gabriel Mittag and Sebastian Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech 2020*, 2020, pp. 1748–1752.

[12] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech 2019*, 2019, pp. 1541–1545.

[13] Yeunju Choi, Youngmoon Jung, and Hoirin Kim, "Deep mos predictor for synthetic speech using cluster-based modeling," *Interspeech 2020*, Oct 2020.

[14] Yeunju Choi, Youngmoon Jung, and Hoirin Kim, "Neural mos prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," 2020.

[15] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 391–395.

[16] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech 2021*, 2021, pp. 2127–2131.

[17] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] Ming Li and Wenbo Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Proc. Interspeech 2014*, 2014, pp. 1120–1124.

[19] Rui Xia and Yang Liu, "Using i-vector space model for emotion recognition," in *Proc. Interspeech 2012*, 2012, pp. 2230–2233.

[20] Anderson R Avila, Jahangir Alam, Douglas O'Shaughnessy, and Tiago H Falk, "On the use of the i-vector speech representation for instrumental quality measurement," *Quality and User Experience*, vol. 5, no. 1, pp. 1–14, 2020.

[21] Anderson R. Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 631–635.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 6000–6010, Curran Associates Inc.

[23] Rui Xia and Yang Liu, "DBN-ivector Framework for Acoustic Emotion Recognition," in *Proc. Interspeech 2016*, 2016, pp. 480–484.

[24] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[25] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Speaker Odyssey 2018*. June 2018, pp. 195–202, ISCA.

[26] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech 2016*, 2016, pp. 1632–1636.

[27] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[28] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[29] Karl Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.

[30] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[31] Andrew Senior and Ignacio Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 225–229.