SPEECH ENHANCEMENT WITH NEURAL HOMOMORPHIC SYNTHESIS

Wenbin Jiang, Zhijun Liu, Kai Yu*, Fei Wen*

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Most deep learning-based speech enhancement methods operate directly on time-frequency representations or learned features without making use of the model of speech production. This work proposes a new speech enhancement method based on neural homomorphic synthesis. The speech signal is firstly decomposed into excitation and vocal tract with complex cepstrum analysis. Then, two complex-valued neural networks are applied to estimate the target complex spectrum of the decomposed components. Finally, the time-domain speech signal is synthesized from the estimated excitation and vocal tract. Furthermore, we investigated numerous loss functions and found that the multi-resolution STFT loss, commonly used in the TTS vocoder, benefits speech enhancement. Experimental results demonstrate that the proposed method outperforms existing state-of-the-art complex-valued neural network-based methods in terms of both PESQ and eSTOI.

Index Terms— Speech enhancement, speech synthesis, source-filter model, complex neural network

1. INTRODUCTION

Monaural speech enhancement is one of the most challenging tasks in speech signal processing, of which the goal is to suppress the interfering noise in observed noisy speech to improve the speech perceptual quality [1]. As a fundamental module of speech application systems [2, 3, 4], it has been widely used in mobile telecommunication, automatic speech recognition, and hearing aids, etc.

Benefited from recent advances in deep learning, speech enhancement with supervised deep neural network (DNN) has attracted much attention and achieved much progress recently [5, 6, 7]. Generally speaking, existing DNN-based methods can be classified into two categories. The first transforms time domain speech signal into time-frequency (TF) representation via short-time Fourier transform (STFT) and feeds the TF spectrogram to a neural network, while the second directly feeds the time domain speech signal to a neural network. Typically, the time-domain approaches either adopt an encoder-decoder network architecture to directly learn a regression mapping from noisy waveform to the target speech [8, 9], or utilize an additional separator network between the encoder and decoder to model the temporal context [7].

The more popular TF-domain approaches utilize neural networks either to estimate the clean speech spectrogram (i.e., *mapping*) [10] or to estimate the mask of the target spectrogram (i.e., *masking*) [11, 12, 13]. For a long time, the TF-domain methods only estimate the speech magnitude spectrum, leaving the phase spectrum unprocessed [14]. More recent studies show that, building complex-valued neural network blocks [15] to deal with complex-valued spectrograms can achieve significant improvement in speech enhancement [16, 17, 18].

However, supervised speech enhancement methods, whether time-domain or TF-domain based ones, are prone to over-suppress the speech and under-suppress the noise, especially in low signal to noise ratio (SNR) cases. To overcome this limitation, a joint framework combining denoising autoencoder and text-to-speech (TTS) vocoder has been proposed for speech denoising in [19, 20]. Similarly, a parametric resynthesis method that combines acoustic feature prediction model and vocoder has been developed to synthesize clean speech from noisy observation in [21, 22]. Though the effectiveness of such methods has been well demonstrated, the computational complexity of these methods is too high to be applied in practice.

Motivated by our recent work on low computational complexity neural homomorphic vocoder [23], we propose a new speech enhancement method utilizing neural homomorphic synthesis. Firstly, the speech signal is decomposed into excitation and vocal tract in complex cepstrum domain. Then, a complex-valued neural network is applied to estimate the target complex spectrum of the decomposed components. Finally, time-domain speech signal is synthesized from the estimated excitation and vocal tract. To the best of our knowledge, integrating speech enhancement and the digital signal processing (DSP) based vocoder into an ensemble has not been studied yet. Our contributions are as follows: *1)* We propose a new speech enhancement method, which combines the advantages of DSP-based vocoder and complex-valued neural network based spectrum denoiser, yielding state-of-the-art performance in terms of PESQ and eSTOI. *2)* We investigate numerous loss functions and found that multi-resolution STFT loss, commonly used in TTS vocoder, benefits speech enhancement.

2. NEURAL HOMOMORPHIC SYNTHESIS

2.1. Source-filter model

The source-filter model is widely used in speech synthesis [24]. A simplified version of the source-filter model of speech production is illustrated in Fig. 1. The excitation signal $e[n]$ is assumed to be either voiced speech that generated by convo-

*corresponding author

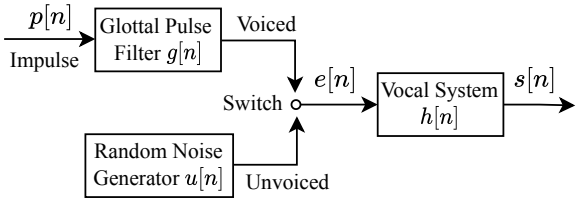

Fig. 1. A simplified source-filter model of speech production.

lution of a periodic impulse $p[n]$ and a glottal pulse filter $g[n]$, or unvoiced speech that generated by a random noise generator $u[n]$. The linear filter $h[n]$ is a combination of a vocal tract model and a radiation model in speech production, and here we call it the vocal system for short. Thus, the discrete-time speech signal is generated by a convolution of the excitation signal $e[n]$ and the vocal system $h[n]$.

2.2. Neural homomorphic synthesis for speech enhancement

Generally, TTS vocoder synthesizes speech using acoustic features as input, while speech enhancement uses reference noisy speech. A vocoder can produce high-quality speech from acoustic features (e.g., log magnitude spectrum and f0), but the quality of the synthesized speech drops dramatically when the acoustic features are degraded by noise. Speech enhancement algorithms can usually obtain moderate quality speech from noisy speech, even if the SNR is relatively low. Thus, integrating speech enhancement and vocoder into an ensemble can be expected to obtain high-quality and high-intelligibility speech.

Speech signal can be decomposed into excitation and vocal tract in the cepstrum domain, whether the speech is clean or degraded by noise. The left part in Fig. 2 illustrates the decomposition of time-domain speech into excitation and vocal tract via cepstral liftering. Firstly, speech waveform is segmented into frames with windows. Then, a complex cepstrum analysis pipe-line is applied, including Fourier transform, complex logarithm, and inverse Fourier transform. Finally, a lifter is applied to separate the excitation and vocal tract.

The right part in Fig. 2 illustrates the procedure of neural homomorphic synthesis. The decomposed excitation and vocal tract cepstrum are transformed into time-domain via complex cepstrum inverse pipe-line, respectively. The inverse pipe-line includes Fourier transform, complex exponential, neural network forward propagation, and inverse Fourier transform. The speech signal is obtained by time-domain circular convolution of the excitation and vocal tract, followed by post-processing, e.g., truncating and overlap-adding.

Most of the blocks in Fig. 2 are non-trainable DSP-based components, and only the neural networks contain trainable parameters. Given noisy-clean training pairs, noisy speech $\mathbf{x}$ and clean target speech $\mathbf{y}$, gradients are backward propagated along with the post-processing, time-domain circular convolution, and inverse Fourier transform. Details for estimating the excitation and vocal tract from noisy speech with neural networks are presented in the following section.

3. EXCITATION AND VOCAL TRACT ESTIMATION WITH COMPLEX NEURAL NETWORK

As illustrated in Fig. 2, the input and output of the neural network are both complex-valued data. Therefore, it is natural to use complex neural networks to process the complex-valued spectrum.

3.1. Complex neural network

The complex neural network is an extension of the real-valued neural network, all building blocks (e.g., convolution, activation, normalization, etc.) are extended to corresponding complex-valued blocks. Here, we take the complex convolution block as an example, and other blocks are similar.

A complex-valued Conv2d block consists two real-valued Conv2d blocks, and four real-valued Conv2d multiplication [15]. Let $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$ be the input complex vector, and $\mathbf{W} = \mathbf{W}_r + j\mathbf{W}_i$ be the weight matrix of the complex-valued Conv2d, where $\mathbf{W}_r$ and $\mathbf{W}_i$ are weight matrices of the two real-valued Conv2d blocks, respectively. Then, the complex-valued convolution is defined as

$$\mathbf{W} * \mathbf{x} = (\mathbf{W}_r * \mathbf{x}_r - \mathbf{W}_i * \mathbf{x}_i) + j(\mathbf{W}_i * \mathbf{x}_r + \mathbf{W}_r * \mathbf{x}_i). \quad (1)$$

Fig. 3 illustrates the structure of the deep complex neural network we used in the experimental section. The network follows an U-Net architecture [16] and applies complex recurrent network blocks for temporal modeling [6, 18]. For the sake of concise description, other building blocks (e.g., batch normalization, activation) are omitted in the figure. It should be noticed that the masking connection from the noisy input to the clean target output is optional. When the connection is applied, the output of the neural network is considered to be a mask $\hat{\mathbf{M}}_{t,f}$, and an additional bounding process [16] is applied to the output of neural network $\mathbf{O}_{t,f}$ as follows

$$\begin{aligned} \hat{\mathbf{M}}_{t,f} &= \left|\hat{\mathbf{M}}_{t,f}\right| \cdot e^{j\theta_{\hat{\mathbf{M}}_{t,f}}} = \hat{\mathbf{M}}^{mag}_{t,f} \cdot \hat{\mathbf{M}}^{phase}_{t,f} \\ \hat{\mathbf{M}}^{mag}_{t,f} &= tanh(|\mathbf{O}_{t,f}|),\ \hat{\mathbf{M}}^{phase}_{t,f} = \mathbf{O}_{t,f} / \left|\mathbf{O}_{t,f}\right|. \end{aligned} \quad (2)$$

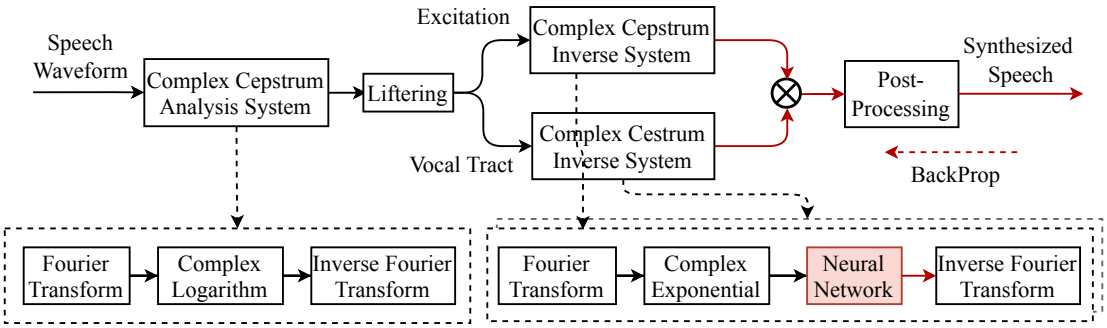

Fig. 2. (Color Online) Block diagram of the proposed method. Gradients are backward propagated along red lines.

The estimated spectrum $\hat{\mathbf{Y}}_{t,f}$ is computed by multiplying the estimated mask $\hat{\mathbf{M}}_{t,f}$ on the input spectrum $\mathbf{X}_{t,f}$, i.e., $\hat{\mathbf{Y}}_{t,f} = \hat{\mathbf{M}}_{t,f} \cdot \mathbf{X}_{t,f}$. Whereas, when the masking connection is not applied, the network directly output the estimated complex spectrum, i.e., $\hat{\mathbf{Y}}_{t,f} = \mathbf{O}_{t,f}$.

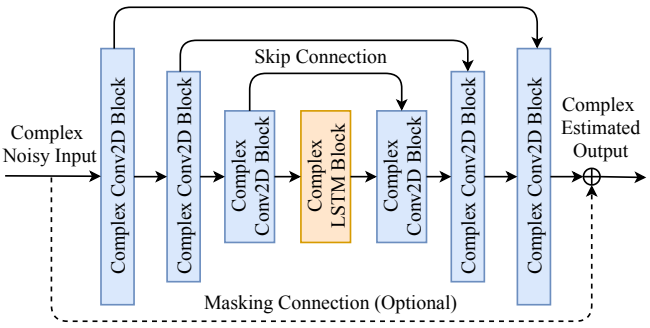

Fig. 3. Structure of deep complex neural network.

3.2. Loss functions

Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{n}$ denote noisy speech, clean speech, and additive noise, respectively, with which the corresponding signal model is $\mathbf{x} = \mathbf{y} + \mathbf{n}$. The goal of speech enhancement is to get an estimation $\hat{\mathbf{y}}$ of the clean speech $\mathbf{y}$, given the observed noisy speech $\mathbf{x}$. That is, finding a function f such that $\hat{\mathbf{y}} = f(\mathbf{x}) \approx \mathbf{y}$. We consider the following loss functions for neural network training.

3.2.1. Scale Invariant SNR loss

Scale-Invariant SNR (SI-SNR) loss [25] is shown to be more robust than the classical SNR and has been widely used as an evaluation metric for speech enhancement and speech separation, which is given by

$$L_{SI-SNR}(\mathbf{y}, \hat{\mathbf{y}}) = -10 \log 10 \left(\frac{\|\mathbf{y}_{\text{target}}\|_2^2}{\|\mathbf{e}_{\text{noise}}\|_2^2} \right) \quad (3)$$

where $\mathbf{y}_{\text{target}} = (\langle \mathbf{y}, \hat{\mathbf{y}} \rangle \cdot \mathbf{y}) / \|\mathbf{y}\|_2^2$ is the projection of the estimated speech $\hat{\mathbf{y}}$ onto clean speech $\mathbf{y}$, and $\mathbf{e}_{\text{noise}} = \mathbf{y} - \hat{\mathbf{y}}$ is the estimated noise.

3.2.2. Weighted SDR loss

Weighted SDR (wSDR) loss uses a normalization term to bound the loss function within the range [-1, 1] and an additional noise prediction term to complement the noise errors [16]. The two terms are balanced with an energy ratio as

$$L_{wSDR}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) = \alpha L_{SDR}(\mathbf{y}, \hat{\mathbf{y}}) + (1-\alpha) L_{SDR}(\mathbf{n}, \hat{\mathbf{n}}) \quad (4)$$

where $L_{SDR}(\mathbf{y}, \hat{\mathbf{y}}) = -\langle \mathbf{y}, \hat{\mathbf{y}} \rangle / (\|\mathbf{y}\| \|\hat{\mathbf{y}}\|)$ is the bounded SDR loss, $\alpha = \|\mathbf{y}\|^2 / \left(\|\mathbf{y}\|^2 + \|\mathbf{n}\|^2\right)$ is the energy ratio, and $\hat{\mathbf{n}} = \mathbf{x} - \hat{\mathbf{y}}$ is the estimated noise.

3.2.3. Multi-resolution STFT loss

Multi-resolution STFT (MR-STFT) loss is commonly used for speech and audio synthesis [26, 23]. Let $\mathbf{Y}_i$ and $\hat{\mathbf{Y}}_i$ be the ith resolution STFT of the time-domain speech $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively. The STFT loss of ith resolution is defined as

$$L_{stft}^{(i)}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\| |\mathbf{Y}_i| - |\hat{\mathbf{Y}}_i| \|_F}{\| |\mathbf{Y}_i| \|_F} + \| \log |\mathbf{Y}_i| - \log |\hat{\mathbf{Y}}_i| \|_1 \quad (5)$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ are the Frobenius and L_1 norm, respectively.

Our preliminary experiment results show that adding L_1 loss of the time-domain speech can further improve the speech quality. This trick has been also applied in the waveform domain speech enhancement [9]. Overall, the multi-resolution STFT loss is defined as

$$L_{MR-STFT} = -\left(\|\mathbf{y} - \hat{\mathbf{y}}\|_1 + \sum_{i=1}^{I} L_{stft}^{(i)}(\mathbf{y}, \hat{\mathbf{y}}) \right) \quad (6)$$

where I is the number of STFT resolutions. For the MR-STFT configuration, we use Hanning window with sizes (256, 512, 768, 1024, 1536, 2048, 3072, 4096) with 75% overlap, and the FFT sizes are set to twice the window sizes.

4. EXPERIMENTS

4.1. Datasets

We conduct experiments using the Chinese Standard Mandarin Speech Corpus (CSMSC)[1] and the DEMAND [27] noise corpus. CSMSC contains 10000 recorded sentences read by a female speaker, totaling about 12 hours of high-quality speech. DEMAND contains six categories and three environments for each category, totally 18 types of noise. Both CSMSC and DEMAND are originally sampled at 48 kHz, which are down-sampled to 16 kHz in our experiment.

Four noise categories and one environment of each are selected as the training, validation, and seen-noise test set, i.e., DKITCHEN, OMEETING, PCAFETER, TBUS. Another two noise categories, Nature and Street, are selected as unseen test data to evaluate the noise generalization. The SNR levels for mixing the noisy speech are randomly sampled from a uniform distribution [-5dB, 10dB]. The noise segment is also randomly sampled from the whole noise data of each selected noise type. Thus, we get 100,000 (10000 utterances $\times$ 10 noises) mixtures. For the seen-noise case, 9000, 500, and 500 utterances are selected as training, validation, and test sets, respectively. For the unseen nosie case, 500 utterances of each are selected as the test set. In total, there are 36000 training mixtures with a total duration of about 43 hours, 2000 validation mixtures, 2000 seen-noise test mixtures, and 3000 unseen-noise test mixtures.

4.2. Setups

All the utterances are framed by a hamming window with a length of 32 ms and a hop size of 8 ms, and the FFT length is 512. The quefrency for separating the excitation and vocal tract is 29. The algorithms are implemented with the help of PyTorch Lightning tools[2], and all the experiments are carried out on a High Performance Computing (HPC) center.

We use the following state-of-the-art complex DNN based speech enhancement models for comparison. Deep complex neural network with masking connection and different losses are considered, including SI-SNR loss, wSDR loss, and MR-STFT loss. The network architectures follow the setting of the DCCRN method [18], which ranked first in the Interspeech 2020 deep noise suppression challenge. The number

[1] https://www.data-baker.com/en/#/data/index/source
[2] https://github.com/PyTorchLightning/pytorch-lightning

Table 1. PESQ and eSTOI scores of the compared methods for seen noise types and unseen noise categories.

	Metrics	KITC	MEET	CAFE	BUS	Seen	Nature	Street	Unseen	Overall
Noisy	PESQ	1.345	1.117	1.098	1.821	1.345	1.211	1.180	1.196	1.270
	eSTOI	0.925	0.721	0.682	0.930	0.815	0.812	0.793	0.803	0.809
OMLSA	PESQ	2.181	1.195	1.256	2.577	1.802	1.757	1.591	1.674	1.738
	eSTOI	0.936	0.732	0.721	0.947	0.834	0.849	0.835	0.842	0.838
DCCRN-SI-SNR	PESQ	2.993	2.370	2.163	3.199	2.681	2.397	2.430	2.413	2.547
	eSTOI	0.971	0.930	0.906	0.975	0.946	0.924	0.930	0.927	0.936
DCCRN-wSDR	PESQ	2.827	2.325	2.142	2.965	2.565	2.297	2.339	2.318	2.441
	eSTOI	0.966	0.927	0.906	0.969	0.942	0.920	0.926	0.923	0.932
DCCRN-MR-STFT	PESQ	3.270	2.475	2.280	3.229	2.814	2.495	2.518	2.507	2.660
	eSTOI	0.970	0.932	0.914	0.972	0.947	0.928	0.933	0.931	0.939
NHS-SE	PESQ	3.444	2.850	2.598	3.582	**3.119**	2.708	2.859	**2.783**	**2.951**
	eSTOI	0.969	0.943	0.925	0.974	**0.953**	0.933	0.944	**0.939**	**0.946**

of channel, kernel size and stride are set to {16, 32, 64, 128, 128, 128}, (5,2) and (2,1), respectively. We denote the three models as DCCRN-SI-SNR, DCCRN-wSDR, DCCRN-MR-STFT, respectively. A traditional minimum mean squared error (MMSE) based speech enhancement method [28], denoted as OMLSA, is also considered for comparison. The proposed Neural Homomorphic Synthesis based Speech Enhancement method, denoted by NHS-SE, uses two complex neural networks, one modeling the excitation and another modeling the vocal tract. The MR-STFT loss is adopted in NHS-SE. In order to alleviate cepstrum aliasing, the FFT size is increased to 2048. Accordingly, the strides of the first two layers in the encoder and the last two layers in the decoder are increased to (4, 1).

The weights of all layers are initialized by a Xavier initializer [29]. Batch normalization for all layers except the output layer is applied. An Adam optimizer [30] with a fixed learning rate set to 0.001 is adopted to optimize the models. The learning rate is reduced by a factor 0.5 of when the validation error has stopped decrease within successive seven epochs. The training process is stopped when the validation error has no longer decreased in successive 30 epochs or reached the maximum of 200 epochs.

4.3. Results and analysis

We use speech quality (PESQ) [31] and extended short-time objective intelligibility (eSTOI) [32] to evaluate the quality of the enhanced speech. Audio samples are provided online[3].

The results are shown in Table 1, in which the best results in each case are highlighted by boldface. The third to sixth collumns are the results of seen-nosie types, and the seventh column is the corresponding mean score. Each categories of the unseen noise (i.e., Nature and Street) contains three noise types, and the results are averaged in the eighth and ninth columns, respectively. The mean scores for the unseen noise are listed in the penultimate column, and the overall mean scores are listed in last column.

The results demonstrate that, *1)* the SI-SNR loss and wSDR loss yield comparable results in term of PESQ and eSTOI; *2)* the MR-STFT loss outperforms the SI-SNR and wSDR losses in terms of PESQ; *3)* the proposed NHS-SE obtains the highest score in terms of PESQ and eSTOI metrics. For the unseen noise types, the performance of all methods is degraded to some extent. However, the proposed NHS-SE still outperforms all the others.

In addition to the results provided in the Table 1, we also conducted extra experiments using the proposed NHS-SE model with the SI-SNR and wSDR losses. Yet, the both models failed to converge. We also notice that the DCCRN-SI-SNR, in which the training loss matches the evaluation metric, obtains the highest SI-SNR score, while the proposed method failed to improve the SI-SNR score. This is due to all the synthesis-based method will cause phase mismatch [26, 20, 19, 21, 22], which results in very low SNR. Such methods usually only use perception scores (e.g., PESQ and eSTOI) as the evaluation metrics.

The proposed NHS-SE is more efficient than the cascaded systems [19, 21], in which a front-end denoising network and a back-end TTS vocoder are applied in combination. The number of model parameters of NHS-SE is about 7.4M, while that of a typical TTS vocoder is more than 20 M (e.g., WaveNet[33] is 21.56 M, WaveGlow[34] is 87.88 M). Furthermore, with the front-end denoising network, the model parameters of the cascaded systems will increase additionally.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a novel speech enhancement method based on homomorphic analysis and synthesis. It makes use of the advantages of the classical vocoder method and the recently popular speech enhancement method based on complex-valued neural networks. We conducted extensive experiments to investigate the popular SI-SNR, weighted SDR, and multi-resolution STFT losses. The results demonstrated that the multi-resolution STFT loss performs better than the others in terms of PESQ and eSTOI. Using the multi-resolution STFT loss, the proposed method outperforms state-of-the-art methods on both seen noise and unseen noise. Future works include removing some artificial noise introduced by the synthesis procedure and studying speaker generalization of the models.

[3] https://jiang-wenbin.github.io/NHS-SE/

6. REFERENCES

[1] Philipos C Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2013.

[2] Wenbin Jiang, Fei Wen, and Peilin Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," *IEEE Access*, vol. 6, pp. 52385–52392, 2018.

[3] Wangyou Zhang, Christoph Boeddeker, Shinji Watanabe, et al., "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *Proc. IEEE ICASSP*. IEEE, 2021, pp. 6898–6902.

[4] Chenda Li, Jing Shi, Wangyou Zhang, et al., "Espnet-se: end-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT*. IEEE, 2021, pp. 785–792.

[5] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. ASLP.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] Ke Tan and DeLiang Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. ISCA Interspeech*, 2018, pp. 3229–3233.

[7] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. ASLP.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[8] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. ISCA Interspeech*, 2017, pp. 3642–3646.

[9] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. ISCA Interspeech*, 2020, pp. 3291–3295.

[10] Yong Xu, Jun Du, Li-Rong Dai, et al., "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. ASLP.*, vol. 23, no. 1, pp. 7–19, 2014.

[11] DeLiang Wang, Ulrik Kjems, Michael S Pedersen, et al., "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[12] Hakan Erdogan, John R Hershey, Shinji Watanabe, et al., "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE ICASSP*. IEEE, 2015, pp. 708–712.

[13] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. ASLP.*, vol. 24, no. 3, pp. 483–492, 2015.

[14] Dequan Wang and Jae Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Audio, Speech, anguage Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[15] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, et al., "Deep complex networks," in *International Conference on Learning Representations*, 2018.

[16] Hyeong-Seok Choi, Janghyun Kim, Jaesung Huh, et al., "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2019.

[17] Ke Tan and DeLiang Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE ICASSP*. IEEE, 2019, pp. 6865–6869.

[18] Yanxin Hu, Yun Liu, Shubo Lv, et al., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. ISCA Interspeech*, 2020, pp. 2472–2476.

[19] Zhihao Du, Xueliang Zhang, and Jiqing Han, "A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1493–1505, 2020.

[20] Zhihao Du, Ming Lei, Jiqing Han, et al., "Self-Supervised Adversarial Multi-Task Learning for Vocoder-Based Monaural Speech Enhancement," in *Proc. ISCA Interspeech*, 2020, pp. 3271–3275.

[21] Soumi Maiti and Michael I Mandel, "Speech denoising by parametric resynthesis," in *Proc. IEEE ICASSP*. IEEE, 2019, pp. 6995–6999.

[22] Soumi Maiti and Michael I Mandel, "Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement," in *Proc. IEEE ICASSP*. IEEE, 2020, pp. 206–210.

[23] Zhijun Liu, Kuan Chen, and Kai Yu, "Neural Homomorphic Vocoder," in *Proc. ISCA Interspeech*, 2020, pp. 240–244.

[24] Lawrence Rabiner and Ronald Schafer, *Theory and applications of digital speech processing*, Prentice Hall Press, 2010.

[25] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, et al., "Sdr–half-baked or well done?," in *Proc. IEEE ICASSP*. IEEE, 2019, pp. 626–630.

[26] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, et al., "Ddsp: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020.

[27] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19, p. 035081.

[28] Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Audio, Speech, anguage Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[29] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.* JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[30] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[31] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[32] Jesper Jensen and Cees H Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. ASLP.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[33] Jonathan Shen, Ruoming Pang, Ron J Weiss, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE ICASSP*. IEEE, 2018, pp. 4779–4783.

[34] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE ICASSP*. IEEE, 2019, pp. 3617–3621.