INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Self-Supervised Learning Based Phone-Fortified Speech Enhancement

Yuanhang Qiu, Ruili Wang[*]*, Satwinder Singh, Zhizhong Ma, Feng Hou*

School of Natural and Computational Sciences, Massey University, New Zealand
{y.qiu1,ruili.wang,s.singh4,z.ma,f.hou}@massey.ac.nz

Abstract

For speech enhancement, deep complex network based methods have shown promising performance due to their effectiveness in dealing with complex-valued spectrums. Recent speech enhancement methods focus on further optimization of network structures and hyperparameters, however, ignore inherent speech characteristics (e.g., phonetic characteristics), which are important for networks to learn and reconstruct speech information. In this paper, we propose a novel self-supervised learning based phone-fortified (SSPF) method for speech enhancement. Our method explicitly imports phonetic characteristics into a deep complex convolutional network via a Contrastive Predictive Coding (CPC) model pre-trained with self-supervised learning. This operation can greatly improve speech representation learning and speech enhancement performance. Moreover, we also apply the self-attention mechanism to our model for learning long-range dependencies of a speech sequence, which further improves the performance of speech enhancement. The experimental results demonstrate that our SSPF method outperforms existing methods and achieves state-of-the-art performance in terms of speech quality and intelligibility.
Index Terms: speech enhancement, self-supervised learning, contrastive predictive coding, self-attention mechanism

1. Introduction

Speech enhancement is an important speech processing task aiming to improve the intelligibility and overall perceptual quality of a contaminated speech signal [1]. The intelligibility is a measurement of how comprehensible a speech signal is, while the perceptual quality measures how easy it is for a listener to perceive the content of a speech signal. Normally, a perceptual high-quality speech sounds more natural, rhythmic, yet less raspy, hoarse or scratchy, etc [1].

In the real world, additive noise (e.g., fan noise) and convolutional noise (e.g., room reverberation) are two common noise types that can degrade speech signals drastically. Correspondingly, many approaches (e.g., denoising, dereverberation) have been proposed and widely used in practical applications such as mobile communication [2], hearing-aids [3], and noise-robust speech recognition [4]. However, the performance of speech enhancement, especially in a real-life environment, still needs to be improved further.

Currently, data-driven deep learning based methods have been thriving in speech signal processing [5, 6], computer vision [7], and natural language processing [8]. For speech enhancement, existing deep learning based models, which pursued the optimal structures with dozens of network layers, have improved the performance of speech enhancement in different scenarios [9, 10, 11, 12, 13]. However, the performance might not be improved all along if we just stack the network layers exhaustively [14]. In addition, ignoring essential information of speech signals (e.g., phase and phonetic information) is also a challenging issue in speech enhancement [15].

For phase-aware speech enhancement, deep complex network based methods have demonstrated their effectiveness in dealing with complex-valued spectrums [16]. A deep complex network was originally proposed to construct richer and more versatile representations of an image or audio signal [17]. Based on this network, recently, Hsieh et al. [18] proposed a Phone-Fortified Perceptual loss (PFP) for enhancing network optimization. They indicated that the phonetic characteristic information is the key to optimizing speech enhancement with respect to human perception, but the latent features for speech characteristics learning in previous models seemed to be lacking in phonetic characteristic information. Moreover, they also indicated that the objective functions based on point-wise distances might not fully reflect the perceptual difference between noisy and clean speech signals. Thus, the phonetic characteristics extracted by a pre-trained Contrastive Predictive Coding (CPC) model were introduced in their model but just for loss calculation.

Inspired by [18], we propose a new speech enhancement method, which focuses on improving speech representation learning via importing the phonetic characteristics into an improved deep complex network explicitly. We adopt a self-supervised learning based CPC model for speech phonetic information extraction because of CPC's great speech representation learning capability. To import phonetic characteristics, we propose a new feature embedding network to re-embed the extracted features and then fuse them with the original frequency spectrum features. We consider that explicitly supplementing speech phonetic information can effectively enhance speech representation learning. Moreover, we also apply the self-attention mechanism to the deep complex network specifically, which aids in learning long-range dependencies of a speech sequence and improving the performance of speech enhancement further. In our experiments, we explore multiple CPC-based pre-trained models for speech phonetic information extraction and compare their performance fully. We also investigate the impact of the size of training data on our enhancement model and unfold the results in terms of signal-to-noise ratios (SNR) and noise types.

In the following, we give the related work in Section 2. We describe the details of our model in Section 3. In Section 4, we describe the setup of experiments. The experimental results are presented in Section 5. Finally, we report our conclusions and acknowledgements in Section 6 and Section 7.

2. Related Work

The existing work related to speech phase information preservation and representation learning is introduced in this section.

[*]Corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-734

2.1. Phase Information Preservation

For phase information preservation, an end-to-end speech processing framework has been considered as a plausible solution, which receives a raw speech waveform as input and outputs the processed speech waveform directly [19]. Since a raw speech waveform naturally contains phase information, the end-to-end speech enhancement can preserve the phase information from the contaminated speech sequence without extra handcrafted feature pre-processing. Generally, the handcrafted preprocessing operation such as traditional speech feature extraction may only capture acoustic information, but ignore other important information such as the speech phase [20]. The end-to-end framework can alleviate this problem by taking in the raw speech waveform. However, for speech enhancement, reusing the phase information of noisy speech generally causes a serious mismatch between reconstructed speech and clean speech, especially under extremely noisy conditions [16].

Further, jointly estimating the speech magnitude and phase information with a complex-valued network is another approach [20]. Unlike the real-valued network that only changes the scale of the magnitude spectral mapping without the phase information processing [15], the complex-valued network [17] learns speech magnitude and phase information with the real and imaginary part, respectively, which has been proven to be an effective framework [16, 18] for speech enhancement. Thus, we also adopt the complex-valued network for speech magnitude and phase response preservation in this research.

2.2. Speech Representation Learning

Learning appropriate speech representation is a fundamental and effective way to improve speech signal processing. With the development of speech signal processing, different speech feature representations were proposed such as Mel-frequency cepstral coefficients (MFCCs), a general all-purpose frame-level acoustic feature [21]; Identity vector (I-vector), a high-dimensional utterance-level speech representation [22]; Speech2vec (i.e., speech version of word2vec [23]), a semantic representation of an audio segment with a fixed-length vector [24]. These speech feature representations have been used widely in specific speech tasks.

Recently, a self-supervised learning based CPC model was developed to extract useful data representation from high-dimensional data space with a powerful autoregressive model [25]. Specifically, the probabilistic contrastive loss was proposed to induce the latent space to capture information that was maximally useful to predict future samples. The self-supervised learning mechanism enables CPC to learn a general and effective representation with massive unlabelled data. CPC was tested in different data modalities such as speech, images, natural language and obtained promising results [25, 26, 27]. In this paper, we introduce two CPC-based models (i.e., wav2vec [26] and vq-wav2vec [28]) for speech representation learning.

2.2.1. wav2vec

Wav2vec [26], a pre-trained CPC model as shown in Figure 1, can learn a general speech representation by training with large amounts of unlabelled raw audio data. The model consists of two convolutional neural networks (i.e., an encoder network and a context network). The encoder network embeds the raw audio signal in a latent space and outputs a low-frequency representation to the context network (also known as an aggregator), which creates a contextualized vector representation by combining the latent representation from multiple time steps. The model can be trained to distinguish a future sample from the distractor samples, which is drawn from a proposal distribution, by minimizing the contrastive loss for each step. After training, the output of the context network can be considered as the desired representation of the input audio.

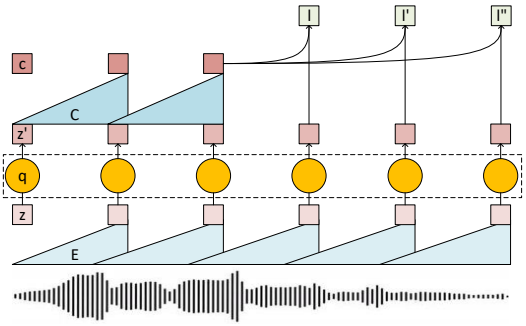

Figure 1: *The framework of wav2vec and vq-wav2vec. For wav2vec, the encoder (E) network maps the raw audio to a dense representation z. z is aggregated into a context (C) network for representation c, which refers to the contrastive loss calculation (l) with the future samples. With the addition of a quantized (q) layer, this framework represents vq-wav2vec.*

2.2.2. vq-wav2vec

Vq-wav2vec [28] is based on wav2vec, as shown in Figure 1. It has an architecture like wav2vec but with an additional quantization module between the encoder network and the context network. The quantization module replaces the original representation z by z' from a fixed size codebook, of which the one-hot representation can be computed by using the Gumbel-Softmax or K-means clustering approaches [28]. Vq-wav2vec, which learns the discrete representations of fixed length segments of an audio signal, enables well-performing language processing algorithms [29] to be applied directly to speech data. In this paper, we use both wav2vec and vq-wav2vec as phonetic characteristics extractors for speech enhancement model training and compare their performances in the Results Section.

3. The proposed method

In this paper, we propose a new self-supervised learning based phone-fortified (SSPF) method for speech enhancement. Our method adopts a deep complex convolutional network to estimate a complex ratio mask for noisy information filtering. As shown in Figure 2, the deep complex network is a refined U-Net architecture [30] and incorporates multiple well-defined complex-valued blocks to deal with complex-valued spectrum [17]. In detail, the complex U-Net adopts multiple complex convolutional and transposed convolutional layers with skip-connections [31], complex batch normalization, and LeakyRelu activation [18] as the main components, which are admittedly functional parts to learn representations of multiple data modalities effectively such as image and speech. For speech enhancement, the complex-valued architecture with real and imaginary parts is used to learn speech magnitude and phase information simultaneously [17].

Referring to Figure 2, our speech enhancement model converts a noisy speech signal to an enhanced speech signal with a learnable complex mask derived from the complex U-Net

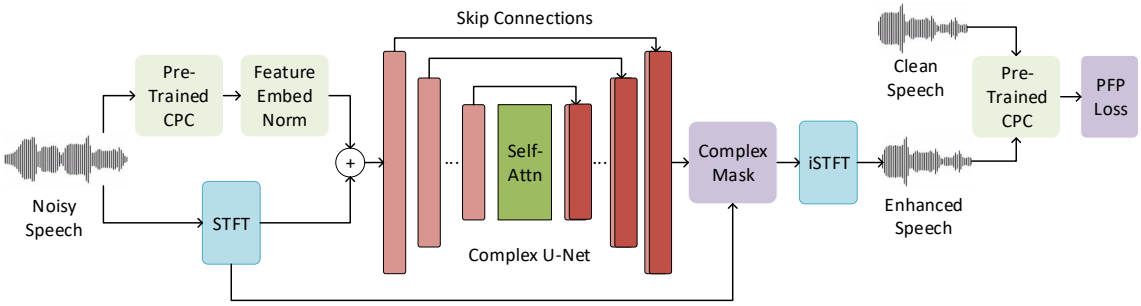

Figure 2: *The framework of our proposed speech enhancement model based on a complex U-Net and CPC. The complex U-Net estimates a complex-valued ratio mask with the fused noisy speech representation. The mask can filter the noisy spectrum and achieve the enhanced spectrum with inverse STFT. The CPC-based pre-trained model extracts linguistic information of an enhanced waveform and paired clean waveform for PFP loss calculation.*

model. To improve the speech representation learning of our speech enhancement model, we employ a pre-trained CPC-based model to extract the phonetic characteristics, which are then fused with the standard frequency spectrum feature converted by short-time Fourier transform (STFT). Here, a simple feature embedding network is proposed to re-embed and normalize the representation of the phonetic characteristics so that it can be fused with the frequency spectrum feature by point-wise addition. The proposed feature embedding network consists of multiple transposed convolutional layers followed by Relu activation and max-pooling [32] and the network parameters are trained along with the complex U-Net simultaneously. We apply a self-attention layer in the middle of complex U-Net since the self-attention layer can learn the long-range dependencies of a speech sequence and further improve speech representation learning. At last, the frequency spectrum feature is point-wise multiplied with the complex-valued ratio mask again to derive the enhanced spectrum, and then the inverse STFT module transforms the enhanced spectrum to a speech waveform.

During the loss calculation, the same pre-trained CPC model is applied again to transform the waveform into a batch of sequence vectors, which are rich in phone-fortified information for a speech evaluation. We follow the effective phone-fortified perceptual loss proposed in [18]. The formulation can be described as below:

$$L_{pfp}(x, \hat{x}) := E_{x,\hat{x}\sim D}[\|\phi_{cpc}(x) - \phi_{cpc}(f(\hat{x}))\|_1], \quad (1)$$

where x denotes the clean speech; $\hat{x}$ denotes the paired noisy speech; ϕ_{cpc} is the pre-trained CPC model for phonetic representation extraction; f denotes the enhancement procedure. The PFP loss calculates the absolute distance between a clean and an enhanced speech phonetic vector.

4. Experiments

4.1. Database

The VCTK database [33, 34] is an open and standard speech corpus for performance evaluation of speech enhancement systems. The original clean speech was selected from the Voice Bank corpus [35]. In addition, eight real noise files and two artificially generated noise files were used to generate paired noisy speech.

For training, two sub-databases were created: one has 28 speakers, which includes 14 males and 14 females with the same accent (England) [33]; another one has 56 speakers, which includes 28 males and 28 females with different accents (Scotland and United States) [34]. The SNR values were set to 15dB, 10dB, 5dB and 0dB. Moreover, each clean speech waveform was normalized, and the silence segments were trimmed off at the beginning and the ending when the silence segments were longer than 200ms.

Another two speakers (a male and a female), not included in the training data, were selected as the test data with an England accent. Five other noise types, different from training data, were selected from the DEMAND database[1], including a domestic noise (in a living room), an office noise (in an office space), a transport noise (in a bus) and two street noises (in an open area cafeteria and a public square). The SNR values were set to 17.5dB, 12.5dB, 7.5dB and 2.5dB, respectively.

Table 1: *The evaluation results of various methods with the 28-speaker VCTK training data. The compared methods are Wiener filtering [36], BiLSTM [37], CRN-MSN [5], NAAGN [11], U-NetC [12, 13], PHASEN [20], HiFi-GAN [38], T-GSA [39]. "-" denotes that the result is not reported or not available. "†" denotes that we reproduced the results with the provided open resource. The best scores are highlighted in bold.*

Models	PESQ	CSIG	CBAK	CVOL	STOI
Noisy	1.97	3.35	2.44	2.63	0.921
Wiener	2.22	3.23	2.68	2.67	-
BiLSTM	2.70	3.99	2.95	3.34	0.925
CRN-MSE	2.74	3.86	3.14	3.30	0.934
NAAGN	2.90	4.13	3.50	3.51	**0.948**
U-NetC	2.90	**4.22**	3.32	3.58	0.938
HiFi-GAN	2.94	4.07	3.07	3.49	-
PHASEN	2.99	4.21	3.55	3.62	-
T-GSA	**3.06**	4.18	**3.59**	3.62	-
W2V†	2.98	4.01	3.46	3.50	0.945
W2V$_F$	3.02	4.11	3.52	3.60	0.943
W2V$_{FSA}$	3.04	4.17	**3.59**	**3.63**	0.945
W2V-G	2.99	4.09	3.46	3.55	0.944
W2V-G$_F$	3.00	4.12	3.48	3.57	0.944
W2V-G$_{FSA}$	2.96	4.04	3.47	3.50	0.944
W2V-K	2.93	4.06	3.45	3.50	0.942
W2V-K$_F$	2.95	4.11	3.54	3.57	0.943
W2V-K$_{FSA}$	3.00	4.14	3.46	3.58	0.942

[1] http://parole.loria.fr/DEMAND/

Table 2: *The evaluation results of W2V$_{FSA}$ with 56-speaker and 84-speaker (the mixture of 28-speaker and 56-speaker). Meanwhile, all scenes with different SNR and noise types are demonstrated separately. The best scores are highlighted in bold.*

Metric	W2V$_{FSA}$	2.5dB	7.5dB	12.5dB	17.5dB	Living	Office	Bus	Cafe	Square	Average
PESQ	56-spk	2.61	3.04	3.29	**3.55**	2.86	**3.57**	3.52	2.61	2.97	3.10
	84-spk	**2.67**	**3.08**	**3.31**	3.53	**2.93**	3.55	3.52	**2.65**	**3.02**	**3.13**
CSIG	56-spk	3.75	4.22	4.47	**4.70**	4.02	**4.70**	**4.63**	3.82	4.20	4.27
	84-spk	**3.86**	**4.26**	4.47	4.66	**4.09**	4.68	4.62	**3.87**	**4.24**	**4.30**
CBAK	56-spk	3.12	3.50	3.73	3.99	3.38	3.98	3.74	3.25	3.53	3.57
	84-spk	**3.19**	**3.55**	**3.76**	3.99	**3.44**	3.98	**3.79**	**3.28**	**3.57**	**3.61**
CVOL	56-spk	3.17	3.64	3.90	**4.16**	3.44	**4.17**	**4.10**	3.21	3.59	3.70
	84-spk	**3.26**	**3.67**	3.90	4.11	**3.51**	4.14	4.09	**3.26**	**3.64**	**3.72**
STOI	56-spk	0.920	0.950	0.960	0.966	0.939	0.967	0.964	0.924	0.946	0.947
	84-spk	**0.923**	**0.952**	**0.961**	**0.967**	**0.943**	0.967	**0.965**	**0.927**	**0.948**	**0.950**

4.2. Setup

Before training, we randomly separate a validation part from the training data at a ratio of 9:1. The model is optimized using the RAdam optimizer [40] with a learning rate of 0.0001 and weight decay of 0.1. The model is trained for 100 epochs with a batch size of 8. For each epoch, we save the best model according to the performance evaluated with the validation data. During the inference stage, each noisy speech is point-wise multiplied with the complex mask for noisy information filtering and then is converted to an enhanced speech waveform.

4.3. Evaluation Metrics

Although subjective evaluation is accurate and reliable, it is costly and time-consuming [41]. Many objective evaluation measures can evaluate enhanced speech performance with high correlation. For speech quality evaluation, we use an effective full-reference speech quality evaluation algorithm [41] namely Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5), which compares each sample of the reference signal (clean speech) to each corresponding sample of the degraded signal, and analyses sample-by-sample after a temporal alignment of corresponding excerpts of reference and testing signals. Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (COVL: from 1 to 5). For speech intelligibility evaluation, we adopt the Short-Time Objective Intelligibility (STOI) [42] that is based on a correlation coefficient between the temporal envelopes of the clean and degraded speech.

5. Results

As shown in Table 1, we explore three CPC-based models for phonetic information extraction and loss calculation: wav2vec (W2V), vq-wav2vec with Gumbel-Softmax (W2V-G), and vq-wav2vec with K-means clustering (W2V-K). To implement the ablation experiments, we test our speech enhancement model with the frequency spectrum feature only (i.e., W2V, W2V-G, W2V-K as shown in Table 1), with the phonetic embedding fused feature (i.e., W2V$_F$, W2V-G$_F$, W2V-K$_F$), and further with self-attention mechanism (i.e., W2V$_{FSA}$, W2V-G$_{FSA}$, W2V-K$_{FSA}$).

Compared with the previous methods, W2V$_{FSA}$ (i.e., wav2vec model with fused features and self-attention mechanism) obtains the best score in CBAK (3.59), and CVOL (3.63). In terms of other metrics, W2V$_{FSA}$ also achieves competitive results. Focusing on the ablation experiments, we find that the performance can be improved gradually when we import the phonetic information and employ the self-attention mechanism with both wav2vec and vq-wav2vec. Thus, we conclude that our proposed method can effectively learn speech representation and the phone-fortified method has a huge potential to improve speech enhancement. In addition, we find that the wav2vec based models outperform vq-wav2vec models. We infer that the non-discrete representations learned by wav2vec outperform the discrete representations learned by vq-wav2vec for speech enhancement.

To further explore the effectiveness of our proposed method, we conduct another experiment with different sizes of training data. Meanwhile, we reveal the results in four SNR and five noise type scenes respectively as shown in Table 2. In this experiment, we present the evaluation results on the best W2V$_{FSA}$ model trained with 56-speaker and 84-speaker (the mixture of 28-speaker and 56-speaker). As we can see, with more training data, the W2V$_{FSA}$ model further improves the performance and achieves state-of-the-art performance in most scenes. Moreover, we also find that our method can perform better in scenarios where SNR is as low as 2.5dB.

6. Conclusions

In this paper, we propose a novel self-supervised learning based phone-fortified method (SSPF) for speech enhancement. Our SSPF method can effectively estimate a complex ratio mask for noisy speech filtering with a self-attention mechanism boosted complex U-Net model. SSPF explicitly imports the phonetic characteristics into the enhancement model via a self-supervised learning based CPC model to further improve speech phase estimation and representation learning. The experimental results demonstrate that our method outperforms previous methods in most evaluation metrics and achieves state-of-the-art performance with more training data in terms of speech quality and intelligibility. In future work, we will further explore the effectiveness of inherent speech information for speech enhancement. In addition, we will also extend our proposed method to noise-robust speech recognition or speech synthesis.

7. Acknowledgements

This work was supported in part by the China Scholarship Council (CSC) and the Catalyst: Strategic – New Zealand-Singapore Data Science Research Programme.

8. References

[1] P. C. Loizou, *Speech enhancement theory and practice*. CRC press, 2013.

[2] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[3] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM TASLP*, vol. 27, no. 1, pp. 99–113, 2019.

[4] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2080–2091, 2019.

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

[6] S. Singh, R. Wang, and Y. Qiu, "DEEPF0: end-to-end fundamental frequency estimation for music and speech signals," in *Proc. of ICASSP*, 2021.

[7] P. Shamsolmoali, M. Zareapoor, R. Wang, D. K. Jain, and J. Yang, "G-GANISR: gradual generative adversarial network for image super resolution," *Neurocomputing*, vol. 366, pp. 140–153, 2019.

[8] F. Hou, R. Wang, J. He, and Y. Zhou, "Improving entity linking through semantic reinforced entity embeddings," in *Proc. of ACL*, 2020, pp. 6843–6848.

[9] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Workshop of HSCMA*. IEEE, 2017, pp. 136–140.

[10] Y. Qiu and R. Wang, "Adversarial latent representation learning for speech enhancement," in *Proc. of INTERSPEECH*, 2020.

[11] F. Deng, T. Jiang, X. Wang, C. Zhang, and Y. Li, "NAAGN: noise-aware attention-gated network for speech enhancement," in *Proc. of INTERSPEECH*, 2020.

[12] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using U-Net convolutional neural networks," in *Proc. of ICASSP*. IEEE, 2020, pp. 6214–6218.

[13] D. N. Tran and K. Koishida, "Single-channel speech enhancement by subspace affinity minimization," in *Proc. of INTERSPEECH*, 2020, pp. 2447–2451.

[14] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. of NeurIPS*, 2018, pp. 6391–6401.

[15] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 380–390, 2019.

[16] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. of INTERSPEECH*, 2020, pp. 2472–2476.

[17] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *Proc. of ICLR*, 2018.

[18] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," *arXiv preprint arXiv2010.15174*, 2020.

[19] S. Pascual, A. Bonafonte, and J. Serr'a, "SEGAN: speech enhancement generative adversarial network," in *Proc. of INTERSPEECH*, 2017, pp. 3642–3646.

[20] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: a phase-and-harmonics-aware speech enhancement network," in *Proc. of AAAI*, vol. 34, no. 05, 2020, pp. 9458–9465.

[21] A. Zolnay, R. Schluter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. of ICASSP*, vol. 1. IEEE, 2005, pp. I–457.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE TASLP*, vol. 19, no. 4, pp. 788–798, 2011.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. of NIPS*, 2013, pp. 3111–3119.

[24] Y.-A. Chung and J. Glass, "Speech2vec: a sequence-to-sequence framework for learning word embeddings from speech," in *Proc. of INTERSPEECH*, 2018, pp. 811–815.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv1807.03748*, 2018.

[26] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: unsupervised pre-training for speech recognition," in *Proc. of INTERSPEECH*, 2019, pp. 3465–3469.

[27] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-supervised representations improve end-to-end speech translation," in *Proc. of INTERSPEECH*, 2020, pp. 1491–1495.

[28] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec; self-supervised learning of discrete speech representations," in *Proc. of ICLR*, 2020.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of ACL*, 2019, pp. 4171–4186.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*. Springer, 2015, pp. 234–241.

[31] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet a generative model for raw audio," *arXiv preprint arXiv1609.03499*, 2016.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of ICLR*, 2015.

[33] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. of INTERSPEECH*, 2016, pp. 352–356.

[34] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech." in *Proc. of SSW*, 2016, pp. 146–152.

[35] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus design, collection and data analysis of a large regional accent speech database," in *Proc. of O-COCOSDACASLRE*. IEEE, 2013, pp. 1–4.

[36] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE TASLP*, vol. 26, no. 3, pp. 197–210, 1978.

[37] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of ICASSP*. IEEE, 2015, pp. 708–712.

[38] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. of INTERSPEECH*, 2020.

[39] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. of ICASSP*. IEEE, 2020, pp. 6649–6653.

[40] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. of ICLR*, 2020.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE TASLP*, vol. 16, no. 1, pp. 229–238, 2007.

[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.