Vector-quantized Variational Autoencoder for Phase-aware Speech Enhancement

Tuan Vu Ho† *Quoc Huy Nguyen*† *Masato Akagi*† *Masashi Unoki*†

† Japan Advanced Institute of Science and Technology, Japan

{tuanvu.ho, hqnguyen, akagi, unoki}@jaist.ac.jp

Abstract

Recent speech enhancement methods based on the complex ideal ratio mask (cIRM) have achieved promising results. These methods often deploy a deep neural network to jointly estimate the real and imaginary components of the cIRM defined in the complex domain. However, the unbounded property of cIRM poses difficulties when it comes to effectively training a neural network. To alleviate this problem, this paper proposes a phase-aware speech enhancement method by estimating the magnitude and phase of a complex adaptive Wiener filter. In this method, a noise-robust vector-quantized variational autoencoder is utilized for estimating the magnitude Wiener filter by using the Itakura-Saito divergence on time-frequency domain, while the phase of the Wiener filter is estimated by a convolutional recurrent network using the scale-invariant signal-to-noise ratio constraint in the time domain. The proposed method was evaluated on the open Voice Bank+DEMAND dataset to provide a direct comparison with other speech enhancement studies and achieved the PESQ score of 2.85 and STOI score of 0.94, which is better than the state-of-art method based on cIRM estimation in the 2020 Deep Noise Challenge.

Index Terms: Speech enhancement, vector-quantized variational autoencoder, complex Wiener filter, noise reduction

1. Introduction

There have been various techniques proposed for speech enhancement to improve speech intelligibility and quality under the effects of noise or reverberation. From the traditional concepts such as spectral subtraction [1], ideal binary mask [2], and minimum mean squared error estimation [3], which can only handle stationary noise, the enhancement methods have evolved to deal with various other types of noise by incorporating deep neural networks.

There are many processing domains in these methods, such as short-time Fourier transform (STFT), in which a complex spectrogram is the output. In general, the output features can be decomposed into magnitude and phase features. Most of the initial methods focused on the enhancement of the amplitude features only [4]. After clarifying the importance of phase in speech quality and intelligibility [5], several studies developed phase-aware enhancement techniques [6, 7, 8, 9], the most successful of which were based on the concept of the complex ideal ratio mask (cIRM) [10]. However, the unbounded property of cIRM makes it difficult for optimization due to the infinite search space [11].

To cope with this issue, we direct our study toward an approach based on the complex adaptive Wiener filter. The Wiener filter is a popular technique that has been applied in many signal enhancement methods. Since the range of Wiener filter is naturally bounded, estimating the Wiener filter should take less effort than estimating cIRM.

In this paper, we propose a phase-aware speech enhancement method that is effective even in an unknown environment through the estimation of a complex Wiener filter. A complex Wiener filter can be constructed using 3 parameters: the speech variance, the noise variance and phase. A vector-quantized variational autoencoder (~~VQVAE~~ [2]) is used to capture the distribution of speech variance by means of a discrete latent space represented by a codebook. Thanks to this discrete latent space, VQ-VAE can sufficiently model the distribution of high-quality speech variance parameters without any unintelligible variation as occurring in the vanilla VAE. Furthermore, the encoder network of the VQ-VAE model is optimized with a noise-robust training strategy, which aims to minimize the variation of the latent variables due to the presence of noise in input speech. The noise variance is estimated by a feed-forward convolutional network conditioned on the estimated speech variance. A convolutional recurrent network is used to estimate the phase of the complex Wiener by maximizing the scale-invariant signal-to-noise ratio (SI-SNR) in time-domain.

In Section 2 of this paper, we introduce the speech model upon which our proposed method is based. We then propose the speech enhancement method based on VQ-VAE in Section 3 and discuss the experimental setup and results in Section 4. We conclude in Section 5 with a brief summary.

2. Complex Wiener filter

In the ~~short-term Fourier transform (~~STFT~~)~~ domain, let the noisy complex spectrogram $\mathbf{X} \in \mathbb{C}^{F \times T}$ be the sum of clean speech complex spectrogram $\mathbf{S}$ and noise complex spectrogram $\mathbf{N}$. Let x_{ft}, s_{ft}, and n_{ft} represent the complex coefficients of $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{N}$, respectively, as

$$x_{ft} = s_{ft} + n_{ft}, \tag{1}$$

where F is the number of frequency bins, T is the number of frames, $f \in [0, F)$ is the frequency bin index, and $t \in [0, T)$ is the frame index. We assume the complex coefficients of the speech and noise spectrogram follow the circularly symmetric complex normal distribution, i.e, $s_{ft} \sim \mathcal{N}_{\mathcal{C}}(0, \sigma^2_{s,ft})$ and $n_{ft} \sim \mathcal{N}_{\mathcal{C}}(0, \sigma^2_{n,ft})$, where $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represents the variances of speech and noise, respectively. Since the speech and noise are uncorrelated, the noisy signal then follows the complex normal distribution as $x \sim \mathcal{N}_{\mathcal{C}}(0, \sigma^2_s + \sigma^2_n)$. By applying the Wiener filter, the power spectral density of clean speech can be estimated from the mixture signal as

$$\|\hat{s}_{ft}\|^2 = \|x_{ft}\|^2 \frac{\sigma^2_{s,ft}}{\sigma^2_{s,ft} + \sigma^2_{n,ft}}, \tag{2}$$

where $\|\hat{s}_{ft}\|$ is the predicted magnitude spectrum of clean speech. To estimate the complex speech spectrum $\hat{s}_{ft}$, the phase

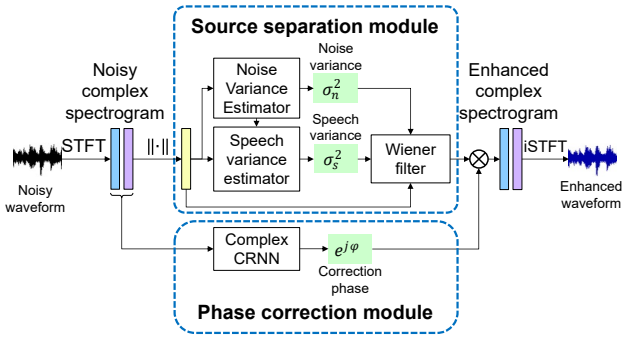

Figure 1: *Overview of proposed method.*

$e^{j\varphi_{ft}}$ of the complex Wiener filter can be introduced to E as:

$$\hat{s}_{ft} = x_{ft}\sqrt{\frac{\sigma^2_{s,ft}}{\sigma^2_{s,ft}+\sigma^2_{n,ft}}}e^{j\varphi_{ft}} \tag{3}$$

3. Proposed ~~Speech Enhancement Model~~

In this section, we describe our proposed method estimating the parameters of the complex Wiener filter, which are: speech variance $\sigma^2_{s,ft}$, noise variance $\sigma^2_{n,ft}$, and phase $e^{\varphi ft}$. For estimating the speech variance from the noisy mixture, we utilize a VQ-VAE model that is pre-trained on a clean dataset. The noise variance and the phase of the clean speech are then estimated on the basis of the predicted speech variance. The overview of our proposed model is shown in Fig. 1. Its training process consists of two phases: the VQ-VAE pretraining on clean speech and the training on noisy speech. The obtained codebook after the pretraining step is utilized to capture the characteristics of the clean speech. In the main training phase, the whole model is trained on the noisy speech mixture except for the latent codebook. The overview of the training flow is shown in Fig. 2.

3.1. Noise-robust vector-quantized variational autoencoder

3.1.1. Vector-quantized variational autoencoder

The ~~vector-quantized variational autoencoder (~~VQ-VAE~~)~~ is a generative model that consists of an encoder and a decoder network. The VQ-VAE basically resembles a communication system, where the encoder compacts the input feature vector into a continuous latent vector $\mathbf{z}$ by means of a non-linear transformation. The continuous latent vector $\mathbf{z}$ is then quantized to discrete variable $\mathbf{q}$ based on its distance to the pseudo-vectors in the codebook $\mathbf{e}_k, k \in 1...K$.

$$\mathbf{q} = \mathbf{e}_k, \text{ where } k = \arg\min_k \|\mathbf{z} - \mathbf{e}_j\| \tag{4}$$

Finally, the decoder outputs the estimated speech variance $\hat{\sigma}_{s,ft}$ by minimizing the Itakura-Saito (IS) divergence [12] between $\|s_{ft}\|^2$ and $\hat{\sigma}_{s,ft}$. The latent codebook is updated simultaneously with other parameters of the model during training process. Due to the use of $\arg\min$ function in quantization process, the computation graph is disconnected, and the model cannot be trained with back-propagation. Therefore, straight-through reparameterization trick [13] is used to avoid this problem:

$$\mathbf{z}_t = \text{Enc}(\mathbf{x}_t), \tag{5}$$
$$\mathbf{q}_t = \text{Quantize}(\mathbf{z}_t), \tag{6}$$
$$\hat{\mathbf{q}}_t = \mathbf{z}_t + \text{sg}(\mathbf{q}_t) - \mathbf{z}_t), \tag{7}$$
$$\hat{\sigma}_{s,t} = \text{Dec}(\hat{\mathbf{q}}_t), \tag{8}$$

where $\hat{\sigma}_{s,t}$ is the estimated speech variance vector, $\hat{\mathbf{q}}_t$ is straight-through variable from which gradient is copied to $\mathbf{z}_t$, $\text{Enc}(\cdot)$ is the encoder function, $\text{Dec}(\cdot)$ is the decoder function, $\text{Quantize}(\cdot)$ is quantization function, and $\text{sg}(\cdot)$ is the stop-gradient operator. The model parameters are obtained by minimizing the following objective function:

$$\mathcal{L}_{\text{vq}} = \text{d}_{\text{IS}}(\mathbf{s}_t, \hat{\sigma}^2_{s,t}) + \|\text{sg}(\mathbf{z}_t) - \mathbf{q}_t)\|_2^2 + \beta\|\mathbf{z}_t - \text{sg}(\mathbf{q}_t)\|\|_2^2, \tag{9}$$

where $\|\text{sg}(\mathbf{z}_t) - \mathbf{q}_t\|_2^2$ is the quantization loss, $\|\mathbf{z}_t - \text{sg}(\mathbf{q}_t)\|_2^2$ is the commitment loss, β is a hyper-parameter to control the weight of commitment loss, and $\text{d}_{\text{IS}}(\cdot,\cdot)$ is the IS divergence defined as

$$\text{d}_{\text{IS}}(\mathbf{s}_t^2, \sigma_x^2) = \sum_f \left(\frac{s^2_{f,t}}{\sigma^2_{s,ft}} - \ln\frac{s^2_{ft}}{\sigma^2_{s,ft}} - 1\right). \tag{10}$$

3.1.2. Method for achieving noise-robustness

The key point of VQ-VAE for speech enhancement is the noise-robustness property, with which the model can accurately estimate the speech variance from the noisy speech input. The most straightforward approach for achieving noise-robustness is to directly train the model to estimate speech variance from noisy speech. However, we observed that a VQ-VAE model trained with a noisy mixture from the beginning has a very low latent perplexity, which means that fewer spectrogram patterns are encoded in the latent codebook. Due to the low latent perplexity, the decoder cannot accurately estimate the speech variance even with clean speech input. In contrast, a VQ-VAE model trained on clean speech can achieve higher latent perplexity and lower reconstruction loss. On the basis of this observation, we propose pretraining the VQ-VAE on clean speech first. In other words, we set $\mathbf{x}_t = \mathbf{s}_t$ in the pretraining phase. Then, except for the latent codebook, the parameters of the encoder and decoder are fine-tuned on the noisy mixture to achieve the noise-robustness. Moreover, we propose the training objective with noise-robust commitment loss defined as follows:

$$\mathcal{L}_{\text{vq}} = \text{d}_{\text{IS}}(\mathbf{s}_t, \hat{\sigma}^2_{s,t}) + \beta\|\hat{\mathbf{z}}_t - sg(\mathbf{z}_t)\|\|_2^2, \tag{11}$$
$$\mathbf{z}_t = \text{Quantize}(\text{Enc}(\mathbf{s}_t)), \tag{12}$$
$$\hat{\mathbf{z}}_t = \text{Enc}(\mathbf{x}_t). \tag{13}$$

Note that the quantization loss is omitted in the main training phase so as to preserve the pre-trained latent codebook.

3.2. Noise variance estimator

The noise variance σ_n needs to be estimated for the Wiener filter. The noise variance estimator is trained to reduce the IS divergence between predicted noise variance σ_n and the noise log-power spectrogram $\mathbf{n}_{ft}$ as

$$\mathcal{L}_{\text{noise}} = \text{d}_{\text{IS}}(\mathbf{n}_t, \sigma^2_{n,t}) = \sum_f \left(\frac{n^2_{f,t}}{\sigma^2_{n,ft}} - \ln\frac{n^2_{ft}}{\sigma^2_{n,ft}} - 1\right). \tag{14}$$

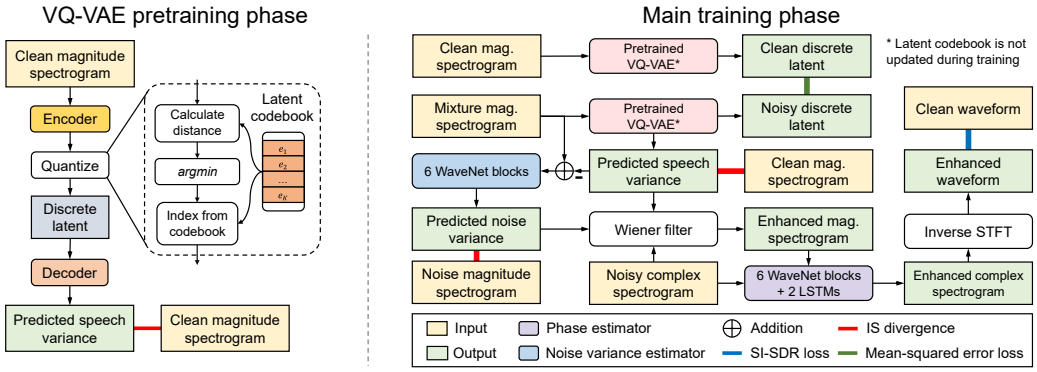

Figure 2: *Block diagram of the proposed method. The blocks in the pretraining phase corresponds to the red block in the main training phase.*

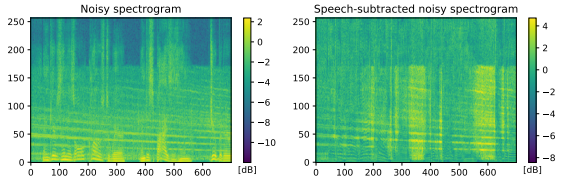

Figure 3: *Left: the noisy speech log-power spectrogram. Right: Noisy speech log-power spectrogram subtracted with estimated speech log-variance that resemble the the noise log-power spectrogram.*

To condition the noise variance estimator on the estimated speech variance σ_n^2, we empirically subtract the noisy speech log-power spectrogram with the estimated speech log-variance. Although it is not entirely accurate, this results to a representation which better resemble the noise log-power spectrogram as shown in Fig.3.

3.3. Phase estimator

Direct phase estimation is difficult due to the phase warping property. To overcome this problem, several studies have proposed using the SI-SNR as the objective function. From the estimated phase $e^{j\varphi ft}$, the clean speech complex spectrum $\hat{s}_{ft}$ can be derived using Eq. Then the speech waveform $\hat{\mathbf{y}}$ is derived using inverse STFT. The phase estimator is trained to maximize the SI-SNR defined as

$$\begin{aligned} \mathbf{y}_{target} &= \frac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle \cdot \mathbf{y}}{\|\mathbf{y}\|_2^2}, \\ \mathbf{e}_{noise} &= \hat{\mathbf{y}} - \mathbf{y}_{target}, \\ \text{SI-SNR} &= 10 \log_{10} \frac{\|\mathbf{y}_{target}\|_2^2}{\|\mathbf{e}_{noise}\|_2^2}, \end{aligned} \tag{15}$$

where $\hat{\mathbf{y}}$ and $\mathbf{y}$ is the enhanced and clean speech waveforms, respectively, $\langle \cdot, \cdot \rangle$ denotes the dot-product between two vectors, and $\|\cdot\|$ is the Euclidean norm of the vector.

In summary, the total loss to train the model is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{noise}} - \text{SI-SNR} \tag{16}$$

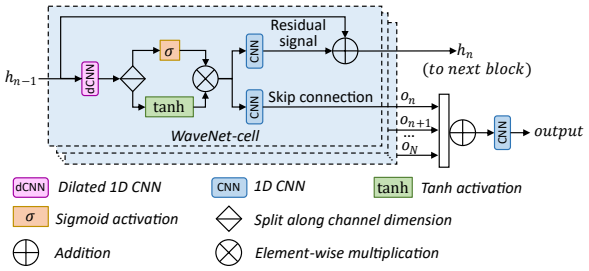

Figure 4: *Architecture of WaveNet block.*

4. Experimental Evaluation

4.1. Dataset

We use the open dataset released by Valentini et al. [16] to evaluate our proposed method. This dataset has been used in several recent speech enhancement studies, which we choose to include here as baselines. The clean train set is composed of 28 speakers and the test set of two speakers from the Voice Bank dataset. The noisy train set is constructed by mixing the clean train set with ten types of noise data from the DEMAND dataset at four SNRs: 15 dB, 10 dB, 5 dB, and 0 dB. The noisy test set is constructed by mixing the clean test set with five other types of noises at four SNRs: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. All speech waveforms are resampled from 48 kHz to a 16 kHz sampling rate. The signal is transformed to the STFT domain by applying the Hann window function with a frame length of 400 and hop length of 100, followed by 512-bin FFT.

4.2. Data augmentation

To improve the robustness to variation of input speech, we randomly scale the input speech between -35 dB and -20dB. Similar to [14], we apply random masking to the STFT spectrogram at blocks of consecutive frequency bands and blocks of consecutive time frames. These augmentation steps are applied to both the pretraining and main training phases.

4.3. Model configuration and training procedure

The WaveNet-like structure shown in Fig. 4 is used as the basic block to construct our models. The VQ-VAE with a hierarchical structure similar to [22] is used to estimate the speech variance. In the VQ-VAE model, each encoder consists of six WaveNet blocks and each decoder consists of 12 WaveNet blocks. As for the noise variance estimator, we use a stack of six WaveNet

Table 1: *Results of proposed and baseline methods trained on Voice Bank+DEMAND dataset.*

Method	PESQ-WB	STOI
Noisy	1.97	0.91
SEGAN, 2017 [15]	2.16	0.93
MMSE-GAN, 2018 [16]	2.53	0.93
Wave U-Net, 2018 [17]	2.40	–
MetricGAN, 2019 [18]	2.86	0.92
DCT-UNet, 2019 [19]	2.70	–
μ-law SGAN, 2020 [20]	2.86	0.94
DCCRN, 2020 [21]	2.68	–
DCCRN+, 2021 [21]	2.84	–
Proposed method	**2.85**	**0.94**

Table 2: *Performance of proposed method with- and without phase correction at different SNRs.*

SNR	W/o phase correction		W/ phase correction	
	PESQ-WB	STOI	PESQ-WB	STOI
0 dB	1.714	0.880	1.889	0.885
3 dB	1.913	0.906	2.156	0.910
5 dB	2.096	0.919	2.333	0.922
10 dB	2.509	0.942	2.761	0.943
20 dB	3.254	0.965	3.480	0.966

blocks. The phase correction network is constructed by stacking six WaveNet blocks and two long short-term memory layers.

The training procedure consists of two steps. In step one, the VQ-VAE model is trained to estimate speech variance using the clean train set for 1000 epochs. In step two, the latent codebook of the pretrained VQ-VAE is kept unchanged, and the phase correction network and noise variance estimator are trained together with the pretrained VQ-VAE model for 1000 epochs.

4.4. Evaluation metrics

The Perceptual Evaluation of Speech Quality (PESQ) and ShortTime Objective Intelligibility (STOI) metrics are used to evaluate the proposed method. The PESQ scores, which range from −0.5 to 4.5, measure the speech quality by comparing the enhanced speech signal to the clean reference speech signal. The STOI metric is highly correlated to the perceptual speech intelligibility. The STOI scores range between 0 and 1. For both metrics, a higher score indicates a better result.

4.5. Results on Voice Bank dataset

To ensure a fair comparison, we only compare the performance of our proposed model against other baseline models that are trained using the same Voice Bank+DEMAND dataset. As we can see from the results in Table 1, our proposed method outperforms the non-GAN-based approaches, notably the state-of-art DCCRN method in the 2020 Deep Noise Suppression Challenge (DNS2020) and the improved DCCRN+ model in Interspeech 2021 [21].

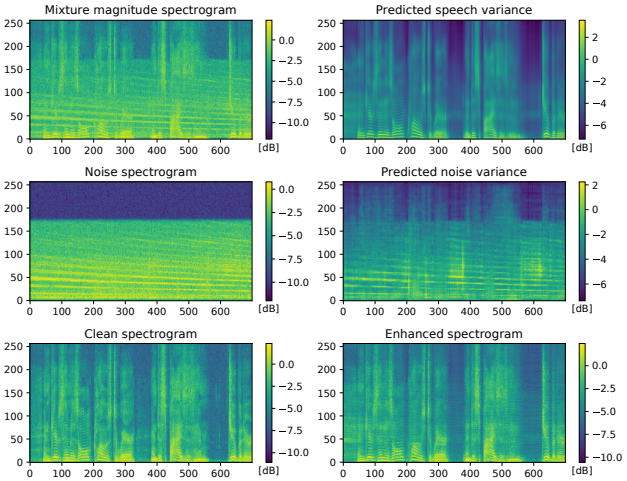

Figure 5: *Predicted variance of speech power spectrogram and noise power spectrogram from the proposed method.*

4.6. Ablation study

To clarify the contribution of the phase correction network, we evaluate the performance of the proposed model with and without the phase correction at different SNRs using the test set of the Voice Bank+DEMAND dataset. As shown in Table 2, the model with the phase correction network obtains higher PESQ and STOI scores. This result demonstrates the advantage of using the phase correction network for speech enhancement.

To verify the accuracy of the predicted speech variance and noise variance, we visualize the output of the speech variance estimator and noise variance estimator networks. Here, we mix an unseen clean speech utterance from the Librispeech dataset with an unseen non-stationary periodic noise sample (siren noise) from the MUSAN dataset at the SNR level of 10 dB to create the noisy speech. As we can see in Fig. 5, the spectrograms of the predicted speech variance and noise variance resemble the harmonic structure of the original speech and noise, respectively. This result demonstrates that the speech and noise components can be correctly separated by the proposed model.

5. Conclusion

This paper proposed a phase-aware speech enhancement method by estimating a complex Wiener filter using a noise-robust VQVAE and a phase correction convolution recurrent network. The results of an ablation test demonstrated that the phase correction is crucial for speech enhancement. Moreover, the objective results indicated that the proposed model outperforms the state-of-art DCRNN model in DNS2020, which proves the effectiveness of our proposed method in enhancing speech quality. The sound samples of our proposed model are available at https://tuanvu92.github.io/IS2022-CVQ.

6. Acknowledgement

This work was supported by a JSPS Grant for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233), by the SCOPE Program of Ministry of Internal Affairs and Communications (Grant Number: 201605002), and by WESTUNITIS CO., LTD.

7. References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

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

[6] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising." in *Proceedings of Interspeech*, 2020, pp. 2467–2471.

[7] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband LSTM," in *Proceedings of Interspeech*, 2020, pp. 2462–2466.

[8] 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 *Proceedings of Interspeech*, 2020, pp. 2472–2476.

[9] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proceedings of Interspeech*, 2020, pp. 2477–2481.

[10] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[11] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[12] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *The 6th International Congress on Acoustics*, 1968, pp. C–17–C–20.

[13] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6309–6318.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[15] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proceedings of Interspeech*, 2017.

[16] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018.

[17] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *ArXiv*, vol. abs/1806.03185, 2018.

[18] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[19] C. Geng and L. Wang, "End-to-end speech enhancement based on discrete cosine transform," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2020, pp. 379–383.

[20] H. Li, Y. Xu, D. Ke, and K. Su, "μ-law sgan for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, 2021.

[21] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise Subband DCCRN with SNR Estimation for Speech Enhancement," 2021.

[22] T. V. Ho and M. Akagi, "Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144.