VISINGER 2: HIGH-FIDELITY END-TO-END SINGING VOICE SYNTHESIS ENHANCED BY DIGITAL SIGNAL PROCESSING SYNTHESIZER

Yongmao Zhang[1], *Heyang Xue*[1], *Hanzhao Li*[1], *Lei Xie*[1*], *Tingwei Guo*[2], *Ruixiong Zhang*[2], *Caixia Gong*[2]

[1]Audio, Speech and Language Processing Group (ASLP@NPU)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]DiDi Chuxing, Beijing, China

ABSTRACT

End-to-end singing voice synthesis (SVS) model VISinger [1] can achieve better performance than the typical two-stage model with fewer parameters. However, VISinger has several problems: *text-to-phase problem*, the end-to-end model learns the meaningless mapping of text-to-phase; *glitches problem*, the harmonic components corresponding to the periodic signal of the voiced segment occurs a sudden change with audible artefacts; *low sampling rate*, the sampling rate of 24KHz does not meet the application needs of high-fidelity generation with the full-band rate (44.1KHz or higher). In this paper, we propose VISinger 2 to address these issues by integrating the digital signal processing (DSP) methods with VISinger. Specifically, inspired by recent advances in differentiable digital signal processing (DDSP) [2], we incorporate a DSP synthesizer into the decoder to solve the above issues. The DSP synthesizer consists of a harmonic synthesizer and a noise synthesizer to generate periodic and aperiodic signals, respectively, from the latent representation z in VISinger. It supervises the posterior encoder to extract the latent representation without phase information and avoid the prior encoder modelling text-to-phase mapping. To avoid glitch artefacts, the HiFiGAN is modified to accept the waveforms generated by the DSP synthesizer as a condition to produce the singing voice. Moreover, with the improved waveform decoder, VISinger 2 manages to generate 44.1kHz singing audio with richer expression and better quality. Experiments on OpenCpop corpus [3] show that VISinger 2 outperforms VISinger, CpopSing and RefineSinger in both subjective and objective metrics. Our audio samples are available on the demo website [1], and we will release our source code upon the acceptance of this paper.

***Index Terms*—** Singing voice synthesis, variational autoencoder, adversarial learning

1. INTRODUCTION

Singing voice synthesis (SVS) is a task that generates singing voices from the given music score and lyrics like human singers. Deep learning based SVS approaches [4, 5, 6, 7, 8, 9] have attracted tremendous attention in recent years for their extraordinary performances and wide applications. Similar to text-to-speech (TTS), most of these SVS systems consist of two stages, the acoustic model first generates low-dimensional spectral representations of vocal signals, typically mel-spectrogram, from the music score and lyrics, and the vocoder subsequently converts these intermediate representations into the singing waveform. Although these systems achieve decent performances, the two-stage models are separately trained, and the human-crafted intermediate representations, such as the mel-spectrogram, may limit the expressiveness of the synthesized singing voice.

We have recently proposed VISinger [1] – an end-to-end (E2E) learned SVS approach based on VITS [10] to mitigate the problems of two-stage systems. Specifically, VITS adopts the structure of CVAE to realize end-to-end speech synthesis. The posterior encoder extracts the latent representation z from the linear spectrum, the decoder restores z to the waveform, and the prior encoder provides a prior constraint for z according to the text. To better model singing, VISinger provides z with more accurate frame level prior constraints under the guidance of F0 and provides extra prior information for the duration predictor. VISinger achieves superior performance over the typical two-stage systems such as Fastspeech [11] + HiFiGAN [12].

Although VISinger advances the end-to-end SVS, it still has some drawbacks preventing its further application in real-world applications. First, the quality artefacts of the two-stage systems still exist in VISinger. Specifically, the audible glitches, such as spectral discontinuities and occasional mispronunciations, reduce the naturalness of the generated singing voice. Second, the sampling rate of the generated singing voice of VISinger is 24KHz, which does not meet the needs of high-fidelity (HiFi) applications which desire full-band audio (44.1KHz or higher).

To address these inadequacies, we reanalyzed the architecture and components of the VISinger. The first and most significant issue is that the latent representation z extracted by the posterior encoder may contain phase information due to the gradients passed back by the decoder when modelling the waveform. This could lead to mispronunciation because it is extremely challenging to predict the phase from the linguistic input reasonably. Secondly, the HiFiGAN [12] architecture adopted in VISinger is not well designed for the SVS task. Its absence of modelling capabilities of rich variations on singing voice may lead to the glitches problem. Finally, a higher sampling rate SVS system relies on an improved decoder to provide better modelling capabilities.

In this paper, we propose VISinger 2, a digital signal processing (DSP) synthesizer enhanced end-to-end SVS system for high-fidelity 44.1KHz singing generation. Specifically, inspired by recent advances in differentiable digital signal processing (DDSP) [2], we incorporate a DSP synthesizer into VISinger to solve the above issues. Specifically, the DSP synthesizer consists of a harmonic synthesizer and a noise synthesizer to generate periodic and aperiodic signals from the latent representation z, respectively. The periodic and aperiodic signals are concatenated as conditional inputs to HiFiGAN, while the sum of the two produces a waveform to calculate the loss function. This design has sufficient advantages. First, both synthesizers need only amplitude information as input to generate the signals, thus fully compressing the phase component in z and

* Corresponding author.

[1]Demo: `https://zhangyongmao.github.io/VISinger2/`

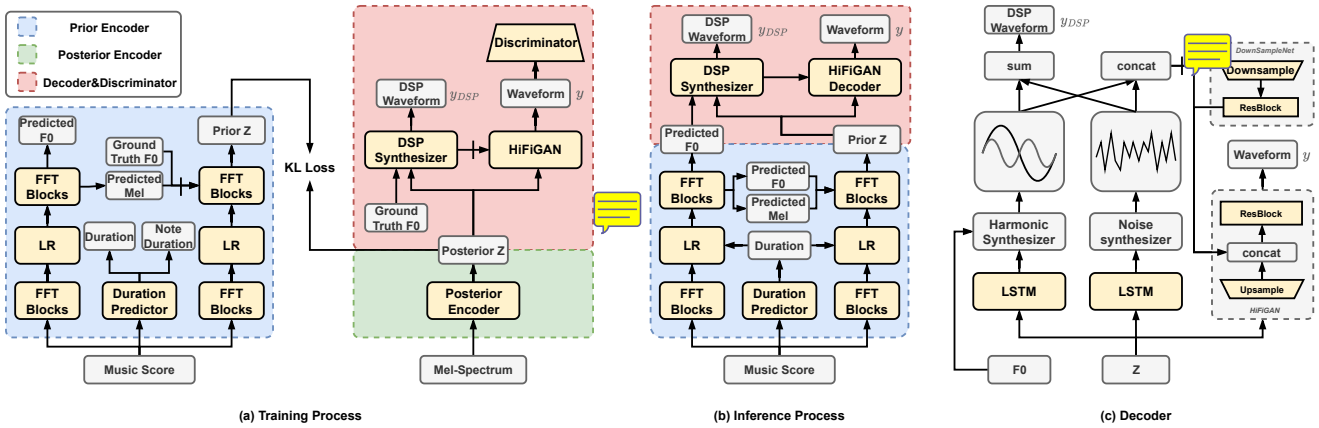

Fig. 1. Architecture of VISinger 2. Yellow components are part of the neural network architecture, and grey components are features or differentiable operations. The short line on the arrow indicates gradient truncation.

avoiding the text-to-phase challenge. Second, the representation of the periodic and aperiodic signal composition provides a strong condition for HiFi-GAN, substantially enhancing its modelling capability and allowing it to model a higher sampling rate. Finally, due to these improved modelling capabilities, the number of parameters in VISinger 2 can be substantially reduced by about 30% compared to VISinger, further facilitating its use in real-world applications. Experiments show that VISinger 2 can generate a high-fidelity singing voice at a 44.1kHz sampling rate, with better naturalness and fewer glitches than VISinger and the traditional two-stage system.

We notice that there has been a recent trend to leverage the advances of conventional DSP to neural audio generation [13, 14, 15]. For example, in [15], harmonic signals are used to improve the stability of GAN and avoid pitch jitters and U/V errors in singing voice conversion. RefineGAN [13] calculates the speech template according to the pitch and then generates waveform according to the speech template. SingGAN [14] adopts the source excitation with the adaptive feature learning filters to alleviate the glitch problem. These works usually focus on the periodic signal because the glitches problem comes from the defect of the periodic signal. Although motivated by these works aiming for better generation quality, our approach has substantial differences in terms of methodology. First, the above revisions are all made on vocoders, and the whole system still faces the two-stage mismatch problem. We mitigate this problem by proposing a fully end-to-end system VISinger 2. Second, to ensure that the extracted latent representation z in VISinger 2 contains full amplitude information (periodic and aperiodic parts), we leverage both periodic and aperiodic signals generated by the DSP synthesizer in our system design.

2. METHOD

The overall model architecture of VISinger 2 is shown in Fig. 1. The proposed model adopts the conditional variational autoencoder (CVAE) structure, which includes three parts: a posterior encoder, a prior encoder and a decoder, the same as VITS [10] and VISinger [1]. The posterior encoder extracts the latent representation z from spectral features, the decoder generates waveform y from z, and the prior conditional encoder constrains the extraction process of z. We will introduce the posterior encoder, decoder and prior encoder, respectively.

2.1. Posterior Encoder

The posterior encoder is composed of multi-layer 1-D convolution, which aims to extract the latent representation z from the mel-spectrum. The last layer produces the mean and variance of the posterior distribution, and the resampling method is used to obtain the posterior z.

2.2. Decoder

The decoder generates waveform from the latent representation z as shown in Fig.1(c). To avoid text-to-phase and glitches problems, we incorporate a DSP synthesizer into the decoder. Specifically, we use a harmonic synthesizer and a noise synthesizer to generate periodic and aperiodic parts of the waveform from the posterior z. The generated waveforms are used as an auxiliary condition for HiFi-GAN as input to enhance its modelling capabilities relieving the glitch problem. Meanwhile, since the inputs of both two synthesizers contain only amplitude information, the posterior z will lean towards not including phase information and thus alleviate the text-to-phase problem.

2.2.1. *Harmonic Synthesizer*

We use the harmonic synthesizer to generate harmonic components of audio the same as the harmonic oscillator in DDSP [2]. The harmonic synthesizer uses sin signals to simulate the waveform of each formant of the single sound source audio. The k-th sinusoidal component signal y_k generated by the harmonic synthesizer can be expressed as:

$$y_k(n) = H_k(n)sin(\phi_k(n)) \tag{1}$$

where n represents the time step of the sample sequence, and H_k is the time-varying amplitude of the k-th sinusoidal component. The phase $\phi_k(n)$ is obtained by integrating on the sample sequence:

$$\phi_k(n) = 2\pi \sum_{m=0}^{n} \frac{f_k(m)}{Sr} + \phi_{0,k} \tag{2}$$

where f_k represents the frequency of the k-th sinusoidal component, Sr represents the sampling rate, and $\phi_{0,k}$ represents the initial phase. We can get the phase of the sin signal y_k through an accumulation operation according to the fundamental frequency f_k. The frequency f_k can be calculated by $f_k(n) = kf_0(n)$, where f_0 is the fundamental frequency. The time-varying f_k and H_k are interpolated from frame-level features. We extract the fundamental frequency using Harvest [16] algorithm.

2.2.2. *Noise Synthesizer*

In the noise synthesizer, we use inverse short-time Fourier transform (iSTFT) to generate the stochastic components of audio, similar to the filtered noise in DDSP. The aperiodic components are

closer to noise, but the energy distribution is uneven in different frequency bands. The stochastic component signal y_{noise} generated can be expressed as:

$$y_{noise} = iSTFT(N, P) \tag{3}$$

where the phase spectrogram P of iSTFT is uniform noise in domain $[-\pi, \pi]$, and the amplitude spectrogram N is predicted by the network.

2.2.3. Loss Function of Decoder

The DSP waveforms generated by the DSP synthesizer contain both harmonic and stochastic components. The complete DSP waveform y_{DSP} and the loss L_{DSP} of the DSP synthesizer are defined as

$$y_{DSP} = \sum_{k=0}^{K} y_k + y_{noise} \tag{4}$$

$$L_{DSP} = \lambda_{DSP} \left\| \text{Mel}(y_{DSP}) - \text{Mel}(y) \right\|_1 \tag{5}$$

where K represents the number of the sinusoidal component and Mel represents the process of extracting mel-spectrum from waveform.

We use a downsampling network gradually downsamples the DSP waveforms to the frame-level features. The HiFi-GAN accepts the posterior z and the intermediate features generated by the downsampling network as input and generates the final waveform $\hat{y}$. Following HiFi-GAN, the GAN loss for the generator G is defined as:

$$L_G = L_{adv}(G) + \lambda_{fm} Lfm + \lambda_{Mel} L_{Mel} \tag{6}$$

where L_{adv} is the adversarial loss, L_{fm} is the feature matching loss, and L_{Mel} is the Mel-Spectrogram loss.

2.2.4. Discriminator

We combine two sets of discriminators to improve the ability of the discriminator. One set of discriminators is multi-resolution spectrogram discriminator (MRSD) in UnviNet [17], and the other is Multi-Period Discriminator (MPD) and Multi-Scale Discriminator (MSD) in HiFi-GAN [12].

2.3. Prior Encoder

The prior encoder takes the music score as input to provide a prior constraint for CVAE. As mentioned in Section 2.2, the posterior z will be used to predict H, N in the decoder, where H represents the amplitude of the sinusoidal formant and N represents the amplitude spectrum of aperiodic components. Both H and N only contain amplitude information but not phase information, so the posterior z will not contain phase information accordingly. In this way, the prior encoder will not model the text-to-phase mapping when predicting the posterior z based on the music score.

Similar to VISinger [1], the prior encoder adopts the same structure as Fastspeech [11]. The flow [18] module plays an important role in VITS [10], but it occupies a large number of model parameters. For a more practical structure, we calculate the KL divergence L_{kl} directly between the prior z and the posterior z without using flow.

We use a separate FastSpeech [11] model to predict the fundamental frequency and mel-spectrum to guide the frame-level prior networks. The loss for the auxiliary feature is defined as:

$$L_{af} = \left\| LF0 - \widehat{LF0} \right\|_2 + \left\| Mel - \widehat{Mel} \right\|_1 \tag{7}$$

where $\widehat{LF0}$ is the predicted log-F0, and $\widehat{Mel}$ is the predicted mel-spectrogram.

We take the predicted mel-spectrum as the auxiliary feature for the frame-level prior network in the training and inference process, so the auxiliary mel-spectrum does not bring a mismatch in the training and inference process. The frame-level prior network predicts the prior z with the guide of auxiliary mel-spectrum to alleviate the text-to-phase problem further. We prove later in the experiment that VISinger 2 does not rely too much on this auxiliary mel-spectrum. The harmonic synthesizer accepts the predicted fundamental frequency as input to guide the generation of periodic signals in the inference process, while the ground-truth fundamental frequency is adopted in the training process.

The duration predictor accepts the music score as input and adopts the method in XiaoiceSing [8] to simultaneously predict phoneme duration and note duration. The duration loss is expressed as:

$$L_{dur} = \left\| d_{phone} - \widehat{d_{phone}} \right\|_2 + \left\| d_{note} - \widehat{d_{note}} \right\|_2 \tag{8}$$

where d_{phone} is the ground truth phoneme duration, $\widehat{d_{phone}}$ is the predicted phoneme duration, while d_{note} is the ground truth note duration, and $\widehat{d_{note}}$ is the predicted note duration.

2.4. Final Loss

Our final objectives for the proposed model can be expressed as:

$$L(G) = L_G + L_{kl} + L_{DSP} + L_{dur} + L_{af} \tag{9}$$

$$L(D) = L_{adv}(D) \tag{10}$$

where L_G is the GAN loss for generator G, L_{kl} is KL divergence between prior z and posterior z, L_{af} is the loss of the auxiliary feature, and $L_{adv}(D)$ is the GAN loss of discriminator D.

3. EXPERIMENTS

3.1. Datasets

We evaluate VISinger 2 on the Opencpop [3] dataset, which consists of 100 popular Mandarin songs (5.2 hours) performed by a female professional singer. All the audios are recorded at 44.1kHz with 16-bit quantization. Opencpop has a pre-defined training set and test set: 3,550 segments from 95 songs for training while 206 segments from 5 songs for the test. We follow Opencpop's division of the training and test set.

3.2. Model Configuration

We train the following systems for comparison.

- **CpopSing**: the two-stage conformer-based SVS model introduced in the Opencpop [3]. In the CpopSing, the Transformer blocks in Fastspeech 2 [19] are replaced with Conformer blocks. The adversarial training method similar to the sub-frequency adversarial loss in HiFiSinger [20] is used in the CpopSing.
- **VISinger**: an end-to-end SVS system based on VITS. The model configuration is consistent with that in VISinger [1].
- **RefineSinger**: a two-stage SVS system constructed by FastSpeech [11] and RefineGAN [13]. The FFT block in both the encoder and decoder of Fastspeech are 4-layer. The duration predictor consists of a 3-layer 1D-convolutional network and predicts the phoneme-level and note-level duration. RefineGAN, which is designed for high sampling rate scenarios, adopts pitch-guided architecture to improve the ability of the generator. A Mel2F0 module introduced in [21] is used to predict the F0 for RefineGAN. The hidden dimension of RefineGAN is 512, and the data augmentation method proposed in [13] is not employed for simplicity.

Table 1. Experimental results in terms of subjective mean opinion score (MOS) and two objective metrics.

Model	Sample Rate	Model Size (M)	F0 RMSE	Dur RMSE	MOS
Cpopsing	22k	137.5	28.5	6.6	2.97±0.12
VISinger	22k	36.5	33.7	3.6	3.46±0.13
VISinger 2	22k	**25.7**	**26.0**	2.8	3.69 ±0.15
RefineSinger	44k	36.0	39.1	2.8	2.85±0.10
VISinger 2	44k	**25.7**	26.7	**2.7**	**3.81±0.14**
Recording	22k	-	-	-	4.22±0.12
Recording	44k	-	-	-	4.32±0.11

- **VISinger 2**: the proposed end-to-end SVS system, adopting all the contributions introduced in the paper. Each FFT blocks in VISinger2 consist of 4-layer FFTs. The hidden dim and filter dim of FFT are 192 and 768, respectively. The hidden dimension of HiFi-GAN in the decoder is 256. The posterior encoder consists of an 8-layer 1D-convolutional network, and the dimension of potential representation z is 192. The duration predictor consists of a 3-layer 1D-convolutional network with ReLU activation.

All models are trained up to 500k steps with a batch size of 16. The Adam optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and $\epsilon = 10^{-9}$ is used to train all the models.

3.3. Experimental Results

We performed a mean opinion score (MOS) test for the above systems and randomly selected 30 segments from the test set for subjective listening, and ten listeners attended the test. The objective metrics, including F0 Root Mean Square Error (F0-RMSE) and duration Root Mean Square Error (dur-RMSE), are calculated to evaluate the performance of different systems. The results are summarized in Table 1.

To evaluate the performance of the proposed VISinger 2 in a general SVS scenario, we first compared VISinger 2 with CpopSing and VISinger at the 22.05kHz sampling rate. As shown in Table 1, VISinger 2 and VISinger perform significantly better than CpopSing in the MOS test, demonstrating the superiority of the end-to-end model in the general SVS scenario. Meanwhile, the MOS score of VISinger 2 is higher than VISinger by about 0.23, indicating the effectiveness of our design in a general SVS scenario. For further validation of the performance of VISinger2 in high sampling rate SVS scenarios, we compared VISinger 2 with RefineSinger at the 44.1 kHz sampling rate. The evaluation results listed in Table 1 show that VISinger 2 surpasses RefineSinger in MOS score by 33.6% and has a MOS improvement of about 0.15 compared to the 22.05 kHz version of VISinger 2. This improvement shows that VISinger 2 is capable of modelling high sampling rates SVS enables high-fidelity singing voice generation. Note that CpopSing and VISinger did not participate in the 44.1kHz comparison for fairness as they are not designed for high sampling rate SVS. Similar to the MOS results, VISinger 2 outperformed the other systems in terms of objective metrics, validating our assumptions again.

Another observation worth highlighting is that in addition to outperforming the other systems in MOS and objective metrics, VISinger 2 has the smallest number of parameters in all comparison systems at 25.7M. This result demonstrates the effectiveness of our proposed approach and its sufficiency to be applied in real-world scenarios.

We further visualize the waveforms generated by VISinger 2 in Fig. 2 to illustrate the role of the DSP synthesizer. As shown in Fig. 2, the periodic components and aperiodic components are generated by the harmonic synthesizer and noise synthesizer, respectively. The generated periodic and aperiodic components are added to get DSP waveform y_{DSP}. We can also find that the waveform finally generated by HiFi-GAN is guided by the DSP waveform as its conditional input.

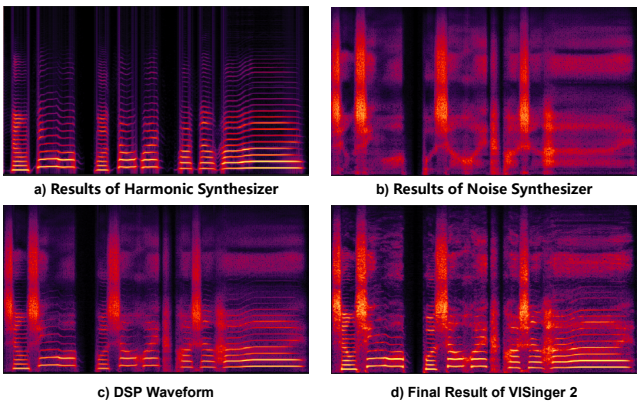

Fig. 2. Visualization of synthesized waveform.

Table 2. Ablation study results in terms of subjective mean opinion score (MOS).

Model	Sample Rate	MOS
Recording	44k	4.47±0.09
VISinger2	44k	3.96±0.11
-auxiliary mel-spectrum	44k	3.85±0.12
-DSP synthesizer	44k	3.02±0.13

3.4. Ablation study

To validate the effectiveness of each contribution, we conduct an ablation study. We remove the DSP synthesizer and auxiliary mel-spectrum feature, respectively. The results are summarized in Table 2. The results show that the model's performance degrades significantly when the DSP synthesizer is deleted, indicating that the DSP synthesizer plays an essential role in solving the text-to-phase problem and glitches problems. At the same time, when the auxiliary mel-spectrum feature is deleted, the model's performance degrades slightly, indicating that the auxiliary mel-spectrum can further solve the text-to-phase problem because a complete mel-spectrum guides the prediction of the prior z.

4. CONCLUSIONS

In this work, we have updated our previous end-to-end singing voice synthesis system VISinger to its new version VISinger 2. Specifically, we solved the text-to-phase problem and the glitch artefacts problem and upgraded the sampling rate from 24KHz to 44.1KHz for a high-fidelity singing generation. These new contributions were achieved by incorporating a differential digital signal processing (DDSP) synthesizer with the VISinger decoder. In this way, the posterior encoder extracts the latent representation without phase information and avoids the prior encoder modelling text-to-phase mapping. To avoid glitch artefacts, we modified the decoder to accept the waveforms generated by the DSP synthesizer as a condition to produce the singing voice. Our experimental results show that, with fewer model parameters, VISinger 2 substantially outperforms CpopSing, VISinger and RefineSinger.

5. REFERENCES

[1] Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 2022, pp. 7237–7241, IEEE.

[2] Jesse H. Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: differentiable digital signal processing," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2020, OpenReview.net.

[3] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi, "Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 4242–4246, ISCA.

[4] Yi Ren, Xu Tan, Tao Qin, Jian Luan, Zhou Zhao, and Tie-Yan Liu, "Deepsinger: Singing voice synthesis with data mined from the web," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[5] Merlijn Blaauw and Jordi Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] Yukiya Hono, Shumma Murata, Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "Recent development of the dnn-based singing voice synthesis system—sinsy," in *Proceedings, APSIPA Annual Summit and Conference*, 2018.

[7] Yu Gu, Xiang Yin, Yonghui Rao, Yuan Wan, Benlai Tang, Yang Zhang, Jitong Chen, Yuxuan Wang, and Zejun 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)*, 2021.

[8] Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou, "Xiaoicesing: A high-quality and integrated singing voice synthesis system," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. 2020, pp. 1306–1310, ISCA.

[9] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. 2022, pp. 11020–11028, AAAI Press.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540, PMLR.

[11] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 3165–3174.

[12] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[13] Shengyuan Xu, Wenxiao Zhao, and Jing Guo, "Refinegan: Universally generating waveform better than ground truth with highly accurate pitch and intensity responses," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 1591–1595, ISCA.

[14] Feiyang Chen, Rongjie Huang, Chenye Cui, Yi Ren, Jinglin Liu, Zhou Zhao, Nicholas Jing Yuan, and Baoxing Huai, "Singgan: Generative adversarial network for high-fidelity singing voice generation," *CoRR*, vol. abs/2110.07468, 2021.

[15] Haohan Guo, Zhiping Zhou, Fanbo Meng, and Kai Liu, "Improving adversarial waveform generation based singing voice conversion with harmonic signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 2022, pp. 6657–6661, IEEE.

[16] Masanori 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*, Francisco Lacerda, Ed. 2017, pp. 2321–2325, ISCA.

[17] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae 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*. 2021, pp. 2207–2211, ISCA.

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real NVP," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[19] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[20] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *CoRR*, vol. abs/2009.01776, 2020.

[21] Heyang Xue, Xinsheng Wang, Yongmao Zhang, Lei Xie, Pengcheng Zhu, and Mengxiao Bi, "Learn2sing 2.0: Diffusion and mutual information-based target speaker SVS by learning from singing teacher," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 4267–4271, ISCA.