



Period Singer: Integrating Periodic and Aperiodic Variational Autoencoders for Natural-Sounding End-to-End Singing Voice Synthesis

Taewoo Kim, Choonsang Cho, Young Han Lee

Korea Electronics Technology Institute, South Korea

{kimtaewoo, ideafisher, yhlee}@keti.re.kr

Abstract

In this paper, we present Period Singer, a novel end-to-end singing voice synthesis (SVS) model that utilizes variational inference for periodic and aperiodic components, aimed at producing natural-sounding waveforms. Recent end-to-end SVS models have demonstrated the capability of synthesizing high-fidelity singing voices. However, owing to deterministic pitch conditioning, they do not fully address the one-to-many problem. To address this problem, we present the Period Singer architecture, which integrates variational autoencoders for the periodic and aperiodic components. Additionally, our methodology eliminates the dependency on an external aligner by estimating the phoneme alignment through a monotonic alignment search within note boundaries. Our empirical evaluations show that Period Singer outperforms existing end-to-end SVS models on Mandarin and Korean datasets. The efficacy of the proposed method was further corroborated by ablation studies.

Index Terms: Singing voice synthesis, end-to-end model, variational autoencoder, pitch modeling, phoneme alignment

1. Introduction

Singing voice synthesis (SVS) aims to generate natural singing voices similar to those of humans based on music scores that include lyrics, note pitch, and note duration. In text-to-speech (TTS) tasks, there is significant uncertainty owing to the wide variation encompassing various emotions and intonations, as it depends solely on the text. By contrast, SVS tends to have relatively less uncertainty than TTS because it is guided by the note pitch and duration information of the music scores. However, modeling narrow variations based on the rich expression of the singing voice is crucial but challenging for achieving natural singing voices. Early deep learning-based SVS models [1, 2] adopted a two-stage approach using a parametric vocoder. For instance, XiaoiceSing [1] successfully generated natural singing voices by leveraging the FastSpeech [3] architecture and WORLD vocoder [4]. Subsequently, neural vocoder-based SVS models [5, 6, 7, 8, 9] were proposed because of the performance ceiling of parametric vocoders, enabling the generation of high-quality singing voices. However, the two-stage model can still be improved to address the mismatch in acoustic features between the training and inference stages.

To address this discrepancy, the end-to-end VISinger series [10, 11] has been introduced. VISinger [10] was the first end-to-end singing voice synthesis model based on VITS [12]. Specifically, they introduced the frame prior network to effectively model narrow variations within a phoneme in singing voices. The improved version, VISinger2 [11], could successfully alleviate pitch glitch issues while achieving high fidelity at a 44.1 kHz sampling rate. However, they did not fully solve

the one-to-many problem because the models were conditioned on deterministic pitches. Moreover, it requires ground truth phoneme duration information for training because it removes the monotonic alignment search to adopt the frame prior network.

In this study, we propose a novel end-to-end SVS model called Period Singer, which enables variational pitch modeling. To address the one-to-many problem, we integrate conditional variational autoencoders (CVAEs) for both the periodic and aperiodic components in the proposed model. Additionally, we apply the smoothed pitch augmentation method to ensure that the latent variables capture both wide and narrow pitch variations. During training, we utilized a normalizing flow to transform the aperiodic posterior distribution into a linguistic representation of a simpler distribution. This allows for the monotonic alignment search while adopting frame prior networks. To enhance the monotonic alignment search, we utilized note boundaries to prevent mismatches that lead to meaningless alignment between the singing voice and the music score. In the experiments, we evaluated our model on the public Mandarin dataset, Opencpop [13] and the internal Korean dataset, demonstrating superior performance compared to existing end-to-end models. Moreover, we validated the effectiveness of the proposed approach through ablation studies.

2. Proposed methods

The proposed model architecture builds on the foundation of VITS [12] by integrating two CVAEs. The proposed model comprises individual posterior encoders and decoders for both periodic and aperiodic components, a prior encoder, and discriminators. The posterior encoders extract latent variables for both the periodic and aperiodic components while being constrained by the prior encoder. In the decoding stage, periodic latent variables are utilized by the periodic decoder to predict fundamental frequency (F_0) sequences. The predicted F_0 sequences along with the aperiodic latent variables are conditioned on the aperiodic decoder to generate the waveform. Subsequently, the generated waveform was adversarially trained using discriminators. The overall architecture of Period Singer is shown in Figure 1, which illustrates both the training and inference processes.

2.1. Posterior encoding

We apply CVAEs separately to model the periodic and aperiodic components of singing voices, using two posterior encoders. First, the posterior encoder for the aperiodic component uses the mel-spectrogram processed from the waveform as the input, whereas the other posterior encoder uses the F_0 sequence as the input. Each posterior encoder is structured similarly to VITS,

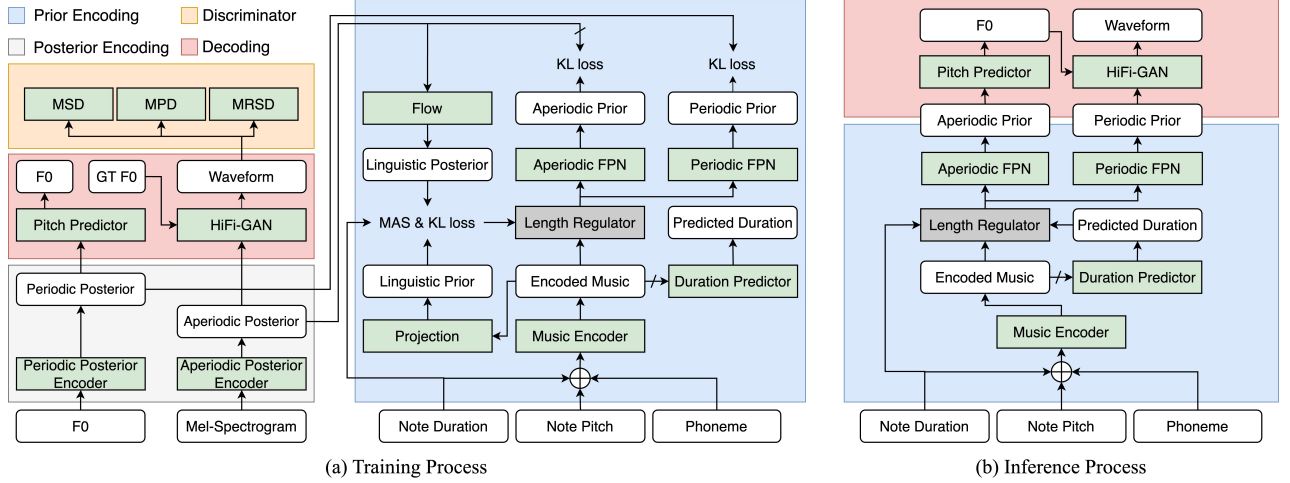


Figure 1: Overall architecture of Period Singer.

comprising non-causal WaveNet residual blocks [14] and a linear projection layer. It generates parameters for the mean and variance of each posterior distribution. They are then sampled using a reparameterization trick and passed to the decoder.

2.2. Prior encoding

The prior encoder derives the latent variables of the linguistic, periodic, and aperiodic prior distributions from the phoneme alignment A and the music score condition c , which includes the lyrics, note pitch, and note duration. For CVAEs, the periodic prior distribution $p(z_p|c, A; \theta_p)$ and the linguistic prior distribution $p(z_l|c, A, \theta_l)$ are regularized. In addition, the aperiodic prior distribution $p(z_a|c, A; \theta_a)$ is optimized to match the detached aperiodic posterior distribution.

As shown in Figure 1(a), the prior encoder structure comprises a music encoder, a duration predictor, a length regulator, frame prior networks (FPNs) for periodic and aperiodic components, and a flow module. Initially, the aperiodic posterior distribution $q(z_a|x_{mel}; \phi_a)$ is transformed into a simplified distribution $q(z_l|x_{mel}; \phi_l)$ for the linguistic component using the flow module. Then, the monotonic alignment search (MAS) [12, 15] is employed between the prior distribution $p(z_l|c, A, \theta_l)$ extracted from the music encoder and the projection layer, and the posterior distribution $q(z_l|x_{mel}; \phi_l)$ to estimate the phoneme alignment A . In the training process, the phoneme duration d , derived from the predicted phoneme alignment A , is utilized to expand the encoded music representation h and linguistic prior distribution $p(z_l|c, A, \theta_l)$ via the length regulator. Conversely, during inference, the predicted phoneme duration \hat{d} obtained from the duration predictor is used. Subsequently, the expanded encoded music representation h is used to extract prior distributions for periodic and aperiodic frames from the FPNs. The linguistic prior distribution $p(z_l|c, A, \theta_l)$ is also expanded for aperiodic CVAE training.

To regularize the periodic and aperiodic CVAEs, the Kullback-Leibler (KL) divergence losses $L_{kl,p}$ and $L_{kl,a}$ are calculated as follows:

$$L_{kl,p} = D_{kl}(q(z_p|F_0; \phi_p) || p(z_p|c, A; \theta_p)), \quad (1)$$

$$L_{kl,a} = D_{kl}(q(z_l|x_{mel}; \phi_l) || p(z_l|c, A; \theta_l)) + \lambda_l D_{kl}(q(\bar{z}_a|x_{mel}; \phi_a) || p(z_a|c, A; \theta_a)), \quad (2)$$

where x_{mel} , λ_l , and \bar{z}_a represent the input mel-spectrogram, the weighting factor and the detached aperiodic posterior latent variables, respectively. In addition, for the duration loss L_{dur} , we employ an L2 loss and apply the stop-gradient, as indicated by the short line in Figure 1(a).

2.2.1. Monotonic alignment search with note boundary

In the TTS task, the monotonic alignment search (MAS) has been successfully adopted in [12, 15] to estimate phoneme alignment without an external aligner. However, the alignment estimation in singing tasks is challenging because of the presence of breath sounds and diverse variations in pitch. This challenge arises from the end-to-end nature of the model, which incorporates reconstruction loss for segmented features, potentially leading to meaningless mappings between the singing voices and music scores. To address this issue, we perform MAS within note boundary information, representing the start and end frames of each note, and integrate it to accurately estimate the alignment. In particular, when there is only one phoneme per note, it is directly enforced in the alignment without the MAS.

2.2.2. Length regulator during inference

The duration predicted by the duration predictor does not guarantee perfect alignment with the length of the singing voice in rhythm. Hence, rhythm adjustment based on note duration is essential. This involves calculating the ratio of the phoneme durations within a single note and scaling them to match the note duration. This process enables the inference of an accurate singing voice synchronized with the rhythm.

2.3. Decoding

Recently, singing voice vocoders [16, 17, 18] have been proposed to alleviate the glitch caused by long vowels in singing voices, which are commonly conditioned on sine waveforms. Therefore, we employ the pitch-conditioned HiFi-GAN [18, 19] as the aperiodic decoder. The pitch predictor reconstructs F_0 from periodic latent variables, which are then conditioned by the aperiodic decoder. During training, the ground truth F_0 is conditioned. Consequently, we use the mel-spectrogram loss L_{mel} and the pitch loss L_{pitch} , both calculated using the L1

loss.

In the SVS task, because the main pitch is determined by the input notes, modeling narrow variations in pitch, such as bending and vibrato, is important. In this study, we propose a pitch augmentation method using smoothed F_0 to guide the periodic CVAE to learn narrow variations in pitch. Through the median filter, short-term bending and vibrato are removed, resulting in a smoothed F_0 contour. To guide the mean of the periodic latent variables for learning the smoothed F_0 , we concatenate the mean with the input batch of the aperiodic decoder and the smoothed F_0 with the target batch, respectively. Therefore, the pitch loss is modified as follows:

$$L_{pitch} = \|\log F_0 - \log \hat{F}_0\|_1 + \lambda_s \|\log f_s(F_0) - \log \tilde{F}_0\|_1, \quad (3)$$

where λ_s , $f_s(\cdot)$, \hat{F}_0 , and \tilde{F}_0 represent a weighting factor for the smoothed F_0 , the median filter, the predicted F_0 , and the predicted smoothed F_0 , respectively.

2.4. Discriminator

To generate high-fidelity singing voice, it is crucial to address high-frequency aliasing and artifacts. VISinger2 [11] addresses this issue by combining the multi-period discriminator (MPD) and multi-scale discriminator (MSD) from HiFi-GAN [20] with the multi-resolution spectrogram discriminator (MRSD) of UnivNet [21]. We also employ the same discriminator configuration, with the loss for adversarial training defined as follows:

$$L_{dis} = \mathbb{E}_{(y, z_a, F_0)} [(D(y) - 1)^2 + (D(G(z_a, F_0)))^2], \quad (4)$$

$$L_{adv} = \mathbb{E}_{(z_a, F_0)} [(D(G(z_a, F_0)) - 1)^2], \quad (5)$$

$$L_{fm} = \mathbb{E}_{(y, z_a, F_0)} \left[\sum_{l=1}^L \frac{1}{N_l} \|D^l(y) - D^l(G(z_a, F_0))\|_1 \right], \quad (6)$$

where z_a denotes the aperiodic posterior latent variable. L , D_l , and N_l represent the number of layers, the l -th layer, and the number of features in that layer of the discriminator, respectively.

2.5. Final loss

The final loss function for optimizing the proposed model, considering the periodic and aperiodic CVAEs, as well as GAN training objectives, is as follows:

$$L_{gen} = L_{adv} + \lambda_{fm} L_{fm} + \lambda_{mel} L_{mel} + \lambda_{pitch} L_{pitch} + \lambda_a L_{kl,a} + \lambda_p L_{kl,p} + \lambda_{dur} L_{dur} \quad (7)$$

where the λ coefficient serves as a weighting factor for each loss.

3. Experiments

3.1. Datasets

In the experiments, we evaluated the performance of Period Singer separately for two languages: Mandarin and Korean.

- **Mandarin:** We utilized the publicly available singing corpus, Opencpop [13], which comprises recordings from a single female singer totaling 100 songs and includes detailed annotations such as lyrics, notes, phoneme alignments, silence, aspirates, and slurs. We used predefined training and testing

sets from Opencpop, consisting of 95 songs (3,550 segments) and 5 songs (206 segments), respectively.

- **Korean:** We selected 100 Korean pop songs and collected singing datasets recorded by a female artist. The dataset was divided into 95 songs (2,461 segments) for training and 5 songs (127 segments) for testing, with each segment ranging from 3 to 12 seconds. It included lyrics and MIDI notes manually annotated by human annotators.

All audio files across both datasets were sampled at 44.1 kHz with 16-bit quantization. In the preprocessing stage, an 80-mel-spectrogram was extracted with settings of 2048 for the FFT size, 2048 for the window size, and 512 for the hop size. The F_0 extraction step was performed using the harvest [22] algorithm from the WORLD toolkit [4]. For pitch augmentation, we extracted the smoothed F_0 using a median filter with a kernel size of 13.

3.2. Model configurations

The proposed Period Singer architecture was developed based on VITS [12], incorporating additional components such as the periodic and aperiodic FPNs, the periodic posterior encoder, and the pitch predictor. The FPNs for periodic and aperiodic components consisted of 6 Conformer blocks [23] and a projection layer. Each Conformer block was set with 2 heads, a hidden size of 192, and a kernel size of 31. Posterior encoders were composed of 8 Wavenet residual blocks [14] with a kernel size of 5. The mean and variance of all the distributions were 192-dimensions. The structure of the pitch predictor mirrored that of the posterior encoder, comprising 4 Wavenet residual blocks and a projection layer, with the output size set to 1. Our decoder module followed the pitch-conditioned HiFiGAN architecture of [19], with 256 hidden dimensions. To extend the proposed model to 44.1 kHz, the upsampling stride and kernel sizes of the HiFi-GAN module [20], originally set at 22.05 kHz, were adjusted to (8,8,4,2) and (16,16,8,4) respectively. The discriminator model configuration was the same as that of VISinger2, while the remaining hyperparameters were consistent with those of VITS. For the weights in the final loss, we set λ_l , λ_s , λ_{fm} , λ_{mel} , λ_{pitch} , λ_a , λ_p and λ_{dur} to 1, 1, 2, 45, 10, 1, 1, and 1, respectively.

During training, we used the AdamW optimizer [24] with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. The learning rate decayed exponentially by a factor of $0.999^{1/8}$ for each epoch with an initial learning rate of 0.0002. The proposed model was trained for 400k iterations with a batch size of 8.

3.3. Baseline systems

For comparison, we trained the following baseline systems:

- **VITS:** Original VITS trained with music score input. We utilized the official implementation¹ with slight modifications for music score input.
- **VISinger:** The end-to-end SVS system based on VITS. It predicts deterministic F0 and conditions it on the frame prior network. We utilized the unofficial implementation².
- **VISinger2:** The improved system of VISinger that predicts the deterministic F0 and mel-spectrogram, and conditions it on the frame prior network and the decoder. We utilized the official implementation³, setting the number of the FFT block

¹<https://github.com/jaywalnut310/vits>

²<https://github.com/jerryuhoo/VISinger>

³<https://github.com/zhangyongmao/VISinger2>

Table 1: The MOS test results with 95% confidence intervals

Model	Mandarin	Korean
VITS	3.67 ± 0.08	3.78 ± 0.07
VISinger	3.71 ± 0.08	3.88 ± 0.07
VISinger2	3.91 ± 0.07	4.45 ± 0.06
Period Singer (DPP)	3.90 ± 0.07	4.56 ± 0.05
Period Singer	3.97 ± 0.07	4.61 ± 0.05
Recording	4.24 ± 0.07	4.65 ± 0.05

Table 2: The CMOS test results for ablation study

Model	CMOS
Period Singer	0
- Normalizing Flow	-0.21
- Note boundary for MAS	-1.13

to 6 instead of the configured 4 as in [11].

- **Period Singer (DPP)**: The Period Singer model with deterministic pitch prediction (DPP). The periodic frame prior network directly predicts F_0 and conditions it on the decoder.

All baseline systems were extended to a sampling rate of 44.1 kHz and trained with the same discriminator and configuration as the Period Singer. To train VISinger and VISinger2 on the Korean dataset, we obtained phoneme durations using the proposed alignment estimation, which utilized note boundary information from a pre-trained VITS.

3.4. Results

We conducted a mean opinion score (MOS) test to assess the overall performance, encompassing the sound quality and naturalness, of our Period Singer and baseline systems. For the evaluation in Mandarin, 20 testers engaged via Amazon Mechanical Turk, whereas 19 Korean testers participated in the Korean assessment. We provided 20 samples per system, and each tester rated 120 samples across the 6 systems, including the recordings. All testers were asked to evaluate the singing voice on a scale of 1 to 5 with intervals of 0.5.

Table 1 summarizes the MOS results. For both datasets, the scores for Period Singer are 3.97 and 4.61, respectively, indicating superior performance compared to all other systems. In particular, for the Korean dataset, Period Singer obtains scores that reached the confidence intervals of the recordings, demonstrating the sound quality and naturalness close to human recordings. Period Singer (DPP) shows slightly lower performance compared to VISinger2 by 0.01 for the Mandarin dataset. On the other hand, it exhibits better performance for the Korean dataset. These outcomes are acceptable given the relatively lower diversity of singing voices in the Openpop dataset compared to the Korean dataset. Moreover, these results are remarkable considering that the Period Singer models were trained without ground truth phoneme duration.

3.5. Ablation study

To verify the effectiveness of the method proposed in Period Singer, we conducted ablation studies. First, we compared the naturalness with and without the normalizing flow module and note boundary for the MAS. To this end, we prepared 20 samples for each system and requested an additional comparative

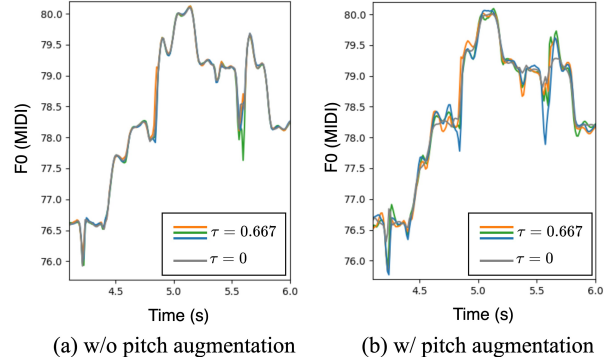


Figure 2: Variational pitch inference of Periodic Singer with and without pitch augmentation.

MOS (CMOS) from the 19 Korean testers. They rated other systems, compared to Period Singer, on a scale of -3 to 3, with intervals of 1.

As shown in Table 2, removing the normalizing flow module results in a decrease of 0.21 points in performance. This suggests that a simpler linguistic distribution, rather than performing MAS with an aperiodic distribution, is more effective for predicting an accurate alignment. Furthermore, when using the original MAS without note boundaries, the performance is 1.13 points lower compared to the proposed method. This indicates that constraining MAS with note boundaries leads to a stable alignment prediction, avoiding meaningless mappings between the singing voice and the music score. Some audio samples are available online⁴.

To confirm the impact of variational pitch inference on pitch augmentation, we compared the results obtained with and without the proposed pitch augmentation. We visualized the results in Figure 2 by setting the temperature τ , which served as the scale factor for variance, to 0.667 and 0, respectively, and conducting the inference four times and once for each setting. Figure 2(a) illustrates that for Period Singer without pitch augmentation, the results are similar when τ is 0 and 0.667, indicating a heavily sparse variance. However, as shown in Figure 2(b), when τ is zero, the predictions tend to cluster around the mean pitch, whereas, at 0.667, we observed a more diverse range of pitches compared with Figure 2(a), indicating a broader distribution of variance.

4. Conclusions

In this study, we introduced Period Singer, a novel end-to-end SVS model designed to produce natural and high-fidelity waveforms. By incorporating variational inference for both periodic and aperiodic components, Period Singer addresses the limitations of previous SVS models, particularly the one-to-many problem stemming from deterministic pitch conditioning. Our architecture not only enhances the synthesis quality but also eliminates the need for an external aligner by estimating the phoneme alignment within the note boundary information. Empirical evaluations on Mandarin and Korean datasets demonstrate that Period Singer surpasses existing end-to-end SVS models.

⁴<https://rlataewoo.github.io/periodsinger>

5. Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00963 and No.2022-0-00608).

6. References

- [1] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System," in *Proc. Interspeech 2020*, 2020, pp. 1306–1310.
- [2] X. Zhuang, T. Jiang, S.-Y. Chou, B. Wu, P. Hu, and S. Lui, "Litesing: Towards fast, lightweight and expressive singing voice synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7078–7082.
- [3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [4] 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.
- [5] G.-H. Lee, T.-W. Kim, H. Bae, M.-J. Lee, Y.-I. Kim, and H.-Y. Cho, "N-Singer: A Non-Autoregressive Korean Singing Voice Synthesis System for Pronunciation Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 1589–1593.
- [6] T.-W. Kim, M.-S. Kang, and G.-H. Lee, "Adversarial Multi-Task Learning for Disentangling Timbre and Pitch in Singing Voice Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 3008–3012.
- [7] Z. Zhang, Y. Zheng, X. Li, and L. Lu, "Wesinger 2: fully parallel singing voice synthesis via multi-singer conditional adversarial training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [8] W. Chunhui, C. Zeng, and X. He, "XiaoiceSing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network," in *Proc. INTERSPEECH 2023*, 2023, pp. 5401–5405.
- [9] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.
- [10] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7237–7241.
- [11] Y. Zhang, H. Xue, H. Li, L. Xie, T. Guo, R. Zhang, and C. Gong, "VISinger2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer," in *Proc. INTERSPEECH 2023*, 2023, pp. 4444–4448.
- [12] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [13] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Openpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 4242–4246.
- [14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
- [15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
- [16] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-filter hifi-gan: Fast and pitch controllable high-fidelity neural vocoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [17] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2525–2535.
- [18] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Period-hifi-gan: Fast and fundamental frequency controllable neural vocoder," *Proc. Acoustical Society of Japan, in Japanese*, pp. 901–904, 2022.
- [19] Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J.-M. Kim, and K. Tachibana, "Period vits: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [20] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [21] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. ISCA, 2021, pp. 2207–2211.
- [22] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2321–2325.
- [23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.