MOSNet: Deep Learning-based Objective Assessment for Voice Conversion

Chen-Chou Lo[1], Szu-Wei Fu[2], Wen-Chin Huang[1,3], Xin Wang[4], Junichi Yamagishi[4],
Yu Tsao[2], Hsin-Min Wang[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3]Graduate School of Informatics, Nagoya University, Japan
[4]National Institute of Informatics, Japan

lochenchou@gmail.com

Abstract

Existing objective evaluation metrics for voice conversion (VC) are not always correlated with human perception. Therefore, training VC models with such criteria may not effectively improve naturalness and similarity of converted speech. In this paper, we propose deep learning-based assessment models to predict human ratings of converted speech. We adopt the convolutional and recurrent neural network models to build a mean opinion score (MOS) predictor, termed as MOSNet. The proposed models are tested on large-scale listening test results of the Voice Conversion Challenge (VCC) 2018. Experimental results show that the predicted scores of the proposed MOSNet are highly correlated with human MOS ratings at the system level while being fairly correlated with human MOS ratings at the utterance level. Meanwhile, we have modified MOSNet to predict the similarity scores, and the preliminary results show that the predicted scores are also fairly correlated with human ratings. These results confirm that the proposed models could be used as a computational evaluator to measure the MOS of VC systems to reduce the need for expensive human rating.

Index Terms: speech naturalness assessment, speech quality assessment, MOS, voice conversion, non-intrusive,

1. Introduction

The quality quantification of generated speech has been a long-standing problem in speech synthesis, speech enhancement, and voice conversion (VC) systems. The evaluation of these systems reports both objective and subjective measurements. In the VC community, objective measures such as the Mel-cepstral distance (MCD) [1] are widely used for automatically measuring the quality of converted speech. However, such metrics are not always correlated with human perception as they measure mainly the distortion of acoustic features. Subjective measures such as the mean opinion score (MOS) and similarity score could represent the intrinsic naturalness and similarity of a VC system, but these type of evaluations are time-consuming and expensive as they need large number of participants to undergo listening tests and provide perceptual ratings.

Many assessment algorithms and models have been proposed to overcome the above-mentioned problem. For instance, in the speech enhancement field, the perceptual evaluation of speech quality (PESQ) [2] released by ITU-T is an intrusive assessment for measuring the quality of enhanced speech because a golden reference is needed for evaluation. There are several non-intrusive assessment metrics [3, 4, 5, 6, 7, 8, 9, 10] for evaluating the quality of enhanced speech and synthetic speech. For instance, Fu *et al.*[7] proposed Quality-Net as a quality as-

sessment model based on bidirectional long short-term memory (BLSTM) to predict the utterance-level quality of enhanced speech in a frame-wise manner. The high correlation between the predicted scores and the PESQ scores confirmed its effectiveness as a non-intrusive assessment model for speech enhancement. Yoshimura *et al.*[6] proposed a naturalness predictor for synthetic speech based on a fully connected neural network and a convolutional neural network (CNN) to predict the utterance-level and system-level MOS. The model was conducted on handcrafted features and trained with large-scale listening test ratings. It is worth noting that both [6] and [8] have reported a high variance in the utterance-level human ratings since the listening test is subjective and listeners may provide variant ratings to the same utterance. Therefore, it is difficult for an assessment model to get a high correlation with human utterance-level ratings. Nevertheless, the system-level prediction would be relatively reliable. Previous works have shown the capability of neural networks in modeling human perception for enhanced synthetic speech. However, there is no such assessment model for VC systems. Our goal is to develop a speech naturalness and similarity assessment model for VC systems using large-scale listening test results of the Voice Conversion Challenge (VCC) 2018 [11].

In this paper, we present a novel deep learning-based end-to-end speech naturalness assessment model, termed MOSNet. To develop such objective measures to model and align with human subjective ratings, we investigated the convolutional neural network (CNN), bidirectional long short-term memory (BLSTM), and CNN-BLSTM, as these architectures have shown their capability to model human perception. We used such architectures for extracting valuable features and then used fully connected (FC) layers to predict the corresponding naturalness scores. In virtue of the capacity of neural networks and the large-scale human naturalness evaluations of the VCC 2018, the MOS prediction of our naturalness assessment model achieved high correlation with human MOS ratings at the system level and fair correlation at the utterance level. In addition, we modified MOSNet to predict the similarity scores, and the preliminary results showed that the predicted similarity scores are fairly correlated with human similarity ratings. The experimental results showed that our proposed models have high capability to measure speech naturalness and similarity of VC systems. As per our knowledge, this is the first deep learning-based speech quality and similarity assessment model for VC.

The paper is organized as follows: The data and its distribution from VCC 2018 are described in Section 2. The proposed models are presented in Section 3. The experiments and results are discussed in Section 4. Finally, the conclusions and future

work are presented in Section 5.

2. The Voice Conversion Challenge Evaluation Data

2.1. The Voice Conversion Challenge 2018

The Voice Conversion Challenge (VCC) 2018, the second edition of VCC, is a large-scale voice conversion challenge. The VCC 2018 corpus was prepared by selecting a part of speakers from the device and production speech (DAPS) dataset [12] recorded by professional US English speakers in a clean and noise-free environment. Participants in the challenge needed to use their VC systems trained by parallel or non-parallel training data to convert speech signals from a source speaker to a target speaker. All the parallel and non-parallel VC systems participating in the challenge had been evaluated in terms of naturalness and similarity scores via crowd-sourcing listening tests.

The evaluation of the VCC 2018 was as follows: There were 2,572 evaluation sets, each consisting of 44 utterances. A total of 113,168 human evaluations fully covered 28,292 submitted audio samples. Each audio sample was rated by 4 listeners. 113,168 evaluations were split into 82,304 naturalness assessments and 30,864 speaker similarity assessments. 82,304 naturalness evaluations covered 20,580 submitted utterances with MOS ranging from 1 to 5, with lowest score of 1 and highest score of 5. The detailed specification of the corpus, listeners and evaluation methods can be found in [11]. We took the average score of the four MOS ratings for each utterance as its ground-truth score.

2.2. Data along with its distribution and predictability

The histograms of the mean and standard deviation of the four MOS ratings for each utterance are shown in Figure 1. It can be seen that the distribution of the mean MOS is closer to a Gaussian, and the mean MOS values are concentrated around 3.0. However, for about half of the submitted utterances, the standard deviation of the four MOS ratings is greater than 1, suggesting a higher degree of variation in the scores. This is expected because the perceptual ratings for the same utterance depend on the listeners' personal experiences and preferences when conducting listening tests. Therefore, we established that it is important to consider the intrinsic predictability of data and the inherent correlation among listeners. In this study, we used the bootstrap method in [6, 13, 14] to verify the inherent predictability of human evaluations in the VCC 2018.

We took 1,000 replications to estimate the MOS correlation between each subset and the whole dataset. Note that the MOS evaluations of natural speech was excluded. For each replication, we randomly sampled 134 listeners from a total of 267 listeners to measure their mean MOS as MOS_{sub}. Then, MOS_{sub} was compared to MOS_{all} (computed using the entire set of MOS ratings) in terms of the linear correlation coefficient (LCC) [15], Spearman's rank correlation coefficient (SRCC) [16] and mean square error (MSE). The average LCC, SRCC, and MSE values are shown in Table 1. Hence, it became clear that LCC and SRCC are quite high at the system level but lower at the utterance level. MSE shows a similar trend. The results indicate that although subjective perceptual ratings of different listeners vary at the utterance level, they have good consistency at the system level. This analysis shows that although the MOS at the system level is predictable, but, the MOS at the utterance
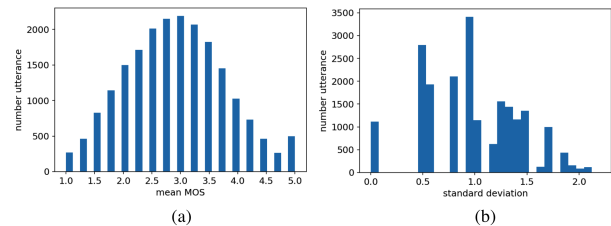

(a)   (b)

Figure 1: *Histograms of the mean (a) and standard deviation (b) of four MOS ratings for each utterance in the VCC 2018.*

Table 1: *Inherent correlation of human evaluations in the VCC 2018 at the utterance and system levels.*

Level	LCC	SRCC	MSE
Utterance	0.805	0.806	0.396
System	0.994	0.978	0.005

level can be predicted only up to a certain extent, although not as good as the prediction at the system level.

3. MOSNet

This paper proposes a deep learning-based objective assessment to model human perception in terms of MOS, referred to as MOSNet. Raw magnitude spectrogram is used as the input feature and three neural network-based models, namely CNN, BLSTM and CNN-BLSTM are used to extract valuable features from the input features of the fully connected (FC) layers and pooling mechanism to generate predicted MOS. In the following sections, we will detail each component of MOSNet.

3.1. Model details

The detailed configuration of different architectures of MOSNet is shown in Table 2, including CNN, BLSTM, and CNN-BLSTM. The BLSTM architecture is the same as the one used in Quality-Net [7]. In virtue of forward and backward processing through time, BLSTM has the ability to integrate long-term time dependencies and sequential characteristics into representative features. CNN has been widely used to model time series data and yielded satisfactory performance. CNN expands its receptive field through stacking more convolutional layers. The architecture of CNN used in this study has 12 convolutional layers, its receptive field for each neuron in the last convolutional layer is 25 frames (around 400 ms in the time-scale). We believe that, by considering the segment of 25 frames, the MOSNet can capture more temporal information to predict the quality scores. Recent studies have confirmed the effectiveness of combining CNN and RNN (BLSTM) for enhancement [17, 18], classification [19, 20], and recognition [21, 22] tasks. Thus, we also investigate the CNN + BLSTM architecture for feature extraction in the MOSNet, which is referred to as CNN-BLSTM in Table 2 and in subsequent discussions. With the extracted features, we use two FC layers to regress frame-wise features into a frame-level scalar to indicate the naturalness score of each frame. Finally, a global averaging operation is applied to the frame-level scores to obtain the utterance-level naturalness score.

3.2. Objective function

As described in the previous section, we formulate MOS prediction as a regression task. The input of the MOSNet is a sequence of spectral features, extracted from a speech utterance. The MOS evaluations of VCC 2018 are used as the ground-truth

Table 2: *Configuration of different model architectures. The convolutional layer parameters are denoted as conv{receptive field size}-{number of channels}/{stride}. The ReLU [23] activation function after each convolutional layer is not shown for brevity. N is for the number of frames.*

model	BLSTM	CNN	CNN-BLSTM
input layer	input (*N X 257 mag spectrogram*)		
conv. layer		$\left\{\begin{matrix} conv3 - (channels)/1 \\ conv3 - (channels)/1 \\ conv3 - (channels)/3 \end{matrix}\right\} X4$ $channels = [16, 32, 64, 128]$	
recurrent layer	BLSTM-128		BLSTM-128
FC layer	FC-64, ReLU, dropout	FC-64, ReLU, dropout	FC-128, ReLU, dropout
	FC-1 (*frame-wise scores*)		
output layer	average pool (*utterance score*)		

Table 3: *Utterance-level and system-level prediction results for different models, where the subscript denotes the batch size.*

	utterance-level			system-level		
Model$_{batchsize}$	LCC	SRCC	MSE	LCC	SRCC	MSE
BLSTM$_1$ [7]	0.511	0.484	0.604	0.826	0.808	0.165
BLSTM$_{16}$	0.487	0.453	0.658	0.818	0.797	0.190
BLSTM$_{64}$	0.251	0.254	0.803	0.412	0.427	0.404
CNN$_1$	0.638	0.587	**0.486**	0.945	0.875	0.058
CNN$_{16}$	0.620	0.573	0.512	0.944	0.890	0.067
CNN$_{64}$	0.624	0.585	0.522	0.946	0.872	0.057
CNN-BLSTM$_1$	0.584	0.551	0.634	0.951	0.873	0.135
CNN-BLSTM$_{16}$	0.607	0.569	0.540	0.944	**0.897**	**0.055**
CNN-BLSTM$_{64}$	**0.642**	**0.589**	0.538	**0.957**	0.888	0.084

to train the model. Fu *et al.*[7] stated that by using frame-level prediction errors in the objective function, utterance-level predictions will be more correlated with human ratings. Thus, we formulate the objective function for training the MOSNet as:

$$O = \frac{1}{S}\sum_{s=1}^{S}[(\hat{Q}_s - Q_s)^2 + \frac{\alpha}{T_s}\sum_{t=1}^{T_s}(\hat{Q}_s - q_{s,t})^2] \quad (1)$$

where $\hat{Q}_s$ and Q_s denote the ground-truth MOS and predicted MOS for the s-th utterance, respectively, α is the weighting factor, $q_{s,t}$ denotes the frame-level prediction at time t, T_s is the total number of frames in the s-th utterance, and S denotes the number of training utterances. Notably, the objective function in Eq. (1) combines the utterance-level MSE and frame-level MSE. In [7], the weighting factor (α) was used to mitigate severe MSE variations across frames in a speech enhancement task. Our preliminary experiments show that the quality of converted speech is much more stable across frames compared to the quality of enhanced speech. Specifically, the converted speech generated from a high MOS VC system typically produces high frame-wise MOSs, and vice versa. Thus, we set the weighting factor α to 1 in this study. To compute the frame-level MSE, the ground-truth MOS is used for all the frames in the speech utterance. From the experiments, it can be established that the frame-level MSE helps MOSNet converge with better prediction accuracy, which will be discussed in the next section.
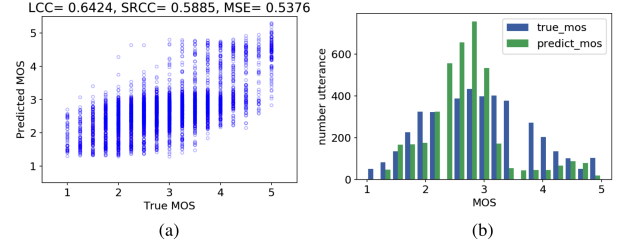

Figure 2: *Scatter plot (a) and histogram (b) of the utterance-level predictions of CNN-BLSTM$_{64}$.*

4. Experiments

In this section, different model architectures will be evaluated and the results at the utterance and system levels will be discussed. We explain our implementation details first. The entire set of 20,580 speech samples (along with the corresponding MOSs) were divided into 13,580, 3,000 and 4,000 samples for training, validation and testing, respectively. All speech samples were down-sized to 16 kHz. For feature extraction, we conducted short-time Fourier transform (STFT) of 512 sample points (i.e., 32 ms frame size) every 256 sample points (i.e., 16 ms frame shift), which resulted in a sequence of frame-based spectral feature of 257 dimensions for each utterance. For the MOSNet, the dropout [24] rate was set to 0.3. The model was trained by the Adam [25] optimizer with a learning rate of 0.0001, without batch normalization. We applied early stopping based on the MSE of the validation set with 5 epochs patience.

4.1. Comparison of different model architectures

First, we intend to compare the prediction performance of different models. Table 3 shows the LCC, SRCC and MSE values of these models with different batch sizes at the utterance and system levels. The model BLSTM$_1$ used in [7] is considered to be the baseline system. From Table 3, BLSTM$_1$ yielded 0.511 in terms of LCC, and its performance dropped drastically when the batch size increased, as zero-padding might affect the prediction performance. Next, the CNN-based models yielded more stable results when the batch size increased and boosted the LCC from 0.511 to 0.638. Among all the models, CNN-BLSTM$_{64}$ achieved the best LCC of 0.642, indicating that the combination of CNN (for feature extraction) and BLSTM (considering time dependencies) could effectively extract features from the spectrogram to perform MOS prediction. CNN-BLSTM$_{64}$ also achieved the highest LCC of 0.957 at the system level, which is very close to the LCC of 0.994 for human evaluation shown in Table 1.

4.1.1. Evaluated results at the utterance level

Figure 2 shows a scatter plot and histogram of the utterance-level predictions of CNN-BLSTM$_{64}$. As seen from Figure 2 (a), the model avoided low predictions although there were quite a few training samples in the low MOS, and the predictions were mainly distributed around scores of 2 to 3. Figure 2 (b) shows that MOSNet rarely predicted MOS between 3.5 to 4 and mostly predicted MOS between 1.5 and 3.5. As the average MOS of the submitted systems fell mainly between 2 to 3.5 at the system level, it is reasonable for the model to avoid prediction of such low and high scores. This is a common limitation when using MSE-based objective functions for the data is in a Gaussian distribution. It can be improved by using other objective functions, which can be a scope of future research.
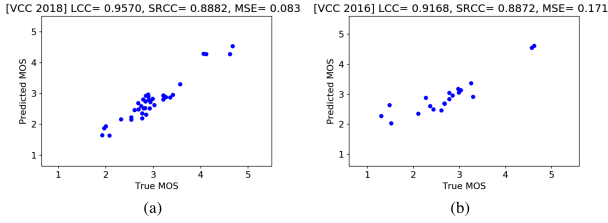
Figure 3: *Scatter plot of system-level predictions on the testing set in (a) the VCC 2018 and (b) the VCC 2016.*
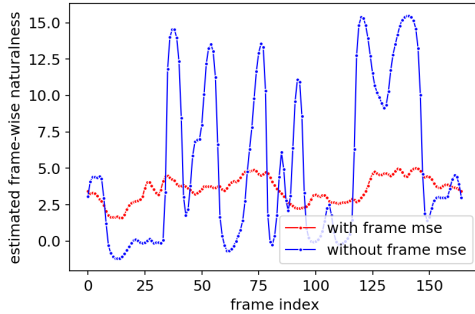

Figure 4: *Frame-wise MOS predictions of CNN-BLSTM$_{64}$.*

4.1.2. Evaluated results at the system level

Figure 3 (a) shows the scatter plot of system-level predictions by CNN-BLSTM$_{64}$ for the VCC 2018. System-level predictions were highly correlated with human evaluations. The results shown in Figure 3 (a) and Table 3 confirmed the effectiveness of the proposed MOSNet as a substitute for human listeners to evaluate VC systems.

To test the generalization of MOSNet, we applied CNN-BLSTM$_{64}$ trained by the VCC 2018 data to the VCC 2016 data. The scatter plot of system-level predictions is shown in Figure 3 (b). Although, the audio samples in both VCC 2016 and VCC 2018 came from the DAPS dataset, the selected speakers for testing of VCC 2016 were different from those of VCC 2018. The participating VC systems for the two challenges were also different. As the pair correspondence results for audio samples were not accessible, we could only report the system-level prediction results. Figure 3 (b) shows that MOSNet achieved a sufficiently high correlation of 0.917 for system-level predictions, thereby confirming its generalization ability in a training-testing mismatched scenario.

4.2. The effect of the frame-level MSE

Here, we intend to investigate the effect of the frame-level MSE used in the objective function in Eq. (1). Figure 4 and Table 4 shows the effect of training with or without the frame-level MSE on the CNN-BLSTM$_{64}$ model (the best system in Table 3). As seen in Figure 4, the estimated frame-wise scores vary greatly when the model was trained without the frame-level MSE, but are relatively stable when the model was trained with the frame-level MSE. Unstable frame-wise predictions affect the final utterance-level predictions. Table 4 shows that with frame-level MSE, the LCC of utterance-level predictions was 0.642, and it dropped significantly to 0.560 when the model was trained without the frame-level MSE.

Table 4: *The effect of frame-level MSE on the utterance-level predictions of CNN-BLSTM$_{64}$.*

	LCC	SRCC	MSE
with frame MSE	**0.642**	**0.589**	**0.538**
without frame MSE	0.560	0.528	2.525

Table 5: *Similarity prediction results of the modified MOSNet model.*

model	Level	ACC	LCC	SRCC	MSE
CNN	utterance	0.696	**0.453**	**0.455**	0.197
(scalar)	system	**0.701**	0.394	0.395	**0.195**
CNN	utterance	0.670	0.329	0.329	0.336
(2 classes)	system	0.674	0.292	0.292	0.326

4.3. Experiments on similarity prediction

The similarity in the measurement of the converted speech and the target speech is another key indicator of the success of VC systems. In addition to predicting the MOS, we also tried to extend the MOSNet to predict the similarity scores. As there were two input audio samples at the same time, we applied a shared CNN (with the same architecture as described in Table 2) to extract features into latent vectors for both the input samples. The paired vectors were combined to form a concatenated feature vector and inferred by two FC layers and a SoftMax layer to generate the final predicted score. Two models with different types of output were created: the first model had a scalar output, and the output of the second model was a two-class vector.

In VCC 2018, there were 30,864 evaluated similarity scores. Each similarity score was obtained by evaluating a pair of speech samples with 4 levels. We merged scores 1 and 2 as label 1 (same speaker), and scores 3 and 4 as label 0 (different speaker). The dataset was divided into 80% and 20% for training and testing, respectively. The similarity prediction results of the two models are shown in Table 5. From the table, we found that when the model is trained with one scalar output, it can achieve an accuracy of 69.6%, which is notably higher than 66.9% yielded by the two-class model. The results in Table 5 confirmed that the modified MOSNet can be used to predict similarity scores with fair correlation values to human ratings.

5. Conclusions

This paper presented a deep learning-based quality assessment model for the VC task, referred to as MOSNet. Based on large-scale human perceptual MOS evaluation results from VCC 2018, our experimental results show that MOSNet yields predictions with a high correlation to human ratings at the system level and a fair correlation at the utterance level. We have shown decent generalization capability of MOSNet by applying the model trained with the VCC 2018 data to the VCC 2016 data. Moreover, with a slight modification, MOSNet can fairly predict the similarity scores of the converted speech relative to the target speech. As per our knowledge, the proposed MOSNet is the first end-to-end speech objective assessment model for VC. In future, we will consider the human perception theory and improve the model architecture and objective function of MOSNet to attain improved correlation with human ratings.

6. References

[1] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[2] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[3] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. Forum Acusticum*, 2005.

[4] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. APSIPA ASC*, 2011.

[5] U. Remes, R. Karhila, and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," in *Proc. SSW*, 2013.

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

[7] 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," in *Proc. Interspeech*, 2018.

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

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

[10] T. H. Falk and W. Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1935–1947, Nov 2006.

[11] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[12] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[13] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC Press LLC, 1993.

[14] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, 2004.

[15] K. Pearson, "Notes on The History of Correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 10 1920. [Online]. Available: https://doi.org/10.1093/biomet/13.1.25

[16] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. [Online]. Available: http://dx.doi.org/10.2307/1412159

[17] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, 2018, pp. 2401–2405.

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

[19] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, 2015, pp. 4580–4584.

[20] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Temporal attentive pooling for acoustic event detection," in *Proc. Interspeech*, 2018, pp. 1354–1357.

[21] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. Interspeech*, 2015.

[22] D. Wang, S. Lv, X. Wang, and X. Lin, "Gated convolutional LSTM for speech commands recognition," in *Proc. International Conference on Computational Science*, 2018, pp. 669–681.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*. USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.