INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Self-supervised Adversarial Multi-task Learning for Vocoder-based Monaural Speech Enhancement

Zhihao Du[1], *Ming Lei*[2], *Jiqing Han*[1], *Shiliang Zhang*[2]

[1]School of Computer Science and Technology, Harbin Institute of Technology
[2]Speech Lab, Alibaba DAMO Academy
{duzhihao,jqhan}@hit.edu.cn, {lm86501,sly.zsl}@alibaba-inc.com

Abstract

In our previous study, we introduce the neural vocoder into monaural speech enhancement, in which a flow-based generative vocoder is used to synthesize speech waveforms from the Mel power spectra enhanced by a denoising autoencoder. As a result, this vocoder-based enhancement method outperforms several state-of-the-art models on a speaker-dependent dataset. However, we find that there is a big gap between the enhancement performance on the trained and untrained noises. Therefore, in this paper, we propose the self-supervised adversarial multi-task learning (SAMLE) to improve the noise generalization ability. In addition, the speaker dependence is also evaluated for the vocoder-based methods, which is important for real-life applications. Experimental results show that the proposed SAMLE further improves the enhancement performance on both trained and untrained noises, resulting in a better noise generalization ability. Moreover, we find that vocoder-based enhancement methods can be speaker-independent through a large-scale training.

Index Terms: self-supervised, multi-task learning, vocoder-based speech enhancement

1. Introduction

Monaural speech enhancement aims at separating speeches from the noisy backgrounds by using a single microphone. Since it was formulated as a supervised learning problem, monaural speech enhancement has achieved a huge progress by using the deep learning techniques [1]. Early supervised methods only enhance the magnitude spectrum of noisy speeches but leave the noisy phase spectrum unchanged [2], such as the ideal ratio mask (IRM) [3] and spectral mapping [4]. Recent studies suggest that the phase spectrum is also important for perceptual quality and speech intelligibility [5]. Therefore, many studies are making efforts on enhancing the magnitude and phase spectra simultaneously. In the complex spectrum domain, the magnitude and phase spectra are enhanced by estimating the complex ratio mask (CRM) [6] or their clean counterparts [7, 8]. In the waveform domain, the WaveNet [9] and Wave-U-Net [10] are introduced and trained with the energy-conserving loss [11] and frequency-domain loss [12], respectively.

In general, it is more difficult to enhance the complex spectrum and waveform than the magnitude spectrum due to the lack of clear structures in them. Inspired by the progress in speech synthesis community, we introduce the neural vocoder to monaural speech enhancement in our previous study [13]. Specifically, a denoising autoencoder is trained to reconstruct the clean Mel power spectrum, and a flow-based generative vocoder, Flowavenet [14], is employed to synthesize the speech waveform from the enhanced features without using the noisy phase spectrum or predicting a clean one. As a result, vocoder-based enhancement methods bypass the problem of phase prediction and outperform several state-of-the-art models on a speaker-dependent dataset [13].

According to the results in our previous study, we find that there is a big gap between the enhancement performance on trained and untrained noises. Therefore, in this paper, we propose the self-supervised adversarial multi-task learning (SAMLE) to improve the noise generalization ability for vocoder-based enhancement methods. In addition, since vocoders are originally designed to synthesize waveforms for a specific speaker, previous vocoder-based enhancement methods are always trained and evaluated on a speaker-dependent dataset [13, 15, 16]. However, enhancement methods should be speaker independent in real-life applications, i.e. speakers for training and test are different. Therefore, we also evaluate the speaker dependence of flow-based generative vocoder (Flowavenet) [14] and autoregressive vocoder (WaveRNN) [17] for both the speech synthesis and enhancement tasks.

2. Vocoder-based enhancement method

The proposed method consists of a denoising autoencoder (DAE), an autoregressive vocoder and a noise classifier. As shown in Fig.1, DAE attempts to reconstruct the clean Mel power spectrum from the noisy complex spectrum, and the autoregressive vocoder is used to synthesize the speech waveform from the predicted Mel power spectrum. Meanwhile, a noise classifier is involved to perform the self-supervised adversarial multi-task learning (SAMLE). As in [13], we first train the DAE and vocoder separately, and then stack them together.

2.1. Denoising autoencoder

In DAE, the encoder is trained to learn an intermediate representation h from the complex spectrum of noisy speech, and the decoder aims at reconstructing the clean Mel power spectrum from h. The mean absolute error (MAE) is employed as the loss function, which is defined between the predicted Mel power spectrum $\hat{S}$ and the clean one S:

$$\mathcal{L}_{MAE}(S,\hat{S}) = \frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}\left|S(t,f)-\hat{S}(t,f)\right| \quad (1)$$

where T and F indicate the total numbers of time frames and frequency bins, respectively. $|\cdot|$ means the absolute value. An-

This work was performed while the first author was an intern at Speech Lab, Alibaba DAMO Academy. This research was supported by National Key Research and Development Program of China under Grant 2017YFB1002102, National Natural Science Foundation of China under Grant U1736210.

http://dx.doi.org/10.21437/Interspeech.2020-1496

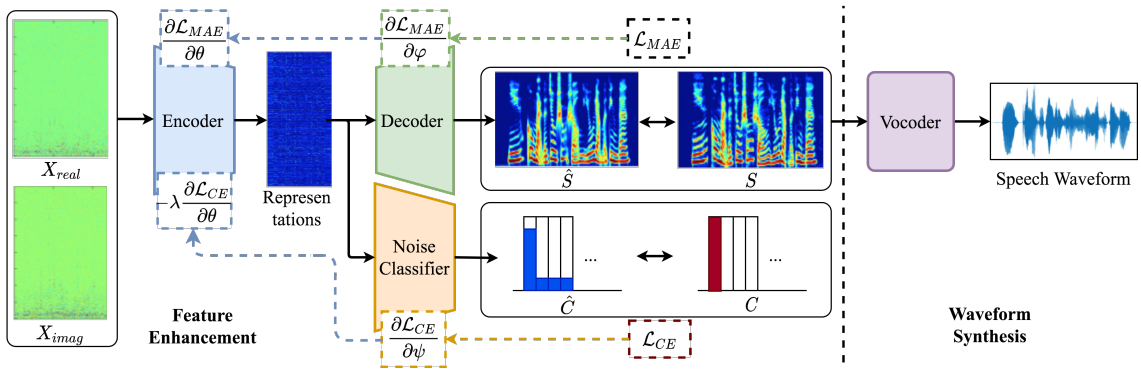


Figure 1: *(Color online) An overview of the proposed vocoder-based enhancement method. The solid and dashed arrows indicate the forward and backward passes, respectively.*

other optional choice here is mean square error (MSE), however, previous studies show that MAE leads to higher speech intelligibility and perceptual quality than MSE [18].

Different from our previous study [13], we feed the DAE with noisy complex spectra rather than the noisy Mel power spectra. The complex spectrum can benefit speech enhancement from two respects. First, it contains not only the magnitude information but also the phases. By using the phase features, such as group delay, the performance of speech recognition and speaker identification has been improved [19, 20]. Therefore, the complex spectrum may provide another cue (phase information) for speech enhancement. Second, the complex spectrum has higher frequency resolution than the Mel power spectrum, which provides more details for speech enhancement.

2.2. Self-supervised adversarial multi-task learning

2.2.1. Adversarial multi-task learning

Only trained with the reconstruction loss $\mathcal{L}_{MAE}$, DAE may suffer the performance degradation on untrained noises, because the intermediate representation h is not restricted, which may contain the information of background noises. To overcome this problem, an additional task, noise classification, is added to the DAE, resulting in a multi-task learning scheme. Specifically, a noise classifier C_ψ is involved, which is trained to distinguish the noise categories according to the intermediate representations by minimizing the cross entropy:

$$\mathcal{L}_{CE}(c, C_\psi(h)) = \sum_{k=1}^{K} c_{ik} \log C_{\psi_k}\left(h_i^{(1)}, h_i^{(2)}, ..., h_i^{(T)}\right) \tag{2}$$

where, K and T are the total numbers of noise categories and time frames. c_i is a one-hot code of the noise category for the noisy utterance i. $h_i^{(t)}$ means the intermediate representation of utterance i at frame t. While the noise classifier attempts to minimize the cross entropy loss $\mathcal{L}_{CE}$, the encoder tries to maximize it by adjusting the intermediate representation of each frame. Through the adversarial multi-task learning, the intermediate representation can be noise-invariant.

There are two ways to implement the adversarial training. One is adding a gradient reverse layer (GRL) [21] between the encoder and noise classifier, like domain adaption methods [22, 23]. The other way is updating the encoder and noise classifier iteratively, like the generative adversarial networks [24]. In our preliminary experiments, we find that the latter one obtains slightly better results. Therefore, we employ the second way in our following experiments.

2.2.2. Self-supervised noise classification

To perform the adversarial multi-task learning, we need to arrange the noise category to each noisy utterance. However, it is expensive and time-consuming to label thousands of noisy utterances manually. In addition, a noise can be arranged with several categories from different respects, which is another obstacle to manually noise labeling. Therefore, we propose a self-supervised noise classification method by developing an automatic noise labeling criterion. According to the energy distribution in the frequency domain, we divide noises into three categories, i.e. "low-frequency", "high-frequency" and "full-frequency" noises. Formally, the labeling criterion is defined as follows:

$$c := \begin{cases} 0 & P_l \geqslant P_a/2 \\ 1 & P_h \geqslant P_a/2 \\ 2 & otherwise \end{cases} \tag{3}$$

$$P_l := \sum_{t=1}^{T} \sum_{f=1}^{\lfloor \alpha F \rfloor} |N|^2(t,f) \tag{4}$$

$$P_h := \sum_{t=1}^{T} \sum_{f=\lfloor \beta F \rfloor}^{F} |N|^2(t,f) \tag{5}$$

$$P_a := \sum_{t=1}^{T} \sum_{f=1}^{F} |N|^2(t,f) \tag{6}$$

where $|N|^2$ is the power spectrum of a noise. P_l, P_h and P_a represent the total energy in the low, high and all frequency-bins, respectively. α and β are hyper-parameters to adjust the range of low and high frequency-bins ($0 < \alpha \leqslant \beta < 1$).

By applying the self-supervised noise classification to the adversarial multi-task learning, we obtain the self-supervised adversarial multi-task learning (SAMLE), and its optimizing target is given as follows:

$$\min_{\theta,\varphi} \max_{\psi} \mathcal{L}_{MAE}(S, D_\varphi(E_\theta(X))) - \lambda \mathcal{L}_{CE}(c, C_\psi(E_\theta(X))) \tag{7}$$

where λ is a hyper-parameter to balance the feature reconstruction and noise classification. Through SAMLE, the intermediate representations can be noise-invariant, which improves the noise generalization ability of DAE.

2.3. Mel power spectrum based WaveRNN

Another component of vocoder-based enhancement methods is the neural vocoder. In [13], we have evaluated a flow-based generative vocoder, Flowavenet, for speech enhancement. However, autoregressive vocoders can achieve higher synthesis performance than the flow-based models [14]. In this paper, we evaluate the autoregressive vocoder, WaveRNN [17], for speech enhancement purpose. Compared with the commonly-used Wavenet [9], WaveRNN has the advantages of fewer model parameters, faster synthesis speed and comparable quality [17].

In the original WaveRNN, speech waveforms are synthesized from the linguistic features and pitch information, which are predicted from the text. However, in speech enhancement, neither the linguistic features nor text are available. Therefore, we adjust the WaveRNN to synthesize the speech waveform from the Mel power spectrum, which is extracted from the clean speech and rescaled in the same manner as [13].

3. Experimental settings

3.1. Datasets

To train a speaker-independent vocoder, we need lots of speakers and enough durations for each speaker [25]. Therefore, we employ a large-scale internal corpus in our experiment, which is recorded by 100 males and 100 females. There are 500 utterances (about 34.25 minutes) for each speaker. Among these speakers, 90 males and 90 females are randomly selected to train the vocoder. Five males and five females are used for cross validation. The remaining 10 speakers are used for test.

For the training set of DAE, 930 noise recordings and six signal-to-noise rates (SNRs) are used to simulate the noisy mixtures. The noise recordings come from MUSAN [26], and the SNR levels are -5, -4, -3, -2, -1, 0 dB. For each SNR level and noise, four utterances are randomly selected from the clean training set and mixed with the noise recording, resulting in 22,320 mixtures (930 noises $\times$ 6 SNRs $\times$ 4 utterances). The validation set of DAE is built in the same manners as the training set, but only one utterance is randomly selected from the clean validation set for each SNR and noise, resulting in 5,580 noisy mixtures. In the test set, five noises (babble, factory, optroom, SSN and café) are selected from the NOISEX-92 [27] and DEMAND [28], which do not appear in the training set. The SNR levels are -5, 0, 5 dB. For each SNR level and noise, 10 utterances are randomly selected from the clean test set, resulting in 150 noisy mixtures.

3.2. Model settings

For the WaveRNN vocoder, we use an open source implementation [1], and the number of hidden units is set to 512. The rescaled Mel power spectrum is extracted with the hanning window, where the window length is 800 and the hop length is 200. There are 80 channels in the Mel filter bank, and the low frequency is set to 40 Hz. All audio files are sampled to 16 kHz.

The model architecture of DAE is shown in Table 1. The convolutional and deconvolutional layers are followed by the batch normalization [29] and exponential linear units (ELU) [30] except the output layer, which is followed by the sigmoid activation function only. The input of DAE, a complex spectrum, is extracted from the noisy speech with the same window settings of Mel power spectrum. Since the value range of complex spectra is too large, we compress the magnitude with cube root and keep the phase unchanged. The desired output of DAE is the rescaled Mel power spectrum of clean speech, which is also the input of WaveRNN. For the noise classifier, the intermediate representations of an utterance are averaged over the time frames first. Then, the average vector is fed to three stacked dense layers with 1024 units in each of them. While the hidden layers are followed by batch normalization and ReLU, the output layer is followed by the softmax activation.

[1] Available at https://github.com/fatchord/WaveRNN

Table 1: *The architecture of the enhancement model. Here T denotes the number of time frames in the acoustic features.*

layer name	input size	kernel, stride	output size
conv2d_0	$2 \times T \times 401$	$3 \times 3, (1, 1)$	$32 \times T \times 401$
conv2d_1	$32 \times T \times 401$	$3 \times 10, (1, 5)$	$32 \times T \times 80$
conv2d_2	$32 \times T \times 80$	$3 \times 3, (1, 1)$	$32 \times T \times 80$
conv2d_3	$32 \times T \times 80$	$3 \times 4, (1, 2)$	$64 \times T \times 40$
conv2d_4	$64 \times T \times 40$	$3 \times 4, (1, 2)$	$128 \times T \times 20$
conv2d_5	$128 \times T \times 20$	$3 \times 4, (1, 2)$	$256 \times T \times 10$
conv2d_6	$256 \times T \times 10$	$3 \times 4, (1, 2)$	$512 \times T \times 5$
conv2d_7	$512 \times T \times 5$	$3 \times 5, (1, 1)$	$1024 \times T \times 1$
reshape_1	$1024 \times T \times 1$	-	$T \times 1024$
blstm_1	$T \times 1024$	1024×1024	$T \times 1024 \times 2$
fc	$T \times 2048$	2048×1024	$T \times 1024$
reshape_2	$T \times 1024$	-	$1024 \times T \times 1$
deconv2d_7	$1024 \times T \times 1$	$3 \times 5, (1, 1)$	$512 \times T \times 5$
deconv2d_6	$512 \times T \times 5$	$3 \times 4, (1, 2)$	$256 \times T \times 10$
deconv2d_5	$256 \times T \times 10$	$3 \times 4, (1, 2)$	$128 \times T \times 20$
deconv2d_4	$128 \times T \times 20$	$3 \times 4, (1, 2)$	$64 \times T \times 40$
deconv2d_3	$64 \times T \times 40$	$3 \times 4, (1, 2)$	$32 \times T \times 80$
deconv2d_2	$32 \times T \times 80$	$3 \times 3, (1, 1)$	$32 \times T \times 80$
deconv2d_1	$32 \times T \times 80$	$3 \times 3, (1, 1)$	$32 \times T \times 80$
output	$32 \times T \times 80$	$3 \times 3, (1, 1)$	$1 \times T \times 80$
reshape_3	$1 \times T \times 80$	-	$T \times 80$

All models are optimized by using the Adam optimizer with the learning rate of 0.001, the batch size of 16 and the early stopping patience of 20. The best model is selected by cross validation. The hyper-parameters λ, α and β are set to 0.01, 0.125 and 0.33, respectively. This setting achieves better results in our preliminary experiments on hyper-parameter selection.

3.3. Compared methods

We compare our method with five recent models on a speaker-independent dataset. The first one is Flowavenet-SE, which is a vocoder-based method for speaker-dependent task in [13]. The second one is TimeMapping [12], which enhances the speech waveform in the time domain with the Wave-U-Net [31]. The third one is based on the convolutional recurrent network (CRN), which maps the noisy complex spectrum to its clean counterpart [32]. The fourth one is PHASEN [8], which comprises two streams, i.e. phase prediction stream and magnitude mapping stream. A Wavenet for speech denoising (Wavenet-SD) [11] is also compared, in which a non-causal Wavenet is trained by minimizing the energy-conserving loss.

4. Results

4.1. Method comparison

We employ the short-time objective intelligibility (STOI) [33] and perceptual evaluation of speech quality (PESQ) [34] as the evaluation metrics. The results of different models are shown in Table 2. The proposed method," Ours (Complex, SAMLE)", consistently outperforms the compared methods under all evaluated SNR levels in terms of PESQ. With respect to STOI, the

Table 2: *Comparisons of different models in STOI and PESQ metrics on untrained speakers and noises.*

Metrics	STOI (%)			PESQ		
SNR (dB)	-5	0	5	-5	0	5
Noisy	60.93	74.60	87.09	1.06	1.44	1.86
Flowavenet-SE [13]	66.54	80.51	87.69	1.50	2.03	2.48
TimeMapping [12]	73.48	87.52	**93.65**	1.78	2.36	2.74
CRN [32]	72.76	86.30	93.60	1.73	2.27	2.73
PHASEN [8]	70.17	85.71	93.57	1.61	2.24	2.73
Wavenet-SD [11]	71.62	86.19	93.11	1.56	2.12	2.51
Ours (Mel, SAMLE)	72.04	83.91	91.69	1.71	2.23	2.71
Ours (Complex, Manual)	**75.28**	87.87	93.31	1.85	2.43	**2.85**
Ours (Complex, SAMLE)	75.16	**87.97**	93.33	**1.86**	**2.44**	2.83

Table 3: *The speaker dependence of vocoder-based enhancement methods in terms of STOI and PESQ scores.*

Models	Trained Speakers		Untrained Speakers	
	STOI (%)	PESQ	STOI (%)	PESQ
Flowavenet-oracle [14]	92.51	3.17	92.52	3.13
WaveRNN-oracle	96.94	3.52	96.95	3.49
Noisy	72.47	1.41	74.17	1.47
Flowavenet+DAE [13]	77.42	1.97	78.25	2.00
WaveRNN+DAE	**84.98**	**2.39**	**85.49**	**2.38**

proposed method achieves better enhancement performance under low and middle SNR levels (-5, 0 dB). For the high SNR level (5 dB), our result is also good enough ($\geqslant 93\%$). By comparing with "Ours (Mel, SAMLE)", we find that the enhancement performance can be much improved by using the noisy complex spectrum as the input of DAE. This indicates that complex spectra can provide more cues to benefit the speech enhancement than Mel power spectra.

In addition, we also tried to use the handcrafted noise labels to perform the adversarial multi-task learning (seen in "Ours (Complex, Manual)"), in which the noisy mixtures are manually divided into 18 categories according to the acoustic scene of background noises. Compared with the handcrafted noise labels, our self-supervised noise classification achieves the comparable STOI and PESQ scores. This indicates that the energy distribution in the frequency domain is a good classification criterion for the adversarial multi-task learning of DAE.

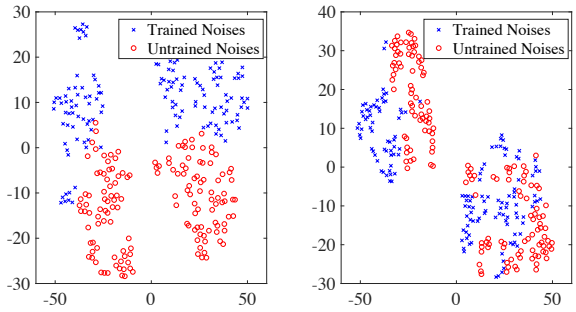
4.2. Speaker dependence

The speaker dependence of WaveRNN and Flowavenet is evaluated for the speech enhancement purpose by using the enhanced Mel power spectrum as an input. To obtain an upper bound of the vocoder-based enhancement methods, vocoders are also fed with the clean Mel power spectrum. The results on trained and untrained speakers are shown in Table 3, where "oracle" means that the vocoder is fed with clean features.

From the table, we can see that, for both Flowavenet and WaveRNN-based methods, the enhancement performance is similar on trained and untrained speakers in terms of the improvements on STOI and PESQ scores. This indicates that vocoder-based enhancement methods can be speaker independent through a large-scale training. In addition, the synthesis performance is also similar on trained and untrained speakers with the clean features as inputs. Compared with Flowavenet, WaveRNN achieves higher STOI and PESQ scores no matter the input is clean or enhanced features, which indicates that the autoregressive vocoder is a better choice for speech enhancement purpose than the flow-based generative vocoder.

Table 4: *The impact of SAMLE on trained and untrained noises.*

SNR	Noise Types	STOI Improvement (%)	PESQ Improvement
-5 dB	Trained	15.32 $\mapsto$ 15.63	0.93 $\mapsto$ 0.94
	Untrained	13.85 $\mapsto$ 14.23	0.76 $\mapsto$ 0.80
0 dB	Trained	12.16 $\mapsto$ 12.38	1.08 $\mapsto$ 1.10
	Untrained	12.80 $\mapsto$ 13.37	0.97 $\mapsto$ 1.00
5 dB	Trained	5.27 $\mapsto$ 5.63	0.97 $\mapsto$ 1.01
	Untrained	6.10 $\mapsto$ 6.24	0.96 $\mapsto$ 0.97



(a) Fully-supervised (b) SAMLE-based

Figure 2: *The intermediate representations learned by (a) fully-supervised and (b) SAMLE-based DAEs.*

4.3. The impact of SAMLE

Table 4 shows the impact of SAMLE on trained and untrained noises. We can see that the proposed SAMLE increases the STOI and PESQ improvements for both trained and untrained noises under all evaluated SNR levels. This indicates that SAMLE can improve the noise generalization ability of vocoder-based enhancement methods.

To have an insight of how SAMLE affects the feature reconstruction process, we compare the intermediate representations learned by the fully-supervised and SAMLE-based DAEs in Figure 2, where the fully-supervised DAE is trained with $\mathcal{L}_{MAE}$ only. For visualization, the dimensionality of learned representations is reduced from 1,024 to 2 by the t-SNE [35]. Only using $\mathcal{L}_{MAE}$, the learned intermediate representations on trained and untrained noises can be distinguished from each other obviously. On the contrary, by performing SAMLE, the representations on untrained noises are "embedded" to those of trained noises. In this manner, the vocoder-based enhancement models achieve better generalization ability on untrained noises.

5. Conclusions

In this paper, the self-supervised adversarial multi-task learning (SAMLE) is proposed for vocoder-based enhancement methods, which can improve the noise generalization ability on both trained and untrained noises. Furthermore, we evaluate the speaker dependence for vocoder-based enhancement methods, which is crucial in real-life applications. Through a large-scale training, the autoregressive vocoder, WaveRNN can be speaker independent and achieve a better performance than the flow-based generative vocoder with respect to synthesis and enhancement performance. By using the SAMLE and WaveRNN, our vocoder-based method outperforms several state-of-the-art models, which presents another optional way for speaker-independent monaural speech enhancement.

6. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[3] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind source separation*, 2014, pp. 349–368.

[4] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[5] K. K. Paliwal, K. K. Wójcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *ICASSP*, 2016, pp. 5220–5224.

[7] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*, 2019, pp. 6865–6869.

[8] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *AAAI*, 2020.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *SSW*, 2016, p. 125.

[10] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018, pp. 334–340.

[11] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP*, 2018, pp. 5069–5073.

[12] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *INTERSPEECH*, 2018, pp. 1136–1140.

[13] Z. Du, X. Zhang, and J. Han, "A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, DOI: 10.1109/TASLP.2020.2991537, 2020.

[14] S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon, "Flowavenet : A generative flow for raw audio," in *ICML*, vol. 97, 2019, pp. 3370–3378.

[15] S. Maiti, J. Ching, and M. I. Mandel, "Large vocabulary concatenative resynthesis," in *INTERSPEECH*, 2018, pp. 1190–1194.

[16] S. Maiti and M. I. Mandel, "Speech denoising by parametric resynthesis," in *ICASSP*, 2019, pp. 6995–6999.

[17] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *ICML*, vol. 80, 2018, pp. 2415–2424.

[18] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *ICASSP*, 2018, pp. 5414–5418.

[19] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[20] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "Dnn-based amplitude and phase feature enhancement for noise robust speaker identification," in *INTERSPEECH*, 2016, pp. 2204–2208.

[21] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, vol. 37, 2015, pp. 1180–1189.

[22] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[23] C. F. Liao, Y. Tsao, H. Y. Lee, and H. M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *INTERSPEECH*, 2019, pp. 3148–3152.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[25] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," in *INTERSPEECH*, 2019, pp. 181–185.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[27] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, 1993.

[28] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, 2013.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, vol. 37, 2015.

[30] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *ICLR*, 2016.

[31] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net," in *ISMIR*, 2018, pp. 5–9.

[32] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*, 2019, pp. 6865–6869.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] 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 *ICASSP*, 2001.

[35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.