Interspeech 2022
18-22 September 2022, Incheon, Korea

IMPROVING GAN-BASED VOCODER FOR FAST AND HIGH-QUALITY SPEECH SYNTHESIS

Mengnan He, Tingwei Guo, Zhengxin Lu, Ruixiong Zhang, Caixia Gong

DiDi Chuxing, Beijing, China

{hemengnan, guotingwei, luzhenxing, zhangruixiong, gongcaixia}@didichuxing.com

Abstract

Following tremendous success in the Generative Adversarial Network(GAN), the GAN-based vocoders have recently shown much faster speed in waveform generation. However, the quality of generated speech is slightly inferior, and the real-time factor (RTF) still can't be satisfied in many devices with limited resources. To address the issues, we propose a new GAN-based vocoder model.Firstly, we introduce the Shuffle-Residual Block into the generator to get a lower RTF. Secondly, we propose a Frequency Transformation Block in the discriminator to capture the correlation between different frequency bins in every frame. To the best of our knowledge, our model achieves the lowest RTF of the GAN-based vocoders under the premise of ensuring the speech quality. In our experiments, our model shows a lower RTF with more than 40% improvement and higher speech quality than MB-MelGAN and HiFi-GAN V2.

Index Terms: neural vocoder, Shuffle-Residual Block, Frequency Transformation Block, speech synthesis

1. Introduction

In recent years, neural waveform generation models (e.g., WaveNet[1], WaveRNN[2]) have demonstrated the potential to produce superb speech quality[3] and significantly outperform conventional vocoders[4, 5, 6] used in traditional statistical parametric speech synthesis (SPSS). However, the RTF of these models is poor, which limits their application in real-time TTS scenarios. Although there exist many optimization methods for RTF, such as LPCNet which combines linear prediction coding with neural vocoder[7], the RTF is still not good enough due to the autoregressive mechanism.

In order to improve the inference speed, the GAN-based neural network has been applied to the vocoder[8], such as Parallel WaveGAN, MelGAN[9, 10]. These vocoders adopt non-autoregressive architecture and have obvious advantages in RTF compared with autoregressive models. However, in many scenarios where require low RTF, such as human-computer dialogue or map navigation, the inference speed of existing models is not fast enough in relatively poor CPUs which are commonly used in[11], and the speech quality also affects the user experiences. Although some GAN-based models[12, 13, 14] have been proposed to improve speech quality, non-autoregressive models are still slightly worse than autoregressive models. Some vocoders like Cargan[15]introduced autoregressive structures to improve the speech quality, but the inference speed is significantly slower.

Beacuse the residual structure in the generator of GAN-based vocoders is similar to ResNet[16], we consider using an acceleration model ShuffleNet[17, 18] which can be seen as a compressed version of ResNet. In computer vision, it was proven to be much faster and its accuracy can be guaranteed. Improving sound quality is similar to a speech enhancement process, making the generated speech closest to the original speech. On one hand, we use SI-SNR loss proposed in TasNet and Conv-TasNet[19, 20], to guarantee the continuity of the signal in the time domain. On the other hand, we utilize the correlation in the time-frequency domain as the additional information which was utilized in PHASEN or DCCRN[21, 22] to further improve the speech quality.

We get inspiration from these improvements and propose a new GAN-based vocoder, which achieves the state-of-the-art RTF and better speech quality. We summarize the contributions as below:

1) We propose Shuffle-Residual Block for the generator to speed up speech generation.

2) To improve speech quality, we propose Frequency Transformation Block into the discriminator to capture the relevant information between frequency bins to get a more effective discrimination.

3) The model we proposed achieves the lowest RTF of the GAN-based vocoders under the premise of ensuring the speech quality. With the deployment of our model, we believe that more users can benefit from high speech quality on a variety of devices.

2. RELATED WORK

Multi-band MelGAN (MB-MelGAN) [23]is a fast waveform generation model targeting to high-quality TTS. The generator uses three layers and an orthogonal filter. The purpose of the orthogonal filter is to downsample the full-band signal into sub-bands signals, making network learn signal information through every sub-band. Then, the sub-bands signals are synthesized into the full-band signal. This is the main contribution of the MB-MelGAN, which improves the RTF by reducing the number of generator layers.

As showed in [9,21], STFT Loss can help the model generate higher quality speech and converge faster in the training stage. The signal STFT Loss contains L_{Sc} and L_{mag} which denote spectral convergence and log magnitude loss respectively. They are defined as:

$$L_{sc}(x, \tilde{x}) = \frac{\| |\operatorname{STFT}(x)| - |\operatorname{STFT}(\tilde{x})| \|_F}{\| |\operatorname{STFT}(x)| \|_F} \quad (1)$$

$$L_{mag}(x, \tilde{x}) = \frac{1}{N} \| \log |\operatorname{STFT}(x)| - \log |\operatorname{STFT}(\tilde{x})| \|_1 \quad (2)$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ represent the Frobenius and 1-norms, respectively. x and $\tilde{x}$ represent the real speech and the generated speech. $|STFT(\cdot)|$ indicates the STFT function to compute magnitudes and N is the number of elements in the magnitude. For the multi-band, it conducts multi-resolution STFT in both full-band and sub-band scales. The multi-resolution STFT Loss

10.21437/Interspeech.2022-730

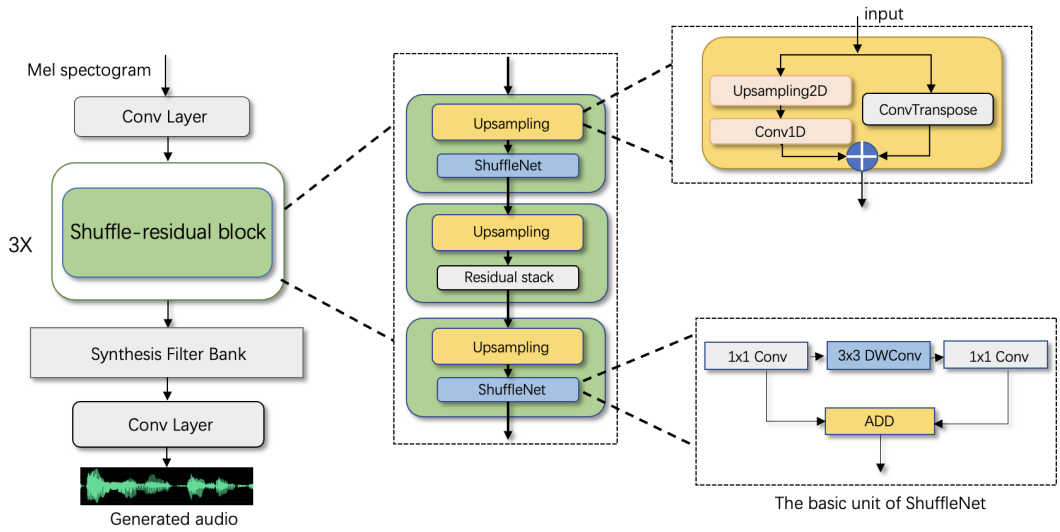

Figure 1: *The GAN-based generator model we proposed. The dashed box in the center is the specific arrangement and details of the generator Layer which we called Shuffle-Residual block. The dashed box in the top right corner is upampling layer with the information of Linear interpolation sampling added. The dashed box on the right corner is the structure of the ShffleNet unit with data calculation process.*

becomes

$$L_{mr_{stft}}(G) = \frac{1}{2}\left(L^{full}_{fmr\_stft}(G) + L^{sub}_{smr\_stft}(G)\right) \quad (3)$$

where $L^{full}_{fmr\_stft}$ and $L^{sub}_{smr\_stft}$ are the multi-resolution STFT Loss in full-band and sub-band, respectively.

3. PROPOSED MODEL

3.1. Generator

3.1.1. Shuffle-Residual Block

Generally, the stacked residual structure is commonly used in the GAN-based vocoder's generator, so we consider optimizing the residual structure to reduce the computation complexity and improve the inference speed. To the best of our knowledge, MB-MelGAN has achieved the fastest inference speed in GAN-based vocoders with the same amount of parameters. So we use MB-MelGAN as our base model. In MB-MelGAN, the residual block has the similar structure as ResNet-3 which was stacked four times. However, we found that the convolution used in these residual structures is convoluted on all input feature maps. This full-channel dense connection method brings a lot of computation.

To address the issue, we consider introducing ShuffleNet to reduce the complexity. Firstly, ShuffleNet uses the group convolution to group different feature maps of the input, and then uses different convolution kernels in different groups, which would reduce the convolution calculations. Secondly, ShuffleNet shuffles different channels to solve the drawbacks caused by the group convolution. The basic unit of ShuffleNet is improved on a residual unit as shown in Fig. 1 in the bottom with dash box, which is a residual unit with 3 layers. The first layer comes a 1x1 convolution, then the second layer is 3x3 pointwise convolution, which is the bottleneck, and the third layer is 1x1 convolution. Finally, a short cut connects the input and the out together to the activation function. Thanks to the pointwise group convolution with channel shuffle, all components in ShuffleNet unit can be computed efficiently.

We split the input into two groups and then divide the channels in each group into several subgroups, then feed each group in the next layer with different subgroups. Our experiment found that two groups are the most suitable. More groups can speed the inference up more quickly, but the speech quality will have a loss. The process is shown in Fig. 1.

Considering the speech quality of the generator, we found that the speech quality would be decreased if we only use ShuffleNet to replace the original structures. We think the reason is that the operation of channel shuffle disrupts the continuous characteristics of the audio to some extent. So we use ShuffleNet, and Residual Stack in turn, which also stacked 4 times based on MB-MelGAN. The three residual structures are combined with upsampling to compose our Shuffle-Residual Block. We use the ShuffleNet in the first and the third layer of the Shuffle-Residual Block. In our experiments, this structure can ensure the speech quality in maximum.

We found that some of the information is inaccurate or lost when the input features were upsampled through deconvolution. Using deconvolution alone does not represent the information between adjacent samples very well. In order to improve the speech quality, more information are added to the upsampling layer.

Therefore, we add a bilinear interpolation together with the deconvolution as the input of the Shuffle-Residual Block, which provides supplementary information. Besides, we add one convolution layer after the bilinear interpolation to get higher-dimensional features. The specific process of interpolation sampling is shown in Fig. 1.

3.1.2. Time domain loss

STFT Loss in MB-MelGAN only considers the difference between signal amplitude and energy in frequency domain, which will lose information on phase and result in discontinuity in speech. Once the speech signal is discontinuous, artifacts will

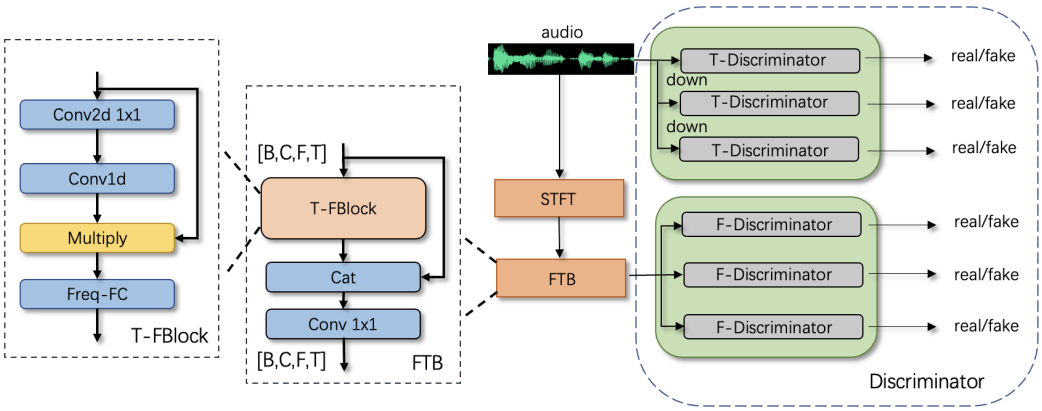

Figure 2: *The GAN based discriminator model we proposed. The discriminator is composed of multi-scale time and multi-scale frequency which is the dashed box in the right. Among them, the input of multi-scale frequency has to go through F-TBlock. The dashed box in the middle represents the details of the F-TBlock. The dashed box in the left represents the details of the key part T-FBlock of the F-TBlock.*

appear on the spectrogram and it sounds like the speech is choppy and jittery. In the time domain, it can reflect the continuity of time periods. Currently, there is no efficient time-domain loss for the generator.

Therefore, we introduce the time domain loss called SI-SNR Loss[19] which is widely used in speech separation. It can be defined as follows:

$$\text{Starget} = \frac{\langle \widehat{s}, s \rangle s}{\|s\|^2} \tag{4}$$

$$e = \widehat{s} - \text{Starget} \tag{5}$$

$$SI - SNR = 10 \log \frac{\| \text{Starget} \|^2}{\|e\|^2} \tag{6}$$

where $\hat{\mathbf{s}} \in \mathbb{R}^{1\times T}$ and $\mathbf{s} \in \mathbb{R}^{1\times T}$ are the generated speech and the original speech, respectively. $\|\mathbf{s}\|^2 = \langle \mathbf{s}, \mathbf{s} \rangle$ denotes the signal energy. Starget is a modified signal, by projecting the generated signal onto the real signal space and making a measure of scale normalization. And e means differences between the generated signal and the modified signal. $\langle \widehat{s}, s \rangle$ means cross power in time domain which carries the phase relationship. The validity of this loss function is proved in our experiments. The multi-resolution loss function becomes:

$$\begin{aligned} L_{mr_{stft}}(G) = \alpha_1 \times L^{full}_{fmr\_stft}(G) + \\ \alpha_2 \times L^{sub}_{smr\_stft}(G) + \alpha_3 \times SI - SNR \end{aligned} \tag{7}$$

The total loss function in the generator constrains the model in both the time and frequency domain.

3.2. Discriminator

We discover that the correlation between frequency bins in each frame has positive influence on a fine-grained discrimination. Because the signals are not independent, the details of the signal waveform reconstruction will lost without considering the correlation of the signal. However, the existing GAN-based vocoders do not take this into account. Therefore, the Frequency Transformation Block(F-TBlock)[21] is adopted in this paper to capture the correlation along the frequency axis. This correlation information between frequency bins is introduced into the discriminator to solve the vertical streak-like pseudo-peaks that appear on the spectrum to some extent. The key part of the F-TBlock is the Time Frequency Block(T-FBlock), which is equivalent to an attention mechanism, so that the output features of the F-TBlock contain information about the other frequency bins in each frame. As shown in Fig.2, firstly, the input goes through two convolution layers to get high dimension features. Secondly, it comes to the multiply module which cross-multiplies each frequency bin. A transition to a high-dimensional feature is then implemented through a full connection so that the output feature has a weight that takes account of all the frequency bins. Finally, the output of T-FBlock and the original input are connected together and then sent to a 1x1 convolution layer. The input passes through F-TBlock after being transformed by STFT, and then would be sent to the discriminator. Therefore, the input of the frequency discriminator not only has the information of original frequency bins, but also has the mutual information between the frequency bins through the F-TBlock. The specific network structure is shown in Fig. 2. The input size is [B,C,F,T] which means [batch size, channel, frequency dims, frames].

Stronger frequency bins correlation between adjacent frames result in smoother and more stable resonance peak in the low and medium frequency domain which can reduce the breakages of the generated speech. Our demo has proved this clearly.

4. EXPERIMENTS

4.1. Data set

The training data and the testing data we used are the Chinese Standard Mandarin Speech Corpus (CSMSC) of 12 hours' recordings about 10000 audios. All the audios were downsampled to 16kHz with 16-bit format. 9500 audios were used for model training and the rest for validation. The feature is an 80-dimensional mel-spectrogram with (fft size, hop length, window length) are (1024, 256, 1024).

Table 1: *Comparison of Model size(Paras size), GFLOPS, RTF, MOS and the MOS of TTS Task*

	Paras.(M)	GFLOPS	RTF	MOS	TTS(MOS)	TTS(RTF)
HiFi-GAN V2	0.92	5.2	0.385	4.12 ± 0.05	3.84 ± 0.04	2.96
MB-MelGAN	1.62	0.95	0.069	4.00 ± 0.04	3.73 ± 0.06	1.13
Ours	1.34	**0.79**	**0.040**	$\mathbf{4.18 \pm 0.04}$	$\mathbf{3.92 \pm 0.05}$	**0.90**
Ground Truth	–	–	–	4.43 ± 0.03	–	–

4.2. Experimental setup

Generator. We use three upsampling layers to realize 64x upsampling, where the up-factors are (2,4,8). In loss function (7), we set α_1, α_2 and α_3 as (0.5,0.3,0.2).
Discriminator. The multi-scale discriminator is the same as that of the MelGAN. For the frequency discriminator, the waveform passes through the STFT using the hanning window whose fft size, hop length and window length are (512,120,600), then the frequency signal is fed into FTB, whose frequency dim, channel are (257, 2). The two channels represent amplitude and phase respectively. The kernal size in FTB is 7 and same padding for each convolution.
Training. The Adam optimizer was adopted with a learning rate varying with the number of steps for both generator and discriminator. We first train the generator about 20k steps, then train the generator and the discriminator together until 100k steps. The model was trained on a single NVIDIA TESLA(R) P40 GPU.
Test. We choose MB-MelGAN and HiFi-GAN V2 as the baseline for comparison to demonstrate the proposed model. Because MB-MelGAN is the lowest RTF model at present and HiFi-GAN V2 is a good balance model between RTF and speech quality. We randomly select 100 audios from the validation. Ten native Chinese Listeners were asked to rate the overall quality of speech samples on the mean opinion score(MOS) which were obtained using the crowdsourcing methodology described in P.808[24] with 95% confidence intervals. 500 audios were used for testing the RTF with three times average on CPU (Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz).

4.3. Results

The results of MOS, the model size and RTF are shown in Tab 1. The results show that the MOS of our proposed model is higher than that of MB-MelGAN and HiFi-GAN V2 with 0.1 MOS improvement. Here, the RTF is as small as possible, less than 1 is real-time. The RTF of our model is almost 40% faster than others. Compared to MB-MelGAN, our model effectively reduces the total computational complexity from 0.95 to 0.79 GFLOPS. Also, the model size is smaller than MB-MelGAN. From this point of view, our proposed vocoder achieves the lowest RTF with guaranteed speech quality.

In order to verify the effectiveness of the proposed vocoder for the TTS task, we combined the vocoder with the Fastspeech2-based acoustic model[25]. FastSpeech2 predicts mel-spectrograms, which are fed into the vocoder to produce waveform. TTS Task in Table 1 summarizes the MOS value for 100 synthesized testing audios. The results indicate that the improved version outperforms the basic MB-MelGAN and HiFi-GAN V2 in TTS task with a higher MOS. Here are some demos for this paper[1].

We also test the RTF in a smart phone whose CPU is Helio A22 with 4 core. The computing resources of this CPU are very low. The result of TTS(RTF) in Tab 1. shows that the RTF of our model is 0.9 which meets the requirements of real-time synthesis. The RTF of MB-MelGAN and Hifi-GAN V2 is over 1. The RTF greater than 1 means that the synthesis takes much more time than the playback, which is intolerable in actual use. In addition to this, a higher real-time rate means a higher CPU usage.

[1] https://cookingbear.github.io/research/

Table 2: *Subjective evaluation results (MOS) of ablation study. (Shuffle-Residual Block, SI-SNR and F-TBlock(FTB))*

Model	MOS (CI)
Ground Truth	4.43($\pm$0.03)
Our model	4.18($\pm$0.05)
w/o Shuffle-Residual Block	4.10($\pm$0.03)
w/o SI-SNR	4.14($\pm$0.06)
w/o FTB	4.09($\pm$0.05)
w/o Shuffle-Residual Block and FTB	4.04($\pm$0.05)
w/o Shuffle-Residual Block and SI-SNR	4.07($\pm$0.04)
w/o FTB and SI-SNR	4.06($\pm$0.07)

4.4. Ablation study

In order to analyze the performances of the proposed model brought by different modules, we set up several ablation experiments and the experimental results are shown in Tab 2. The number of training steps and configurations are the same for each model. As we can see, each improvement combined with the vocoder alone can have a positive effect in speech quality which has 0.04 MOS improvement at least. Among them, the performance of SI-SNR Loss in time domain is not significantly improved. On one hand, constraint in the time domain is averaged at the sample point level, which would weaken the loss function. On the other hand, not all audios have the problem of phase discontinuity, and some of the sound perception are not obvious. Our proposed model is the combination of these three improvements which has the best performance with the MOS of 4.18. The ablation study proves the validity of our proposed model.

5. Conclusion

In this paper, we propose a state-of-art GAN-based vocoder, which is more suitable for edge devices, meanwhile guarantee a better sound quality. We design the Shuffle-Residual Block to replace the original residual structures improving the synthesis speed. Also, We proved that it is effective to add time domain supervision to frequency domain supervision improving the speech quality. What's more, we also propose the F-TBlock in the discriminator, which can capture the correlation information between the frequency bands, so that the sound quality has been further improved. Experimental results indicate that the proposed vocoder has advantages in terms of inference speed and speech quality.

6. References

[1] A. Oord, S. Dieleman, H. Zen, K. Simonyan, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," 2018.

[3] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *10th ISCA Speech Synthesis Workshop*, 2019.

[4] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," vol. 99, no. 7, 2016, pp. 1877–1884.

[5] K. Hideki, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds(introduction to the amazing world of sounds with demonstrations)," 2006.

[6] A. Sharma, P. Kumar, V. Maddukuri, N. Madamshettib, K. Kg, S. Kavurub, B. Raman, and P. P. Roy, "Fast griffin lim based waveform generation strategy for text-to-speech synthesis," 2020.

[7] J. M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[8] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," 2019.

[9] R. Yamamoto, E. Song, and J. M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," 2019.

[10] K. Kumar, R. Kumar, T. D. Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. D. Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[11] D. e. a. Carole-Jean, Wu, "Machine learning at facebook: Understanding inference at the edge," *IEEE International Symposium on High Performance Computer Architecture (HPCA). 0.*, 2019.

[12] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[13] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," 2020.

[14] R. Yang, T. J. Cao, X. Chen, and F. R. Zhang, "A novel and universal gan-based countermeasure to recover adversarial examples to benign examples," 2021.

[15] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," *arXiv preprint arXiv:2110.10139*, 2021.

[16] V. V. e. a. Szegedy C, Ioffe S, "Inception-v4, inception-resnet and the impact of residual connections on learning[j]," 2016.

[17] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017.

[18] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Springer, Cham*, 2018.

[19] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," vol. PP, no. 99, 2019, pp. 1–1.

[20] ——, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," 2018, pp. 696–700.

[21] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," vol. 34, no. 5, 2020, pp. 9458–9465.

[22] Y. Hu, Y. Liu, S. Lv, M. Xing, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," 2020.

[23] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," 2020.

[24] ITU-T, "Recommendation p.808: Subjective evaluation of speech quality with a crowdsourcing approach," 2018.

[25] Y. Ren, C. Hu, T. Qin, S. Zhao, and T. Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," 2020.