Interspeech 2022
18-22 September 2022, Incheon, Korea

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

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 the 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 through estimating the magnitude and phase of a complex adaptive Wiener filter. With this method, a noise-robust vector-quantized variational autoencoder is used for estimating the magnitude of the Wiener filter by using the Itakura-Saito divergence on the time-frequency domain, while the phase of the Wiener filter is estimated using 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 methods and achieved a Perceptual Evaluation of Speech Quality score of 2.85 and ShortTime Objective Intelligibility score of 0.94, which is better than the state-of-art method based on cIRM estimation during the 2020 Deep Noise Challenge.

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

1. Introduction

There have been various methods 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, enhancement methods have evolved to handle with various other types of noise by incorporating deep neural networks.

There are many processing domains with 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 only the enhancement of the amplitude features [4]. After clarifying the importance of phase in speech quality and intelligibility [5], several studies developed phase-aware enhancement methods [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 the cIRM makes it difficult for optimization due to the infinite search space [11].

To address this issue, we focused on an approach based on the complex adaptive Wiener filter. The Wiener filter is a common technique that has been applied to many signal enhancement methods. Since the range of the Wiener filter is naturally bounded, it should take less effort than estimating the cIRM.

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 three parameters: speech variance, noise variance, and phase.

A vector-quantized variational autoencoder [12] (VQ-VAE) 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, the VQ-VAE can sufficiently model the distribution of high-quality speech-variance parameters without any unintelligible variation appearing in the vanilla VAE. Furthermore, the encoder network of the VQ-VAE is optimized with a noise-robust training strategy to minimize the variation of the latent variables due to the presence of noise in input speech. The noise variance is estimated using 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 filter by maximizing the scale-invariant signal-to-noise ratio (SI-SNR) in the time domain.

In Section 2, we introduce the speech model upon which our proposed method is based. We then present our speech-enhancement method in Section 3. Objective evaluations are conducted to measure the performance of proposed method. The experimental setup and results are reported in Section 4. We conclude in Section 5 with a brief summary.

2. Complex Wiener filter

In the STFT domain, let the noisy complex spectrogram $\mathbf{X} \in \mathbb{C}^{F\times T}$ be the sum of the 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., $s_{ft} \sim \mathcal{N}_C(0, \sigma^2_{s,ft})$ and $n_{ft} \sim \mathcal{N}_C(0, \sigma^2_{n,ft})$, where $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represent the variances of speech and noise, respectively. Since the speech and noise are uncorrelated, the noisy signal then follows the complex normal distribution as $x \sim \mathcal{N}_C(0, \sigma^2_s + \sigma^2_n)$. By applying the Wiener filter, the power spectral density of clean speech can be estimated from the mixture signal as

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

10.21437/Interspeech.2022-443

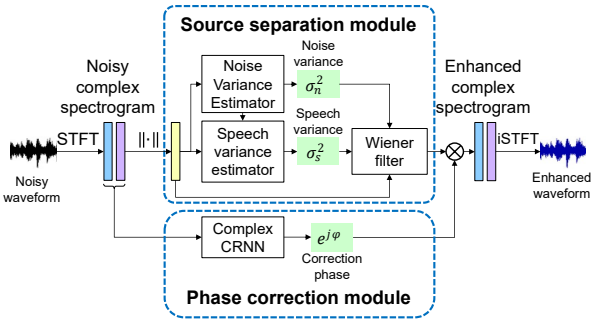

Figure 1: *Overview of proposed method.*

where $\|\hat{s}_{ft}\|$ is the predicted magnitude spectrum of clean speech. To estimate the complex speech spectrum $\hat{s}_{ft}$, the phase $e^{j\varphi_{ft}}$ of the complex Wiener filter can be introduced to Eq. (2) 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 Method

In this section, we describe the proposed method for estimating the parameters of the complex Wiener filter. For estimating the speech variance from the noisy mixture, we use a VQ-VAE pre-trained on a clean dataset. The noise variance and phase of the clean speech are then estimated on the basis of the predicted speech variance. An overview of the proposed method is shown in Fig. 1. Its training process consists of two phases: VQ-VAE pre-training on clean speech and main training on noisy speech. The obtained latent codebook after the pre-training step presumably captures the characteristics of the clean speech. In the main training phase, the latent codebook is preserved while the whole model is trained on the noisy speech mixture. An 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

A VQ-VAE is a generative model that consists of an encoder and decoder network. It basically resembles a communication system, in which 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 a discrete variable $\mathbf{q}$ on the basis of 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 $\hat{\sigma}^2_{s,ft}$ by minimizing the Itakura-Saito (IS) divergence [13] between $\|s_{ft}\|^2$ and $\hat{\sigma}^2_{s,ft}$. The latent codebook is updated simultaneously with other parameters of the model during the training process. Due to the use of the $\arg\min$ function in the quantization process, the computation graph is disconnected, and the model cannot be trained with back-propagation. Therefore, the straight-through reparameterization trick [12] 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 $\hat{\sigma}^2_{s,t}$ is the estimated speech variance vector, $\hat{\mathbf{q}}_t$ is the 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 the 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

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

3.1.2. Method for achieving noise-robustness

The key point of a VQ-VAE for speech enhancement is noise robustness, with which it 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 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 trained on clean speech can achieve higher latent perplexity and lower reconstruction loss. On the basis of this observation, we propose pre-training the VQ-VAE on clean speech first. In other words, we set $\mathbf{x}_t = \mathbf{s}_t$ in the pre-training phase. Except for the latent codebook, the parameters of the encoder and decoder are then fine-tuned on the noisy mixture to achieve noise-robustness. We also introduce 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 to preserve the pre-trained latent codebook.

3.2. Noise-variance estimator

The noise variance $\sigma^2_{n,ft}$ needs to be estimated for the Wiener filter. The noise-variance estimator is trained to reduce the IS divergence between predicted noise variance $\sigma^2_{n,ft}$ and the noise power spectrogram $\|n_{ft}\|^2$ as

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

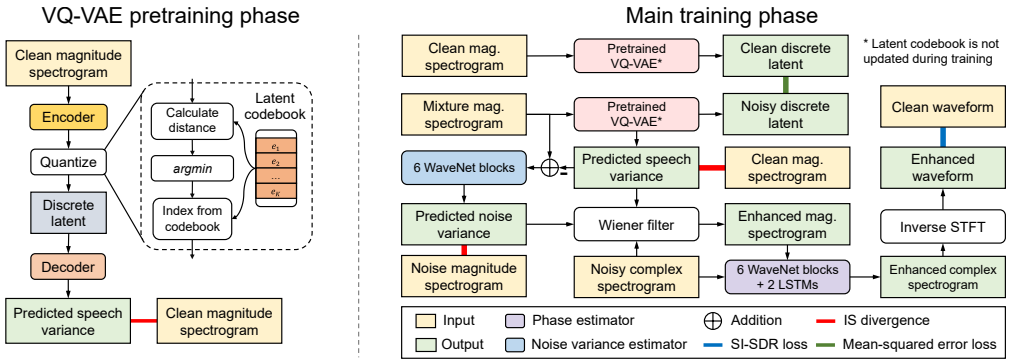

Figure 2: *Block diagram of proposed method. Blocks in pre-training phase corresponds to red block in main training phase.*

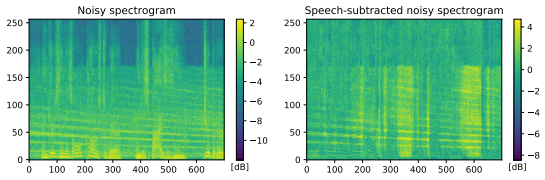

Figure 3: *Left: noisy-speech log-power spectrogram. Right: noisy-speech log-power spectrogram subtracted with estimated speech log-variance that resemble noise log-power spectrogram.*

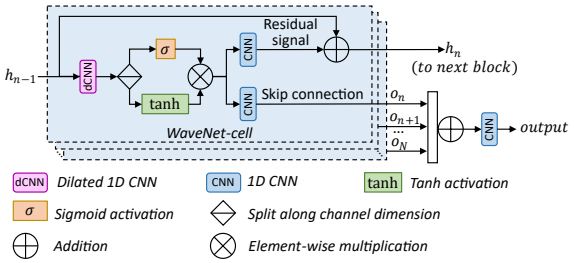

Figure 4: *Architecture of WaveNet block.*

To condition the noise-variance estimator on the estimated speech variance $\sigma^2_{s,ft}$, we empirically subtract the noisy speech log-power spectrogram with the estimated speech log-variance. Although it is not entirely accurate, this results in a representation that better resemble the noise log-power spectrogram, as shown in Fig. 3.

3.3. Phase estimator

Direct phase estimation is difficult due to phase warping. 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. (3). The speech waveform $\hat{\mathbf{y}}$ is then derived using inverse STFT. The phase estimator is trained to maximize the SI-SNR defined as

$$\mathbf{y}_{target} = \frac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle \cdot \mathbf{y}}{\|\mathbf{y}\|_2^2}, \tag{15}$$

$$\mathbf{e}_{noise} = \hat{\mathbf{y}} - \mathbf{y}_{target}, \tag{16}$$

$$\text{SI-SNR} = 10 \log_{10} \frac{\|\mathbf{y}_{target}\|_2^2}{\|\mathbf{e}_{noise}\|_2^2}, \tag{17}$$

where $\hat{\mathbf{y}}$ and $\mathbf{y}$ are 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

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{noise}} - \text{SI-SNR} \tag{18}$$

4. Experimental Evaluation

4.1. Dataset

We used the open dataset released by Valentini et al. [14] to evaluate the proposed method. This dataset has been used with several speech-enhancement methods, which we chose to include as baselines. The clean training set was composed of 28 speakers and the test set of two speakers from the Voice Bank dataset [15]. The noisy training set was constructed by mixing the clean training set with ten types of noise data from the DEMAND dataset at four SNRs: 15, 10, 5, and 0 dB. The noisy test set was constructed by mixing the clean test set with five other types of noise data at four SNRs: 17.5, 12.5, 7.5, and 2.5 dB. All speech waveforms were resampled from a sampling rate from 48 to 16 kHz. The signal was 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 fast FT.

4.2. Data augmentation

To improve the robustness to variation of input speech, we randomly scaled the input speech between -35 dB and -20dB. Similar to a previous study [16], we applied random masking to the STFT spectrogram at blocks of consecutive frequency bands and blocks of consecutive time frames. These augmentation steps were applied to both the pre-training and main training phases.

4.3. Model configuration and training procedure

The WaveNet-like structure shown in Fig. 4 was used as the basic block to construct the modules. A VQ-VAE with a hierarchical structure similar to that in a previous study [24] was used

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

Method	PESQ	STOI
Noisy	1.97	0.91
SEGAN, 2017 [17]	2.16	0.93
MMSE-GAN, 2018 [18]	2.53	0.93
Wave U-Net, 2018 [19]	2.40	–
MetricGAN, 2019 [20]	2.86	0.92
DCT-UNet, 2019 [21]	2.70	–
μ-law SGAN, 2020 [22]	2.86	0.94
DCCRN, 2020 [23]	2.68	–
DCCRN+, 2021 [23]	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	STOI	PESQ	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

to estimate the speech variance. In this VQ-VAE, each encoder consists of six WaveNet blocks, and each decoder consists of 12 WaveNet blocks. We used a stack of six WaveNet blocks for the noise-variance estimator. The phase-correction network was 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 is trained to estimate speech variance using the clean training set for 1000 epochs. In step two, the latent codebook of the pre-trained VQ-VAE is kept unchanged, and the phase-correction network and noise-variance estimator are trained together with the pre-trained VQ-VAE for 1000 epochs. We use One-cylcle Learning Rate scheduler with initial learning rate 5×10^{-4} and maximum learning rate 2×10^{-4} for all training step. To increase the training speed, we leverage mixed-precision training of Pytorch. All models are trained with batch size 256 using 2 Nvidia RTX3090 GPUs.

4.4. Evaluation metrics

The Perceptual Evaluation of Speech Quality (PESQ) [25] and Short-Time Objective Intelligibility (STOI) [26] metrics were used to evaluate the proposed method. The PESQ scores, which range from -0.5 (bad) to 4.5 (excellent), measure speech quality by comparing the enhanced speech signal to the clean reference speech signal. The STOI metric is highly correlated to perceptual speech intelligibility. The STOI scores range between 0.0 (lowest intelligibility) and 1.0 (highest intelligibility). For both metrics, a higher score indicates a better result.

4.5. Results on Voice Bank dataset

From the results reported in Table 1, the proposed method outperformed the many baselines and produced comparative results against other strong baselines, notably the state-of-art Deep Complex Convolution Recurrent Network (DCCRN) method from the 2020 Deep Noise Suppression Challenge (DNS2020) and its improved DCCRN+ method from Interspeech 2021 [23].

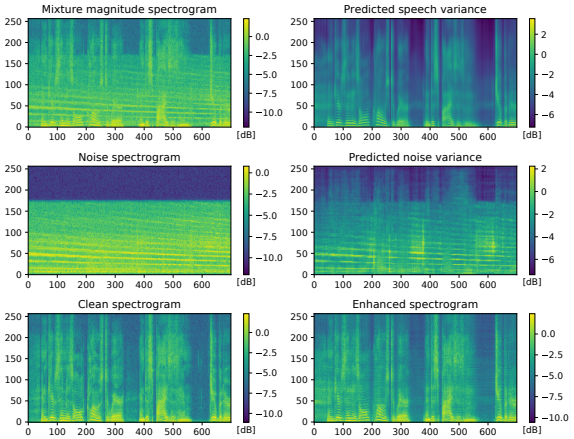

Figure 5: *Predicted variance of speech power spectrogram and noise power spectrogram from proposed method.*

4.6. Ablation study

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

To verify the accuracy of the predicted speech variance and noise variance, we visualized the output of the speech-variance-estimator and noise-variance-estimator networks. We mixed an unseen clean speech utterance from the Librispeech dataset [27] with an unseen non-stationary periodic noise sample (siren noise) from the MUSAN dataset [28] at an SNR 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 indicates that the speech and noise components can be correctly separated with the proposed method.

5. Conclusion

We proposed a phase-aware speech-enhancement method through estimating a complex Wiener filter using a noise-robust VQ-VAE and phase-correction convolution recurrent network. The results of an ablation study indicate that phase correction is crucial for speech enhancement. Moreover, the objective results indicate that the proposed method outperforms the state-of-art method from DNS2020, which proves the effectiveness of the proposed method in enhancing speech quality. Sound samples of the proposed method are available at https://tuanvu92.github.io/IS2022-CVQ.

6. Acknowledgements

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] 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.

[13] 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.

[14] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of Speech Synthesis Workshop (SSW)*, 2016.

[15] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp. 1–4, 2013.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] A. W. Rix, J. G. Beerends, M. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. R. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, 2010.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.