SVSNet: An End-to-End Speaker Voice Similarity Assessment Model

Cheng-Hung Hu , Yu-Huai Peng, Junichi Yamagishi , *Senior Member, IEEE*, Yu Tsao , *Senior Member, IEEE*, and Hsin-Min Wang , *Senior Member, IEEE*

Abstract—Neural evaluation metrics derived for numerous speech generation tasks have recently attracted great attention. In this paper, we propose SVSNet, the first end-to-end neural network model to assess the speaker voice similarity between converted speech and natural speech for voice conversion tasks. Unlike most neural evaluation metrics that use hand-crafted features, SVSNet directly takes the raw waveform as input to more completely utilize speech information for prediction. SVSNet consists of encoder, co-attention, distance calculation, and prediction modules and is trained in an end-to-end manner. The experimental results on the Voice Conversion Challenge 2018 and 2020 (VCC2018 and VCC2020) datasets show that SVSNet outperforms well-known baseline systems in the assessment of speaker similarity at the utterance and system levels.

Index Terms—Neural evaluation metrics, speech similarity assessment, voice conversion.

I. Introduction

THE speech generated in voice conversion (VC) tasks remains challenging to effectively evaluate. In most studies, both objective and subjective evaluation results are reported to compare the performance of VC systems. For objective evaluation [1], measurements borrowed from the speaker recognition task are usually used. For subjective evaluation, a listening test is usually conducted. Compared with objective evaluation, subjective evaluation incurs more time and cost. Moreover, to reach unbiased results, a large amount of subjective tests must be carried out [2]. However, since the target users of VC are humans, the subjective evaluation results are more important than the objective counterparts. In our previous study [3], we proposed MOSNet, which can predict the mean opinion score (MOS) of human subjective ratings of speech quality and naturalness. MOSNet is formed by a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BLSTM) architecture. The results of large-scale human evaluation of Voice Conversion Challenge 2018 (VCC2018) demonstrated that MOSNet achieves a high correlation with

human MOS ratings at the system level and a fair correlation at the utterance level.

In [3], we also slightly modified MOSNet to predict the similarity scores. The preliminary results showed that the predicted similarity scores were fairly correlated with human similarity ratings. In this work, to further improve the similarity prediction, we propose a novel assessment model called SVSNet, which has two features: (1) To more accurately characterize speech signals, SVSNet directly takes the speech waveform as input. (2) SVSNet adopts a co-attention mechanism to deal with length mismatch, content mismatch, and switching of paired utterances. For (1), although hand-crafted features are widely used in many speech processing tasks, such as speaker verification (SV) [4], [5], VC [6], [7], speech synthesis [8], [9], and speech enhancement [10], we believe that the raw waveform contains the most complete information for two reasons. First, for most hand-crafted features, the phase information is ignored. However, many studies have shown that phase can provide useful information [11]–[13]. Second, to compute hand-crafted features, prior knowledge is required about feature extraction specifications, such as window size, shift length, and feature dimension. Improper specifications can lead to ineffective features, which can result in poor prediction performance. For (2), we want to deal with asymmetry. First, the two input utterances may have different lengths and speech content. Second, when the input order is switched, SVSNet should output the same prediction score. Therefore, we design a special co-attention mechanism, which aligns the two input utterances in both directions. Compared to simple alignment methods such as dynamic time warping (DTW), attention can handle content mismatches. We also argue that co-attention can provide more information for similarity assessment than single-sided attention. Our experimental results on the VCC2018 [14] and VCC2020 [15] datasets show that SVSNet can predict the similarity score of a VC system quite accurately. As per our knowledge, this is the first deep learning-based model for similarity assessment for VC tasks.

II. Related Works

A. Neural Evaluation Metrics

Conventional evaluation metrics are generally derived on the basis of signal processing and human auditory theories. For example, perceptual evaluation of speech quality (PESQ) [16] and short-time objective intelligibility (STOI) [17] are commonly used to evaluate the quality and intelligibility of processed speech. The normalized covariance measure (NCM) [18] and its extensions [19], [20] have been shown to be effective in measuring the intelligibility of normal speech and vocoded speech. In addition, some parametric distances are often used

to measure the difference between paired voices, such as speech distortion index (SDI) [21], mel-cepstral distance (MCD) [22], cepstrum distance (Cep) [23], segmental signal-to-noise ratio (SSNR) improvement [24], and scale-invariant source-to-noise ratio (SI-SNR) [25]. Several studies have indicated that these objective evaluation metrics may not truly reflect human perception [22]. Therefore, subjective listening evaluations are usually reported in speech generation studies. Unbiased subjective results, however, require a large number of tests, covering a wide range of listeners (gender, age, and hearing ability) and test samples, which makes listening tests challenging in terms of time and cost.

To address the above issues, several neural evaluation metrics have been proposed. For speech enhancement tasks, Quality-Net [26], DNSMOS [27], and STOI-Net [28] were proposed as non-intrusive tools for measuring speech quality and intelligibility. For VC, MOSNet [3] and MBNet [29] were proposed to measure the naturalness of converted speech. Mittag and Möller [30] proposed an assessment model for the text-to-speech synthesis task. To the best of our knowledge, no prior work has previously established neural evaluation metrics for the similarity assessment of VC tasks.

B. Similarity Prediction

The similarity prediction task resembles an SV task, which aims to determine whether the input speech is pronounced by a claimed speaker. For most SV systems, the test utterance and the enrollment utterance are first converted into embedding vectors through a neural network (NN), and then a similarity score between the two embedding vectors is calculated based on a distance function, such as cosine distance or another NN model [4], [31], [32]. The organizer of VCC2020 reported the results of speaker similarity evaluation using x-vector [1]. In [33], several deep speaker representation learning methods that considered the perceptual similarity among speakers were proposed for multispeaker generative modeling. Experimental results show that compared with speaker-classification-based speaker representation learning, perceptual-similarity-aware speaker representation learning has better performance in several speech generative tasks. The speaker embedding learned in this way may be more suitable for speaker similarity evaluation, but the $N \times N$ speaker-pair similarity matrix of N speakers as the training target is not available for the VC task.

C. Waveform Modeling

Recently, several approaches have been proposed to incorporate waveform modeling into speech processing tasks, such as speech recognition [34], speech enhancement [35], [36], speech separation [25], [37], speech vocoding [38]–[40], and SV [41], [42]. The main idea of these waveform modeling methods is that traditional hand-crafted feature extraction techniques can be substituted by NN models in a data-driven manner. To effectively model speech waveforms, a dilated architecture has been proposed to increase the reception field with the same number of model parameters [39]. Meanwhile, SincNet [43] processes the raw waveform with a set of parameterizable band-pass filters, where only the low and high cutoff frequencies of the band-pass filters are the parameters to be learned. Learning data-dependent and task-dependent filters provides greater flexibility than fixed feature processing procedures. The effectiveness of SincNet has been demonstrated in several studies [43]–[45].
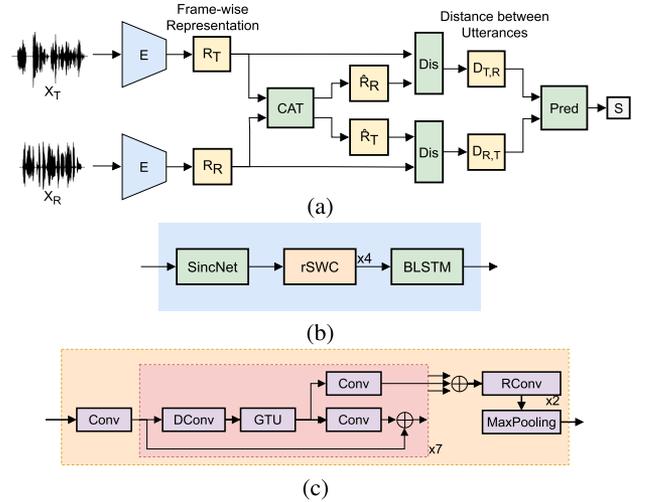
Fig. 1. (a) The architecture of SVSNet. E, CAT, Dis, and Pred blocks denote the Encoder, Co-attention, Distance, and Prediction modules. (b) The Encoder module. (c) The rSWC module. DConv is a dilated convolutional layer. RConv denotes a normal convolutional layer followed by a ReLU layer. The kernel size of the MaxPooling module is set to 3.

III. PROPOSED SVSNET

Fig. 1(a) shows the SVSNet architecture. The encoder (E) module (shared by two inputs) encodes the waveforms of the test and reference utterances into frame-wise representations (R_T and R_R). Unlike the attention module in [32], which only aligns the test utterance with the enrollment utterance in one direction, to maintain the symmetry, the "Co-attention" module aligns the two representations in two directions. Then, two distances, namely $D_{T,R}$ (between R_T and $\hat{R}_R$) and $D_{R,T}$ (between R_R and $\hat{R}_T$), are computed by the "Distance" module and used to calculate the final similarity score by the "Prediction" module. We study two types of prediction modules: regression-based and classification-based. Their outputs are a continuous score and a score-level category, respectively.

A. Encoder

Fig. 1(b) shows the architecture of the encoder in SVSNet. First, the input waveform is processed by SincNet, which contains K learnable band-pass filters, to decompose the input signal to K subband signals. The K subband signals are then processed by four stacked residual-skipped-WaveNet convolution (rSWC) layers and a BLSTM layer. Fig. 1(c) shows the rSWC layer. The core of the rSWC layers is the convolutional layers with dilation sizes of (1, 2, 4, 8, 16, 32, and 64), followed by a gated tanh unit (GTU) [46]. In addition, the maxpooling layer with a stride size of 3 is to downsample the feature sequence. As shown in Fig. 1(a), given the test utterance X_T and the reference utterance X_R, the encoder outputs R_T and R_R, respectively.

B. Co-Attention Module

The co-attention module is used to align the representation of the other input with that of one input:

$$\hat{R}_R = Attention(R_T, R_R, R_R),$$

$$\hat{R}_T = Attention(R_R, R_T, R_T). \quad (1)$$

Two pairs of aligned representation sequences are obtained, namely $(R_T, \hat{R}_R)$ and $(R_R, \hat{R}_T)$, which are then fed to the distance calculation module. We used the scaled dot-product attention mechanism [47] in this study.

C. Distance Calculation and Prediction Modules

We extend the attentive pooling used in SV [48] to our work. We average the representations of an utterance over time to obtain the utterance embedding and compute the 1-norm distance of each dimension of two means:

$$D_{T,R} = ||Mean(R_T) - Mean(\hat{R}_R)||_1,$$
$$D_{R,T} = ||Mean(R_R) - Mean(\hat{R}_T)||_1. \qquad (2)$$

Then, the two distances are fed to the prediction module to obtain the similarity score:

$$\hat{S}_T = \sigma(f_{lin2}(\rho_{ReLU}(f_{lin1}(D_{T,R})))),$$
$$\hat{S}_R = \sigma(f_{lin2}(\rho_{ReLU}(f_{lin1}(D_{R,T})))), \qquad (3)$$

where $\sigma(.)$ denotes an activation function, $f_{lin1}(.)$ and $f_{lin2}(.)$ denote two linear layers, and $\rho_{ReLU}(.)$ denotes a rectified linear unit (ReLU) activation function. The number of nodes of the final linear output layer is 1 for the regression model and 4 for the classification model, i.e., $\hat{S}_T, \hat{S}_R \in R^1$ for regression, and R^4 for classification. The activation function is an identity function for the regression model and a softmax function for the classification model. Finally, the prediction module obtains the final score by $\hat{S} = (\hat{S}_T + \hat{S}_R)/2$.

D. Model Training

SVSNet is trained on a set of reference-test utterance pairs with corresponding human labeled similarity scores. We implemented two versions of SVSNet by using two types of prediction modules: regression and classification. The corresponding SVSNet models are termed SVSNet(R) and SVSNet(C), respectively. Given the ground-truth similarity score S and the predicted similarity score $\hat{S}$, the mean squared error (MSE) loss is used to train SVSNet(R), and the cross entropy (CE) loss is used to train SVSNet(C).

IV. EXPERIMENTS

A. Experimental Setup

Since 2016, the VC challenge (VCC) has been held three times. The task is to modify an audio waveform so that it sounds as if it was from a specific target speaker other than the source speaker. In each challenge, a large-scale crowdsourced human perception evaluation was conducted to test the quality and similarity of the converted utterances. In this study, we focused on the similarity evaluation.

In VCC2018, there were 36 VC systems and two reference systems. Of the two reference systems, one took the source input as the output (used to evaluate the lower performance bound), and the other took the target output as the output (used to evaluate the upper performance bound). The similarity evaluation was conducted on 21,562 converted-natural utterance pairs, with two reference systems each accounting for 360 pairs, and the remaining systems each accounting for 570 to 599 pairs. Each pair was evaluated by 1 to 8 subjects, with a score ranging from 1 (same speaker) to 4 (different speakers). A total of 30,864 speaker

TABLE I
RESULTS OF SVSNET AND MOSNET ON VCC2018

Method	Utterance-level				System-level		
	ACC	LCC	SRCC	MSE	LCC	SRCC	MSE
MOSNet(R)	0.276	**0.439**	0.421	**1.024**	**0.845**	**0.755**	0.134
MOSNet(C)	0.432	0.426	**0.434**	1.463	0.796	0.665	**0.087**
SVSNet(R)	0.438	**0.574**	**0.572**	**0.844**	**0.965**	**0.903**	**0.044**
SVSNet(C)	**0.482**	0.552	0.554	1.137	0.933	0.890	0.057

similarity scores were obtained. The two reference systems were scored 614 and 624 times, and each of the remaining systems was scored 822 to 825 times. In addition, for each system, half of the pairs were converted-target pairs, and the other half of the pairs were converted-source pairs. The score of a system was the average score of its converted-target pairs. The detailed description of the corpus, listeners and evaluation methods can be found in [14]. In this study, the dataset was divided into 24,864 pair-score samples for training and 6,000 pair-score for testing.

We used MOSNet [3] as the baseline. MOSNet was originally proposed for quality assessment, but a modified version was used for similarity assessment. Like SVSNet, the models via regression and classification are termed MOSNet(R) and MOSNet(C), respectively. Performance was evaluated in terms of accuracy (ACC), linear correlation coefficient (LCC) [49], Spearman's rank correlation coefficient (SRCC) [50], and MSE at both utterance and system levels. The utterance-level evaluation was calculated from the predicted score and the ground-truth score for each pair of utterances. Note that if there was more than one score for a pair of utterances, the average value was used as the ground-truth score in testing. The system-level evaluation was calculated based on the average predicted score and the average ground-truth score for each system. When treating similarity prediction as a classification task, the original labels were used as the ground-truth. When treating similarity prediction as a regression task and evaluating performance on the basis of ACC, the outputs of SVSNet(R) and MOSNet(R) were rounded and clipped to the nearest integer (i.e., 1, 2, 3, or 4).

Since two different sampling rates (22,050 and 16,000 Hz) were used in the VCC2018 dataset, we reduced the sampling rate of all utterances to 16,000 Hz. For the encoder, the number of output channels of SincNet, the output size of the WaveNet convolutional layers, and the hidden size of BLSTM were 64, 64, and 256, respectively. The hidden size of the linear layers in the distance module was 128, and the output size was 1 and 4 for the regression output and classification output, respectively. We used the Adam optimizer to train the model, where the learning rate, β_1, and β_2 were 1e-4, 0.5, and 0.999, respectively. The batch size was set to 5. The model parameters were initialized by Xavier Uniform.

B. Experimental Results on VCC2018

First, we compare SVSNet with MOSNet. The results are shown in Table I. From the table, several observations can be drawn. First, SVSNet consistently outperforms MOSNet in all metrics. Second, SVSNet performs better in regression mode than in classification mode, but MOSNet is difficult to determine which mode is better. Third, the high LCC (0.965 based on regression and 0.933 based on classification) and SRCC (0.903 based on regression and 0.890 based on classification) scores indicate that the predicted ranking of the 38 systems by SVSNet is close to that of human evaluation.

TABLE II
RESULTS OF ABLATION STUDY ON SVSNET

Method	Utterance-level				System-level		
	ACC	LCC	SRCC	MSE	LCC	SRCC	MSE
SVSNet(R)	**0.438**	**0.574**	**0.572**	**0.844**	**0.965**	**0.903**	**0.044**
SVSNet(R)$_{spec}$	0.417	0.545	0.544	0.879	0.929	0.857	0.092
SVSNet(R)$_{ss}$	0.404	0.550	0.552	0.874	0.929	0.811	0.092
SVSNet(R)$_{dtw}$	0.353	0.534	0.527	0.906	0.873	0.809	0.093
SVSNet(C)	0.482	0.552	0.554	1.137	0.933	0.890	0.057
SVSNet(C)$_{spec}$	0.459	0.489	0.492	1.211	0.904	0.832	0.053
SVSNet(C)$_{ss}$	0.481	0.531	0.534	1.164	0.932	0.822	0.044
SVSNet(C)$_{dtw}$	0.462	0.512	0.517	1.173	0.886	0.770	0.071

Next, we further evaluate SVSNet(R) and SVSNet(C). We first study the effect of waveform processing. For a fair comparison, we replaced the waveform input and the SincNet in SVSNet(R) and SVSNet(C) with the 257-dimensional spectrogram used in MOSNet. The corresponding models are termed SVSNet(R)$_{spec}$ and SVSNet(C)$_{spec}$. From Table II, we can see that SVSNet(R) is better than SVSNet(R)$_{spec}$ in all metrics, while SVSNet(C) is better than SVSNet(C)$_{spec}$ in all metrics except for the system-level MSE. The results confirm the advantage of using waveforms instead of spectrograms as input. Then, we study the effect of co-attention. In Table II, SVSNet(R)$_{ss}$ and SVSNet(C)$_{ss}$ denote SVSNet with single-sided attention, which aligns the converted utterance with the natural utterance. SVSNet(R)$_{dtw}$ and SVSNet(C)$_{dtw}$ denote SVSNet with DTW, where the alignment is based on the dB-scale 2-norm distance between the spectrograms of two utterances. We can see that SVSNet(R)$_{ss}$ and SVSNet(C)$_{ss}$ are better than SVSNet(R)$_{dtw}$ and SVSNet(C)$_{dtw}$ in all metrics, respectively. Moreover, SVSNet(R) is better than SVSNet(R)$_{ss}$ in all metrics, while SVSNet(C) is better than SVSNet(C)$_{ss}$ in all metrics except for the system-level MSE. The results confirm the effectiveness of co-attention.

C. Experiment Results on VCC2020

Voice Conversion Challenge 2020 (VCC2020), the next edition of VCC2018, includes two tasks, namely intra-language VC and cross-language VC. The intra-language task considered 16 source-target speaker pairs, and the cross-language task considered 24 source-target speaker pairs. Each source-target speaker pair contained 5 converted-target utterance pairs, and each converted-target utterance pair was evaluated by 12 subjects (for intra-language) and 8 subjects (for cross-language). Therefore, there were 960 evaluation scores per system ($16\times5\times12$ for intra-language and $24\times5\times8$ for cross-language). There were 31 VC systems for the intra-language task, 28 VC systems for the cross-language task, and three reference systems for evaluating the lower and upper performance bounds. To study the impact of corpus mismatch, we tested the models trained on the VCC2018 training set on the complete VCC2020 dataset to perform system-level evaluation. Since most systems used conventional vocoders in VCC2018 and neural vocoders in VCC2020, the corpus mismatch is significant. It is also worth mentioning that SVSNet(R)$_{dtw}$ and SVSNet(C)$_{dtw}$ are not applicable because the two utterances to be compared here have different content.

The results are shown in Table III. We can see that the scores of both SVSNet and MOSNet are lower than those reported earlier due to corpus mismatch, while SVSNet still outperforms MOSNet. Following Das *et al.* [1], we tested the performance with another prediction model formed by a cosine similarity measure based on 128-dimensional linear discriminant analysis (LDA) reduced x-vectors. The similarity scores were linearly

TABLE III
RESULTS OF DIFFERENT MODELS ON VCC2020

Method	System-level		
	LCC	SRCC	MSE
MOSNet(R)	0.630	0.504	2.964
MOSNet(C)	0.723	0.736	0.211
x-vector	0.844	0.911	1.213
SVSNet(R)	0.819	0.841	0.361
SVSNet(C)	0.775	0.765	0.244
SVSNet(R)+x-vector$_{SF}$	0.872	0.926	0.867
SVSNet(C)+x-vector$_{SF}$	0.877	**0.936**	0.711
SVSNet(R)+x-vector$_{FF}$	0.863	0.921	**0.141**
SVSNet(C)+x-vector$_{FF}$	**0.885**	0.926	0.373
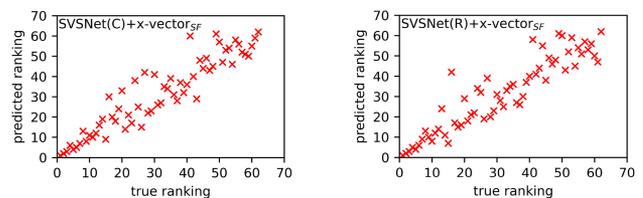
Fig. 2. Predicted system rankings of the score fusion models on VCC2020.

mapped to [1, 4]. The results show that with an extra and massive dataset for pretraining, the x-vector model outperforms both SVSNet and MOSNet on LCC and SRCC. We also constructed fusion models to utilize the information of x-vector into SVSNet. For the score fusion models (SVSNet(R)+x-vector$_{SF}$ and SVSNet(C)+x-vector$_{SF}$), the fusion score was the weighted average of the scores of the SVSNet and x-vector models at a ratio of 3:7. The weight was determined based on the test set of VCC2018. For the feature fusion models (SVSNet(R)+x-vector$_{FF}$ and SVSNet(C)+x-vector$_{FF}$), the 1-norm distance of each dimension between two x-vectors was concatenated to $D_{T,R}$ and $D_{R,T}$ in (2). From Table III, we can see that all fusion models yield improvements over the SVSNet and x-vector models. Since the score fusion models achieve the best SRCC values in Table III, we compare their predicted system rankings with the true ranking in Fig. 2. Obviously, both models achieve fairly good prediction performance on VCC2020, although there is still room for further improvement.

V. CONCLUSIONS

In this paper, we have proposed SVSNet, an end-to-end neural similarity assessment model. The results of experiments on the large-scale human perception evaluation in VCC2018 and VCC2020 show that SVSNet, benefiting from the SincNet and the residual-skipped-WaveNet architecture, performs better than the previous model MOSNet in terms of LCC, SRCC, and MSE. It is also found that directly using the waveform as input without discarding the phase information will increase the prediction ability of our model. In the future, we plan to consider the theory of human perception to design a perception-based objective function to build a more robust quality and similarity prediction model. We will also explore the use of SVSNet in the training objective to guide VC models to generate utterances that are highly similar to the target speaker's speech.

REFERENCES

[1] R. K. Das *et al.*, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 99–120.

[2] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech*, 2015, pp. 3476–3480.

[3] C.-C. Lo *et al.*, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[5] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4052–4056.

[6] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 3364–3368.

[7] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[8] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc AAAI Conf. Artif. Intell.*, 2019, pp. 6706–6713.

[9] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[10] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[11] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 598–602.

[12] E. Loweimi, J. Barker, O. Saz-Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *Proc. Interspeech*, 2017, pp. 414–418.

[13] E. Loweimi, "Robust phase-based speech signal processing from source-filter separation to model-based robust ASR," Ph.D. dissertation, Univ. Sheffield, U.K., 2018.

[14] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey*, 2018, pp. 195–202.

[15] Z. Yi *et al.*, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 80–98.

[16] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[18] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[19] F. Chen and P. C. Loizou, "Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech," *J. Acoust. Soc. Amer.*, vol. 128, no. 6, pp. 3715–3723, 2010.

[20] F. Chen and P. C. Loizou, "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 131, no. 5, pp. 4104–4113, 2012.

[21] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[22] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128.

[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[24] S. R. Quackenbush, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986.

[25] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[26] 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, pp. 1873–1877.

[27] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6493–6497.

[28] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 482–486.

[29] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 391–395.

[30] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Proc. Interspeech*, 2020, pp. 1748–1752.

[31] C. Zeng, X. Wang, E. Cooper, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022.

[32] Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui, and D. Yu, "Seq2Seq attentional siamese neural networks for text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6131–6135.

[33] Y. Saito, S. Takamichi, and H. Saruwatari, "Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1033–1048, 2021.

[34] T. Parcollet, M. Morchid, and G. Linares, "E2E-SINCNET: Toward fully end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7714–7718.

[35] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[36] M. P. Shifas, N. Adiga, V. Tsiaras, and Y. Stylianou, "A non-causal FFTNet architecture for speech enhancement," in *Proc. Interspeech*, 2019, pp. 1826–1830.

[37] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[38] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6199–6203.

[39] A. van. den. Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. ISCA Speech Synth. Workshop*, 2016, p. 125.

[40] K. Kumar *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.

[41] J.-W. Jung, H.-S. Heo, J.-H. Kim, H.-J. Shim, and H.-J. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Proc. Interspeech*, 2019, pp. 1268–1272.

[42] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5349–5353.

[43] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 1021–1028.

[44] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1888–1900, 2020.

[45] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6465–6469.

[46] A. van. den. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 4797–4805.

[47] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[48] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

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

[50] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 39, no. 5, pp. 1137–1150, 1904.