



XiaoiceSing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network

*Chunhui Wang¹, *Chang Zeng^{2,3}, Xing He¹

¹Beijing Bombax XiaoIce Technology Co., Ltd, China

²National Institute of Informatics, Japan ³SOKENDAI, Japan

{wangchunhui, hexing}@xiaobing.ai, zengchang@nii.ac.jp

Abstract

The over-smoothing problem in the middle- and high-frequency areas prevents the acoustic model from generating high-quality singing voices. In this paper, we propose XiaoiceSing2, which is a generative adversarial network consisting of a FastSpeech-based generator and a multi-band discriminator, to generate the full-band mel-spectrogram. Specifically, we improve the feed-forward Transformer (FFT) block by adding multiple residual convolutional blocks in parallel with the self-attention block to balance the local and global features. The multi-band discriminator contains three sub-discriminators responsible for low-, middle-, and high-frequency parts of the mel-spectrogram, respectively. Each sub-discriminator is composed of several segment discriminators (SD) and detail discriminators (DD) to distinguish the audio from different aspects. The experiment on our internal 48kHz singing voice dataset shows XiaoiceSing2 significantly improves the quality of the singing voice over XiaoiceSing.

Index Terms: Singing voice synthesis, feed-forward transformer, generative adversarial network

1. Introduction

Recently neural network for singing voice synthesis [1, 2, 3, 4] has attracted a lot of attention since deep learning has achieved great gain in text-to-speech (TTS) task [5, 6, 7] which has a similar pipeline to SVS. Some studies [1, 3, 8, 9, 10, 11] reported the promising results of the proposed models on synthesizing high-fidelity 48kHz singing voices. For instance, XiaoiceSing [3] modified the architecture of FastSpeech [6] to adapt the task of high-fidelity SVS and it was combined with WORLD [12] vocoder to generate 48kHz singing voices. HifiSinger [11] utilized a sub-frequency GAN in the acoustic model and a multi-length GAN in the vocoder to better reconstruct the high-fidelity singing voices.

However, due to no special design for generating the full-band mel-spectrogram, these studies work not well in high-fidelity SVS scenarios in which middle- and high-frequency parts possess stronger emotion and expressiveness. Further, this over-smoothing problem in the generated mel-spectrogram results in the vocoder failing to reconstruct the high-fidelity waveform from it owing to its low-quality [11, 13, 14].

In order to solve the over-smoothing problem of middle- and high-frequency areas, we present a novel high-fidelity singing voice synthesizer XiaoiceSing2 based on a generative adversarial network [15] to generate a more realistic mel-spectrogram since GAN can theoretically approximate the real data distribution via the adversarial training. The proposed Xi-

oiceSing2 is composed of a FastSpeech-based generator and a multi-band discriminator. For the generator, we follow the design of XiaoiceSing [3] but improve the feed-forward Transformer (FFT) block [16] of it by adding multiple residual convolutional blocks in parallel with the multi-head self-attention (MHSA) block [17] to balance the local and global features. Because we argue that the global features generated by the MHSA block are prone to being over-smoothing for the middle- and high-frequency parts, which can be alleviated by introducing local information from the multiple residual convolutional blocks.

As for the multi-band discriminator, similar to HiFiSinger [11], it consists of three sub-discriminators responsible for low-, middle-, and high-frequency parts of the mel-spectrogram, respectively. Moreover, each sub-discriminator contains several segment discriminators (SD) and detail discriminators (DD) for distinguishing the mel-spectrogram from the segments with different window lengths and local time-frequency patterns, respectively. The segment discriminators are able to cover the different levels of long-term dependencies by applying multiple windows with different lengths on the mel-spectrogram to increase the capability of the discriminator. Similar to PatchGAN [18, 19, 20], the detail discriminator divides the mel-spectrogram into multiple time-frequency patches so that it can pay more attention to the middle- and high-frequency regions and the generator also benefits from the stronger discriminator to produce a more realistic mel-spectrogram.

In the experiment, XiaoiceSing2 is combined with a high-fidelity vocoder HiFi-WaveGAN [21] which is designed to reconstruct the 48kHz waveform and the result shows XiaoiceSing2 significantly improves the quality of the singing voice over XiaoiceSing in term of mean opinion score (MOS) metric. We also make a comparative study of the middle- and high-frequency areas generated by XiaoiceSing2 and XiaoiceSing via visualizing the mel-spectrogram. Besides, an ablation study is conducted to show the contribution of the proposed components.

The rest of this paper is organized as below. Section 2 illustrates the detailed architecture of XiaoiceSing2 including the generator and discriminator. The experimental settings including the dataset, baseline system, and training methodology are shown in Section 3. In addition, the MOS test result* and an ablation study are also reported in this section. Finally, we conclude this paper in Section 4.

2. XiaoiceSing2

In order to generate the mel-spectrogram with more fine-grained middle- and high-frequency parts, we adopt an adversarially training strategy to optimize XiaoiceSing2, which is different

*These authors contributed equally to this work.

*Demo page: <https://wavelandspeech.github.io/xiaoice2>

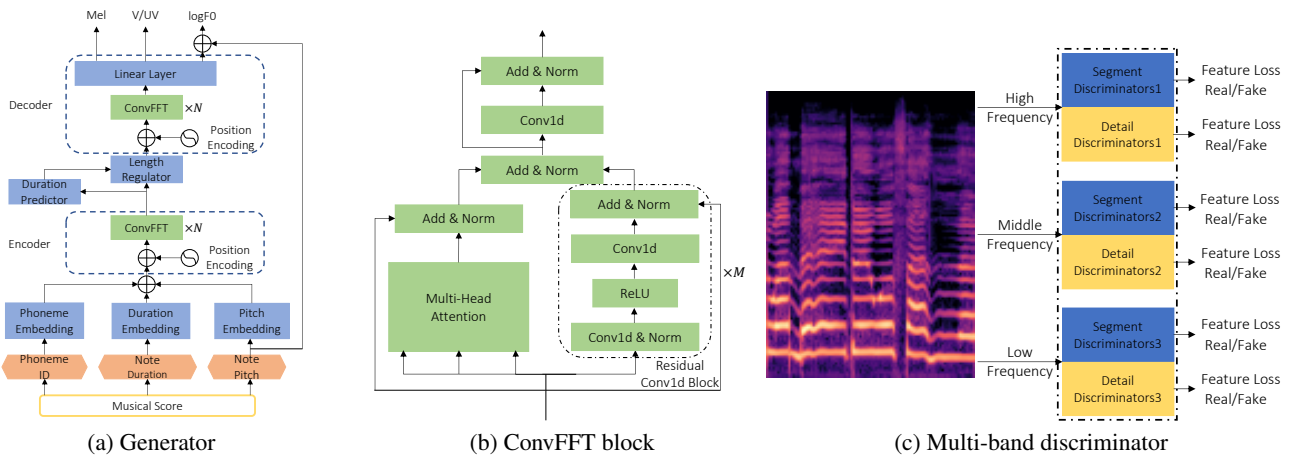


Figure 1: The architecture of XiaoiceSing2. (a). The improved feed-forward Transformer. (b). Feed-forward Transformer with parallel residual convolutional block. (c). Multi-band discriminator, consisting of three sub-discriminators, and each contains several segment discriminators and detail discriminators.

from XiaoiceSing that directly trains the feed-forward Transformer [16] as the acoustic model. The generator of XiaoiceSing2 uses the novel ConvFFT blocks to efficiently leverage the local and global information as Figure 1(a) shows. Additionally, since the generator can benefit from a powerful discriminator, we use three sub-discriminators to distinguish the low-, middle-, and high-frequency parts of the mel-spectrogram. And each sub-discriminator employs several segment and detail discriminators to identify the mel-spectrogram from different aspects.

2.1. Generator

The input to the generator is the musical score consisting of lyrics, note duration sequence, and note pitch sequence. The lyrics are converted to phoneme sequences by the grapheme-to-phoneme (G2P) tool [22]. All sequences are transformed into their own embedding spaces by the corresponding embedding modules and these embedding sequences are concatenated as shown in Figure 1(a).

As for the architecture of the generator, XiaoiceSing2 follows XiaoiceSing which is a FastSpeech-based [6] acoustic model. The generator can be divided into an encoder, a length regulator with a duration predictor, and a decoder as shown in Figure 1(a). The encoder converts the concatenated sequence into a hidden space which is considered to be shared with the mel-spectrogram [6, 7]. Consequently, the output sequence of the encoder can be expanded by the length regulator according to the result of the duration predictor to directly match the length of the target mel-spectrogram. Finally, the expanded sequence is transformed by the decoder to predict the mel-spectrogram, V/UV decision, and the logF0 value. Note there is a residual connection between the input pitch sequence and the predicted logF0 sequence to lower the training difficulty [3].

In this paper, both the encoder and decoder contain 6 ConvFFT blocks which are improved from the FFT block used in [3, 6, 7]. As Figure 1(b) shows, the ConvFFT block incorporates multiple residual convolutional blocks in parallel with the MHSA block since we believe that the over-smoothing problem of middle- and high-frequency areas is intensified if only the global information extracted by MHSA is used. To rectify

this problem in the generated mel-spectrogram, the local information is extracted by the stacked residual convolutional blocks which share the same input with MHSA, then it is added to the global information for fusion. In the encoder, each ConvFFT block has 2 residual convolutional blocks. In the decoder, the number is 5.

2.2. Multi-band discriminator

Because the strong discriminator is beneficial to the generator, we utilize three sub-discriminators to work on the low-, middle-, and high-frequency parts of the mel-spectrogram as Figure 1(c) shows. In this paper, the dimension of the mel-spectrogram is 120 and it is divided into low-frequency (0-60), middle-frequency (30-90), and high-frequency (60-120) parts. Each sub-discriminator has several segment and detail discriminators for identifying the mel-spectrogram from long-term dependencies as well as time-frequency patterns.

2.2.1. Segment discriminator

The idea of the SD is similar to the multi-length GAN (ML-GAN) in HiFiSinger [11]. However, instead of applying ML-GAN on the waveform, we utilize our SD for the mel-spectrogram via randomly clipping the input mel-spectrogram by different window lengths which are set as [200, 400, 600, 800], and the whole segment in this paper. All segment discriminators have the same architecture which is a 10-layers 1-dimensional (1-d) convolutional neural network (CNN) with 3 kernel size and 128-dimensional hidden channel. The 1-d CNN is able to promote the continuity of the mel-spectrogram produced by the generator along the time axis by distinguishing the long-term dependencies with different lengths. In addition to outputting the real/fake decision, the intermediate feature maps generated by the hidden layers are also collected for calculating the feature loss which is described in Section 2.3.3.

2.2.2. Detail discriminator

Although segment discriminators promote the continuity of the generated mel-spectrogram, they cannot benefit to generate the

high-quality middle- and high-frequency parts which are significant to produce high-fidelity singing voices [13]. By taking this into account, we accompany a detail discriminator for each segment discriminator. The motivation for generating more fine-grained middle- and high-frequency areas by using the detail discriminator comes from the PatchGAN [18, 19, 23] which utilizes a fully convolutional discriminator for generating high-resolution images. To be specific, the first 2-dimensional convolutional layer with (3, 3) kernel size upsamples the input mel-spectrogram to 32 channels. The rest of the network consists of 5 convolutional layers with (3, 3) kernel size and (2, 2) dilation for downsampling, and 5 convolutional layers with (1, 3) kernel size for output. These layers are alternately stacked. The outputs from each output layer are collected for calculating the adversarial loss. And the outputs from each downsampling layer are collected for computing the feature loss. Due to the detail discriminator, the mel-spectrogram is divided into multiple time-frequency patches for identifying whether they are real or fake, which is helpful to better construct the middle- and high-frequency parts.

2.3. Loss function

The loss for XiaoiceSing2 is a weighted sum of several loss terms, which is formulated as follows,

$$\mathcal{L}_G = \lambda_1 * \mathcal{L}(G; D) + \lambda_2 * \mathcal{L}_a + \lambda_3 * \mathcal{L}_f, \quad (1)$$

$$\mathcal{L}_D = \mathcal{L}(D; G), \quad (2)$$

where \mathcal{L}_G denotes the generator loss, $\mathcal{L}(G; D)$ is the adversarial loss for the generator, \mathcal{L}_a denotes the acoustic loss similar to the one used in [3], and \mathcal{L}_f represents the feature loss. While \mathcal{L}_D denotes the discriminator loss, which only possesses the adversarial loss term $\mathcal{L}(D; G)$ for the discriminator. As for the weights λ_1 , λ_2 , and λ_3 in Eq. (1), they are set as 0.1, 1, and 1 in this paper, respectively.

2.3.1. Adversarial loss

The adversarial loss proposed in LS-GAN [24] is also used in the training stage of XiaoiceSing2. The formula is shown as

$$\mathcal{L}_{adv}(G; D) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} [(1 - D(G(\mathbf{z})))^2], \quad (3)$$

$$\mathcal{L}_{adv}(D; G) = \mathbb{E}_{\mathbf{x} \sim p_{data}} [(1 - D(\mathbf{x}))^2] + \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} [D(G(\mathbf{z}))^2], \quad (4)$$

where \mathbf{z} denotes the random noise and \mathbf{x} is the real mel-spectrogram. This format of adversarial loss can avoid the gradient vanishing while training the GAN [24].

2.3.2. Acoustic loss

The acoustic loss is also a weighted sum of the loss terms for the predicted acoustic features, which is shown as

$$\mathcal{L}_a = \alpha_1 * \mathcal{L}_{mel} + \alpha_2 * \mathcal{L}_{pitch} + \alpha_3 * \mathcal{L}_{V/UV} + \alpha_4 * \mathcal{L}_{dur}, \quad (5)$$

where \mathcal{L}_{mel} , \mathcal{L}_{pitch} , and \mathcal{L}_{dur} are MSE loss for the mel-spectrogram, pitch, and duration, respectively. While $\mathcal{L}_{V/UV}$ is a binary cross-entropy loss for the V/UV decision. Besides, the weights α_1 , α_2 , α_3 , and α_4 are set as 1, 0.01, 0.01, and 0.1, respectively.

2.3.3. Feature loss

Feature loss was proposed in [25] and was introduced into the speech field in MelGAN [26]. The generator can make full use of the information brought by the discriminator via learning from the L_1 similarity metric between the feature maps of the real and fake data. It can be formulated as

$$\begin{aligned} \mathcal{L}_f = \mathbb{E}_{\mathbf{z}, \mathbf{x}} [& \sum_{k=1,2,3} \left(\sum_{i=1}^{L_s} \frac{1}{N_i} \|D_{ks}^i(\mathbf{x}) - D_{ks}^i(G(\mathbf{z}))\|_1 \right. \\ & \left. + \sum_{j=1}^{L_d} \frac{1}{N_j} \|D_{kd}^j(\mathbf{x}) - D_{kd}^j(G(\mathbf{z}))\|_1 \right), \end{aligned} \quad (6)$$

where $\|\cdot\|_1$ denotes the L_1 distance, $D_{ks}^i(\cdot)$ and $D_{kd}^j(\cdot)$ denote the feature maps of i -th and j -th layer of the k -th SD and DD, respectively. N_i and N_j are the numbers of the corresponding feature map. L_s and L_d denote the number of layers of SD and DD, respectively. Since the multi-band discriminator has three sub-discriminator, the feature loss for each sub-discriminator is merged in Eq. (6).

3. Experiments

3.1. Dataset

We conduct the experiment on our internal singing voice dataset including 6917 pieces of singing voices from a Mandarin female singer, which is identical to the one used in HiFi-WaveGAN [21]. All audios are sampled at 48kHz. The duration of audio in the dataset ranges from 4s to 10s and the total duration is 5 hours. We transform each audio into the corresponding STFT spectrogram by applying 20ms window with 5ms shift. The spectrogram is converted to mel-scale by 120 filters. The pitch and V/UV decision are extracted by using the Parselmouth [27] toolkit which is a Python interface to Praat [28]. As for the division of data, we randomly choose 300 segments for validation and 300 segments for testing. The remaining data is used for training.

3.2. Baseline system

Since XiaoiceSing2 is improved based on XiaoiceSing, XiaoiceSing is selected as the baseline system. XiaoiceSing adapts FastSpeech [6] from TTS to SVS by extending the inputs and outputs of the model. It concatenates the pitch sequence, duration sequence, and phoneme sequence as the input to the model and it outputs the mel-spectrogram, V/UV decision, and logF0 for the vocoder. Besides, it uses a residual connection between input pitch and output logF0 to lower the difficulty of training. We train the model with the same optimization strategy in [3]. As for the vocoder, we utilize the HiFi-WaveGAN [21] to generate high-fidelity singing voices for fair comparison because it is designed for the scenario of 48kHz SVS.

3.3. Training methodology

XiaoiceSing2 is trained on 4 NVIDIA V100 GPUs with 32 batch size for 300 epochs until convergence, which costs 24 hours. We use Adam [29] optimizer with 0.01 learning rate, 0.9 β_1 , 0.98 β_2 , and 10^{-9} ϵ to train the both generator and discriminator. In addition, we adopt a warmup strategy that is identical to the one in [17] to adjust the learning rate for better optimization.

Table 1: MOS test result with 95% confidence interval of the ground truth and different acoustic models for 48kHz singing voice synthesis.

Vocoder	MOS(\uparrow)
Ground truth	4.27 \pm 0.044
XiaoiceSing + HiFi-WaveGAN	3.30 \pm 0.073
XiaoiceSing2 + HiFi-WaveGAN	4.23 \pm 0.044

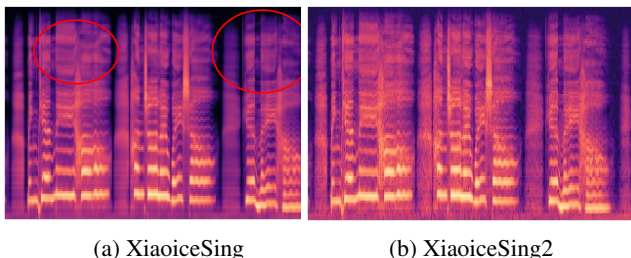


Figure 2: Mel-spectrograms generated by XiaoiceSing and XiaoiceSing2, respectively.

3.4. Subjective evaluation

To show the quality of the singing voices generated by XiaoiceSing2, we conduct a subjective evaluation for the real and synthesized audio. 20 listeners are asked to give their opinion score to 20 segments in terms of quality and naturalness of the singing voices, which indicates we collect 400 scores for the ground truth as well as each system. Table 1 summarized the evaluation result. Compared with the MOS of XiaoiceSing, the proposed XiaoiceSing2 significantly outperform it by over 0.93 in term of the MOS metric. And the number of the fluctuation in the 95% confidence interval shows XiaoiceSing2 is much more stable than XiaoiceSing when synthesizing high-fidelity singing voices. In addition, the MOS of XiaoiceSing2 is very close (-0.04) to the ground truth from the table, which means our model can synthesize the human-level singing voices in the 48kHz scenario.

3.5. Spectrogram analysis

Although the MOS test result indicates the quality of the singing voices generated by XiaoiceSing2 is much better than it of XiaoiceSing, it is necessary to find some evidence from the generated mel-spectrograms to support this conclusion. As Figure 2 shows, it seems the left mel-spectrogram generated by XiaoiceSing has more distinct spectral lines compared with the right mel-spectrogram generated by XiaoiceSing2. However, the distinct mel-spectrogram also indicates the severe over-smoothing problem for high-fidelity singing voice generation. Compared with Figure 2(a), Figure 2(b) obviously reserves more details in the transition regions between the adjacent spectral lines, which demonstrates the over-smoothing problem is alleviated. In addition, the over-smoothing problem in the high-frequency parts circled in Figure 2(a) also leads to audible hissing noise in the generated audio.

3.6. Ablation study

In this paper, we proposed multiple points to improve the quality of high-fidelity singing voices. It is reasonable to figure out

Table 2: Ablation study to show the contribution of the proposed components. The XiaoiceSing model of the first line predict MGC and BA rather than the mel-spectrogram. While other models predict the mel-spectrogram.

Vocoder	MOS(\uparrow)
XiaoiceSing + WORLD [3]	3.39 \pm 0.058
XiaoiceSing + HiFi-WaveGAN	3.30 \pm 0.073
(+ConvFFT) + HiFi-WaveGAN	3.33 \pm 0.072
(++SD) + HiFi-WaveGAN	4.20 \pm 0.043
(+++DD) + HiFi-WaveGAN	4.23 \pm 0.044

the contribution of each proposed component to the quality of the synthesized audio. Therefore, we conduct an ablation study to demonstrate the improvement of each component. As Table 2 shows, the first line of it indicates the result of the original XiaoiceSing system described in [3]. Instead of predicting the mel-spectrogram, it generates the mel-generalized cepstrum (MGC) and band aperiodicity (BA) for the WORLD vocoder [12]. The result of it is slightly better than the combination of XiaoiceSing described in Section 3.2 and HiFi-WaveGAN vocoder [21] because the model of the original XiaoiceSing learns more information from the training data.

When the ConvFFT module is incorporated into the XiaoiceSing as the third line shows, the MOS is boosted by 0.03, which means even only substituting the FFT in XiaoiceSing with the ConvFFT module, it is helpful to generate a better mel-spectrogram. Based on it, we change the sequence-to-sequence (S2S) model of XiaoiceSing to a GAN-based model by adding all segment discriminators as the fourth line shows. The MOS metric is largely promoted from 3.33 to 4.20 as expected, which proves that the GAN-based model has a huge advantage over the S2S model for high-fidelity singing voice synthesis. The last line in the table shows the result of the proposed XiaoiceSing2. By combining the segment and detail discriminators, the MOS is improved by 0.03 further because of the better construction of the middle- and high-frequency parts.

4. Conclusion

We propose a novel GAN-based acoustic model XiaoiceSing2 for SVS in this paper to relieve the over-smoothing problem in the middle- and high-frequency parts of the mel-spectrogram. In the FastSpeech-based generator, the new ConvFFT block combines the MHSA block and multiple residual convolutional blocks in parallel to couple the global and local information, which is beneficial to generate a more fine-grained mel-spectrogram as shown in the experiment. As for the discriminator, we extend the multi-band discriminator used in HiFiSinger by randomly clipping the mel-spectrogram into several segments so that the discriminator can increase the capability from the different long-term dependencies. Additionally, a detail discriminator accompanying the segment discriminator is used to pay more attention to middle- and high-frequency parts of the mel-spectrogram. The powerful discriminator also forces the generator to produce a more realistic mel-spectrogram. The experimental result on the 48kHz singing voice dataset proves that XiaoiceSing2 is able to generate high-quality mel-spectrogram, especially in middle- and high-frequency regions.

5. References

- [1] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6955–6959.
- [2] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, "Adversarially trained end-to-end korean singing voice synthesis system," *Proc. Interspeech 2019*, pp. 2588–2592, 2019.
- [3] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "Xiaoicesing: A high-quality and integrated singing voice synthesis system," *Proc. Interspeech 2020*, pp. 1306–1310, 2020.
- [4] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.
- [5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.
- [6] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [7] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021*, 2021.
- [8] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan," in *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.
- [9] Y. Wu, S. Li, C. Yu, H. Lu, C. Weng, L. Zhang, and D. Yu, "Peking opera synthesis via duration informed attention network," *Proc. Interspeech 2020*, pp. 1226–1230, 2020.
- [10] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Fast and high-quality singing voice synthesis system based on convolutional neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7239–7243.
- [11] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *arXiv preprint arXiv:2009.01776*, 2020.
- [12] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [13] R. Huang, C. Cui, F. Chen, Y. Ren, J. Liu, Z. Zhao, B. Huai, and Z. Wang, "Singgan: Generative adversarial network for high-fidelity singing voice generation," in *30th ACM International Conference on Multimedia (ACMMM)*, 2022. [Online]. Available: <https://doi.org/10.1145/3503161.3547854>
- [14] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting over-smoothness in text to speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8197–8213.
- [15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
- [16] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7229–7233.
- [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
- [20] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European conference on computer vision*. Springer, 2016, pp. 702–716.
- [21] W. Chunhui, Z. Chang, and H. Xing, "Hifi-wavagan: Generative adversarial network with auxiliary spectrogram-phase loss for high-fidelity singing voice generation," *arXiv preprint arXiv:2210.12740*, 2022.
- [22] P. Taylor, "Hidden markov models for grapheme to phoneme conversion," in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [24] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.
- [25] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.
- [26] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.
- [27] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [28] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>, 2021.
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>