SingGAN: Generative Adversarial Network For High-Fidelity Singing Voice Generation

Feiyang Chen*, Rongjie Huang*, Chenye Cui, Yi Ren, Jinglin Liu, Zhou Zhao†
Zhejiang University

Abstract

High-fidelity singing voice synthesis is challenging for neural vocoders due to extremely long continuous pronunciation, high sampling rate and strong expressiveness. Existing neural vocoders designed for text-to-speech cannot directly be applied to singing voice synthesis because they result in glitches in the generated spectrogram and poor high-frequency reconstruction. To tackle the difficulty of singing modeling, in this paper, we propose SingGAN, a singing voice vocoder with generative adversarial network. Specifically, 1) SingGAN uses source excitation to alleviate the glitch problem in the spectrogram; and 2) SingGAN adopts multi-band discriminators and introduces frequency-domain loss and sub-band feature matching loss to supervise high-frequency reconstruction. To our knowledge, SingGAN is the first vocoder designed towards high-fidelity multi-speaker singing voice synthesis. Experimental results show that SingGAN synthesizes singing voices with much higher quality (0.41 MOS gains) over the previous method. Further experiments show that combined with FastSpeech 2 as an acoustic model, SingGAN achieves high robustness in the singing voice synthesis pipeline and also performs well in speech synthesis. Audio samples are available at https://SingGAN.github.io/.

Introduction

Singing voice synthesis (SVS) has been widely employed in various applications with human-computer interaction, such as virtual anchor, artificial intelligence singer and short video score (Blauuw and Bonada 2020; Ren et al. 2020b; Liu et al. 2021; Huang et al. 2021). As shown in Figure 1, similar with text-to-speech (TTS) systems (Ren et al. 2019; Shen et al. 2018; Zhang et al. 2021; Kim et al. 2020; Popov et al. 2021), SVS systems generally adopt an acoustic model to convert the musical score into acoustic features, and a vocoder to generate an audio waveform from acoustic features. However, it is more difficult in SVS to synthesize high-quality waveform than in TTS, because singing voices differ from speech in the following aspects: 1) singing voice has a longer continuous pronunciation and varies in expression and style. 2) Singing voice usually has a higher sampling rate, which results in a wider spectrogram band in the frequency domain and more high-frequency parts. Furthermore, the high sampling rate makes models computationally demanding with increasingly longer durations.

Most previous works focus on optimizing the acoustic model, but usually use speech vocoders for SVS (Gu et al. 2020; Chen et al. 2020). Some speech vocoders have been widely applied to SVS, such as WaveRNN in ByteSing (Gu et al. 2020) and Parallel WaveGAN in HiFiSinger (Chen et al. 2020). However, as an important component in SVS, the vocoder directly impacts the upper bound of generated audio quality. Directly applying those speech vocoders to SVS can result in following issues: 1) speech vocoders have difficulty in modeling long continuous pronunciation of singing voices, and therefore glitches in the spectrogram may emerge; 2) speech vocoders usually lack capacity of high-frequency reconstruction, which leads to noises or unnatural sounds in the high-frequency band. There is still a gap in sample quality between the speech vocoders and singing voice vocoders.

To address the challenges above and synthesize high-quality singing voice, we propose SingGAN, a novel GAN-based vocoder for high-fidelity singing voice synthesis. Specifically, 1) to reduce the glitches long continuous pronunciation of singing voices, SingGAN introduces a source-excitation generator, which is fed with F0 (fundamental frequency) driven source-excitation and can generate stable waveforms with stable and natural pronunciation. 2) To speed up waveform modeling, we further adopt the adaptive feature learning (AFL) filter on the basis of Wavenet. 3) To supervise high-frequency reconstruction in singing voice...
generation, SingGAN introduces Multi-Band Discriminators to distinguish the reconstruction effect of different frequency bands. 4) To achieve faster convergence and better optimization, we further propose the frequency-domain loss, combining sub-band feature matching loss and adversarial loss.

Experimental results show that SingGAN achieves the best performance among existing neural vocoders in terms of 4.1 (MOS) and 0.402 (FDSD). SingGAN, requiring only 1.59M parameters, can achieve a real-time rate of 41X on a single NVIDIA 2080Ti and generate high-fidelity singing voices. In addition, combined with FastSpeech 2 as an acoustic model, SingGAN shows strong robustness in singing voice synthesis. Our further experiments on speech prove that SingGAN also performs well in speech synthesis, which shows the potential of our method to be widely applied to text-to-speech and singing voice synthesis.

Related Works

In this section, we provide the background of singing voice synthesis and briefly review several variations of speech vocoders.

Singing Voice Synthesis  A typical SVS system consists of an acoustic model to convert musical scores into acoustic features and a vocoder to generate an audio waveform from acoustic features. For instance, HiFiSinger (Chen et al. 2020) consists of an acoustic model based on FastSpeech 2 and a vocoder based on Parallel WaveGAN. ML-GAN in HiFiSinger helps supervise waveform reconstruction and achieves good results in single speaker singing voice synthesis, but its F0 embedding reduces model generalizability in multi-speaker singing data. ByteSing (Gu et al. 2020) is a Chinese SVS system based on duration allocated Tacotron-like acoustic model and WaveRNN vocoder. The authors report that those systems can generate natural singing voices. Choi at all (Choi et al. 2020) build a Korean singing voice synthesis system using an auto-regressive algorithm that generates spectrogram with the boundary equilibrium GAN objective.

As the papers say, these previous SVS systems could generate natural singing voices. However, since vocoders in such SVS systems are not designed for singing voice synthesis, vocoder problems can still be observed.

Vocoder  Current vocoders can be categorized into several distinct families: 1) autoregressive-based models: WaveRNN (Kalchbrenner et al. 2018), Paul, Pantazis, and Stylianou (2020) generates waveforms using recurrent neural network layers, which is adopted as vocoder to synthesize waveform in traditional SVS system. However, inference with WaveRNN is inherently slow and hissing noises exit in the synthetic singing voice. 2) Source-excitation-based models: neural source-filter (NSF) (Wang, Takaki, and Yamagishi 2019) uses source excitation combined with a non-autoregressive neural filter to synthesize high-quality speech, while NSF not use the GAN for training, resulting in metallic noise in the generated speech. 3) GAN-based models: Parallel WaveGAN (Yamamoto, Song, and Kim 2020) uses a generator similar to WaveNet in structure and introduces multi-resolution STFT loss and adversarial loss, bringing improvement to waveform reconstruction. MelGAN (Kumar et al. 2019) applies multiple techniques to produce waveforms, but the output quality of MelGAN has room for improvement. A series of improved models (e.g., VocGAN (Yang et al. 2020b), MelGAN (Mustafa, Pia, and Fuchs 2020), Universal MelGAN (Jang, Lim, and Yoon 2020)) have been proposed to further improve voice quality. However, although GAN-based models can synthesize high-quality speech waveform, they have a limited ability in singing voice synthesis due to the above-mentioned differences between speech and singing voice.

Different from the vocoders described above, SingGAN introduces source-exitation and AFL filters for the non-autoregressive generator to better model the characteristics of singing voices. It uses multi-band discriminators as discriminator, and combines frequency-domain loss, sub-band feature matching loss and adversarial loss to conduct adversarial training.

Analysis on Previous Vocoders

Considering the challenges in high-fidelity singing voice synthesis (e.g., long continuous pronunciation, high sampling rate and strong expressiveness), we first conduct experiments and analyze how previous vocoders designed for text-to-speech perform on this task. In this section, we describe two main problems in detail: glitches in the spectrogram and high-frequency reconstruction.

Glitches in the Spectrogram

In contrast to speech, singing voice has a longer continuous pronunciation with a characteristic of smooth harmonics in low frequency. Most existing vocoders take mel-spectrogram as input, but the waveform segments sampled on each mel-spectrogram frame are not continuous. To achieve continuity in waveform segments, a neural network with a large receptive field is required to learn this continuity. As shown in Figure 2, due to the limitations of receptive fields in non-autoregressive vocoders, such as FB-MelGAN and Parallel WaveGAN, glitches in spectrograms easily appear when the singing voice duration exceeds the size of the receptive field. Furthermore, we find that glitches can be
reduced if we increase the receptive field by enlarging the kernel size or adding more dilated convolution layers, but this problem can still be observed if the pronunciation of a phoneme is too long.

**High-Frequency Reconstruction**

![Figure 3: High-frequency reconstruction](image)

Singing voice recordings usually have a high sampling rate (24K or 48K), as well as high pitched sounds and a large frequency range. There exists a short-term fluctuation in high-frequency sounds, which is difficult for vocoders to model in the time domain. Taking waveforms generated by WaveRNN in Figure 3(a) as an example, losing attention on short-term fluctuation usually results in limited capacity of high-frequency reconstruction and hissing noises. Although expanding the receptive field helps the model to learn continuity in waveforms, vocoders like MelGAN easily lose attention on high-frequency reconstruction, leading to noises in the high-frequency band.

Two possible measures have been taken to better reconstruct high-frequency sounds in waveforms. First, vocoders, such as VocGAN and Style MelGAN, take mel auxiliary features as joint input in each upsampling layer. Second, preserving the learned features in each layer could also be taken into consideration, which has been implemented in Parallel WaveGAN and HiFi-GAN.

However, these methods could result in metallic noises, as shown in Figure 3(b). In other words, when generating singing voices using previous vocoders, harmonics can be over-reconstructed in the high-frequency band.

**SingGAN**

Based on the problem analysis, we propose SingGAN and introduce several methods to tackle the challenges in singing voice synthesis. As shown in Figure 4, the generator in SingGAN is driven by F0 as well as log mel-spectrograms. The adversarial loss is computed by an ensemble of five discriminators, including four discriminators operating after a differentiable Pseudo Quadrature Mirror Filter bank (PQMF) and one for full-band discrimination. The discriminators are trained to distinguish ground truth samples from synthesized samples from the generator. The generator is trained to generate samples that maximize the score of the discriminator.

For GAN training, here we denote \( x_k = (x, \text{PQMF}(x)) \), \( y_k = (G(s, f_0), \text{PQMF}(G(s, f_0))) \). For generative adversarial networks, the adversarial losses are defined as:

\[
L_{adv}(D_k; G) = E_{x,s,f_0} \left[ (1 - D_k(x_k))^2 + (D_k(y_k))^2 \right],
\]

\[
L_{adv}(G; D_k) = E_{s,f_0} \left[ \frac{1}{K} \sum_{k=1}^{K} (1 - D_k(y_k))^2 \right],
\]

where \( x \) denotes the ground truth singing voice; \( s \) and \( f_0 \) denote the mel-spectrogram and fundamental frequency of the ground truth waveforms respectively as input condition; \( D_k \) represents k-th discriminators and we set \( K = 5 \) in our paper because there are five discriminators in SingGAN, including one for full-band discrimination and four for multi-band discrimination.

Differences to previous GAN-based vocoders (e.g., Parallel WaveGAN) are: 1) We introduce source excitation and adaptive feature learning (AFL) filter for the generator; 2) We apply multi-band modeling to the discriminator; and 3) We use frequency-domain loss and sub-band feature matching losses in addition to adversarial loss and SFTF loss as training losses. We describe them in detail in this section.

**Generator**

**Source excitation**

To prevent glitches in the spectrogram, we introduce source excitation in the generator. There exist two kinds of F0 driven source excitation structure, which have been applied in the Neural Source Filter (NSF) family of vocoders [Wang, Takaki, and Yamagishi 2019, Wang and Yamagishi 2019]. However, source-excitation combined with simplified WaveNet blocks generates less natural outputs. We combine source-filter with the generative adversarial network here to ensure the stability of singing voice synthesis.

Figure 4(a) depicts the model architecture with source excitation. Assuming that we input \( \tilde{f}_{1:t} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_t\} \) as F0, where \( t \) denotes the total number of frames, we get \( f_{1:T} = \{f_1, f_2, \ldots, f_T\} \) by linear interpolation of \( \tilde{f}_{1:t} \) to the same length as waveform. We get source excitation of \( W \) harmonics by calculating:

\[
F_{ij} = \begin{cases} 
\sin(2\pi j \frac{f_j}{S_T} + \phi_i) & \text{if } f_j > 0, \\
0 & \text{if } f_j = 0 
\end{cases},
\]

where \( i \in \{1, 2, \ldots, W\}, j \in \{1, 2, \ldots, T\}, S_T \) is the sampling rate of audio, and \( F_{ij} \) denotes the excitation value of the \( i \)-th harmonic in \( j \)-th time. \( \phi_i \) is a phase random number.
Figure 4: Architecture of SingGAN: there are three AFL blocks in the generator, each block contains 10 layers of AFL including dilated convolutions for dilation $D_r = \{1, 2, 4, \ldots, 256\}$. The channel size, kernel size, and stride for each dilated convolutional layer, which outputs info-vector as well as control-vector, are set to 64, 5, and 1, respectively. There exist one full-band discriminator and four multi-band discriminators in SingGAN. The number of convolutional layers, kernel size and dilation $D_s$ for high, middle and low frequency bands are set to 6, 7, $\{1, 1,1 \ldots 1\}$; 8, 5, $\{1,2,3 \ldots 8\}$ and 10, 5, $\{1,2,3 \ldots 10\}$, respectively.

in the range of $[-\pi, \pi]$. Because for $f_j = 0$ no frequency period exists, excitation could be replaced by Gaussian noise $Z_{\text{noise}} \sim N(0, \sigma^2)$.

Source excitation transforms the frequency feature in $F_0$ into the initial waveform matrix with low-frequency information through the sine function and oscillator. It ensures that the singing voice is stable and continuous in low frequencies, so as to prevent glitches in the spectrogram.

Adaptive feature learning (AFL) filter — Source excitation contains limited frequency information of singing voice, therefore generator must learn auxiliary features such as timbre and high-frequency information in the mel-spectrogram. Dilated convolutions in WaveNet (Oord et al. 2016) slow down waveform generation, whose details are introduced in Appendix A in the supplementary materials. Here, we design an AFL filter block for the generator on the basis of WaveNet, whose structure is shown in Figure 4(a)(2).

Each AFL block takes both source excitation and the upsampled mel features as input. It outputs the hidden state $h_r$ in multiple receptive fields, passing it to the next AFL block after fusion. To be more specific, as shown in Figure 4(a)(3), each AFL block contains several AFL filter layers with different dilations $D_r$ of dilated convolutions to learn the multi-band feature $\mu$. After upsampled mel features passed through the convolutional layers, we obtain an info-vector $\alpha$ and a control-vector $\beta$. Ultimately, we use sigmoid gated tanh activation functions to calculate the hidden state $h_r$, which is fed into the next AFL filter layer. To conclude, we control the number of AFL filter layers and dilations $D_r$ to adjust the singing voice synthesis in both speed and quality.

Discriminator

In order to better supervise the high-frequency reconstruction in singing voice generation, we adopt multi-band discriminators. Style-MelGAN uses discriminator groups based on multi-frequency random windows. On the contrary, our experiments show that these discriminators increase training time, therefore we do not operate on a random window slice when discriminating multi-frequency bands. Besides, since different frequencies mean different waveform periods, we design discriminators with multiple receptive field scales. Discriminators with a large receptive field evaluate low-frequency waveforms, while those with a small receptive field focus on the short term of high-frequency sounds.

Singing voice segmentations are operated after PQMF, which is free from phase distortion in audio analysis and synthesis. Waveforms segmentation has been used for generator acceleration in Multi-band WaveRNN (Yu et al. 2019) and Multi-band MelGAN, and here we employ it for multi-band discrimination.

Five discriminators have been implemented for waveform judgment, including a full-band discriminator for realness of waveforms and four sub-band discriminators for the accuracy of periodicity reconstruction in the related frequency band. We use dilated convolutions with multiple different dilations $D_s$ to discriminate different periodicity in sub-band and obtain probability $p$ after fusion.

Training Loss

For training loss in SingGAN, we neither apply multi-resolution STFT loss alone to auxiliary loss like in FB-MelGAN (Yang et al. 2020a) and Parallel WaveGAN (Yamamoto, Song, and Kim 2020), nor multi-resolution mel-spectrogram loss alone like in HiFi-GAN (Kong, Kim, and Bae 2020). For one thing, multi-resolution STFT loss pays much attention to the periodicity in the spectrogram. While unvoiced waveforms are almost nonperiodic, it is easy to generate high-frequency pseudo harmonics and metallic noises if we only use multi-resolution STFT loss for training. For another, if only take mel-spectrogram loss to be auxiliary loss, the generator lacks attention on high-frequency information in waveforms, and consequently synthesizes
poor singing voices. Hence, we train SingGAN jointly with multi-resolution STFT loss and mel-spectrogram loss as name frequency-domain loss, so that SingGAN generates singing voice more naturally. The mel-spectrograms of singing voice generated by different training loss have been introduced in figure [5].

Since full-band adversarial loss lacks concentration on details in waveform modeling, we introduce the sub-band feature matching objective based on multi-band discriminators to improve the naturalness of singing voice and help SingGAN learn aperiodic information. Both frequency-domain loss function and sub-band feature matching loss function are different from those in previous vocoders. The specific definitions are as follows:

**Frequency-domain loss** To avoid the drawbacks of multi-resolution STFT loss and mel-spectrogram loss, we add frequency-domain loss to improve the training efficiency of the generator and the fidelity of the synthesized singing voice, which is defined as

\[
L_{m_{-}\text{sc}}(x, y) = \frac{\| MEL(x) - MEL(y) \|_F}{\| MEL(x) \|_F},
\]

\[
L_{m_{-}\text{mag}}(x, y) = \frac{1}{N} \| \log(MEL(x)) - \log(MEL(y)) \|_1,
\]

where \(x\) denotes the ground truth singing voice and \(y\) denotes the synthesis singing voice. \(\| \cdot \|_F\) and \(\| \cdot \|_1\) denote the Frobenius and L1 norms; \(MEL(\cdot)\) and \(N\) denote mel extraction operation and the number of elements in the magnitude, respectively.

The final multi-resolution mel-spectrogram loss is the sum of losses with different analysis parameters (i.e., FFT size, window size, and hop size). The multi-resolution mel-spectrogram loss is represented as follows:

\[
L_{mel}(x, y) = \frac{1}{M} \sum_{m=1}^{M} \left( L_{m_{-}\text{sc}}^{(m)}(x, y) + L_{m_{-}\text{mag}}^{(m)}(x, y) \right).
\]

Our final auxiliary frequency-domain loss is defined as a linear combination of the multi-resolution STFT loss and the multi-resolution mel-spectrogram loss as follows:

\[
L_{aux}(G) = E_{(x,s,f_0)} \left[ \frac{1}{2} (L_{mel}(x, y) + L_{stft}(x, y)) \right],
\]

where \(L_{stft}(x, y)\) denotes the multi-resolution STFT loss, and \(L_{aux}(G)\) denotes the auxiliary frequency-domain loss of the generator.

**Sub-band feature matching loss** We introduce the sub-band feature matching loss to punish the loss of hidden features in each frequency domain, such as timbre and style in the singing voice. In addition to the training objectives above, we use a sub-band feature matching objective to train the generator, which has also been used in MelGAN and HiFi-GAN. It is worth noting that our feature matching loss is different from that in MelGAN with a different discriminator structure. Our proposed sub-band feature matching loss minimizes the L1 distance between the discriminators’ ultimate output map of real and synthetic audio in multi-band rather than full-band.

\[
L_{fm}(G; D_k) = E_{(x,s,f_0)} \left[ \sum_{k=2}^{K} \| D_k(x_k) - D_k(y_k) \|_1 \right].
\]

Our final loss function for the generator is defined as a linear combination of the auxiliary frequency-domain loss, adversarial loss, and sub-band feature matching loss as follows:

\[
L_G = L_{aux}(G) + \lambda_1 L_{adv}(G; D_k) + \lambda_2 L_{fm}(G; D_k),
\]

where \(\lambda_1\) and \(\lambda_2\) denotes the hyperparameters balancing loss terms, and we set \(\lambda_1 = 4\) and \(\lambda_2 = 10\), respectively. By jointly optimizing the waveform-domain adversarial loss, frequency-domain loss, and sub-band feature matching loss, the generator can learn the distribution of the realistic speech waveform effectively.

**Experiments**

In this section, we first describe the experimental setup including the dataset and model configurations. Then, we report our experimental results and conduct some analyses.

**Experimental Setup**

**Dataset** Our singing dataset contains Mandarin pop songs collected from 93 singers in a professional recording studio. The dataset consists of 80 hours of singing voice, including 50 hours from 56 females and 30 hours from 37 males. All songs are saved in wav format, sampled at 24 kHz, and quantized by 16 bits. We randomly choose 340 utterances for validation and 60 utterances from 6 speakers as seen test set. To evaluate the generalization and robustness of SingGAN, we prepare utterances from unseen speakers for the additional test set. The unseen test set contains 5 utterances from each speaker, including five males and five females.

To aid singing voice research in the community, we will release our singing voice dataset after publication.

**Evaluation Metric** To evaluate the audio quality, we crowd-sourced 5-scale MOS tests. 10 native Chinese speakers evaluate the singing voice quality. In the MOS tests, after listening to each stimulus, the subjects are asked to rate the naturalness of the stimulus on a five-point Likert scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Besides, We adopt Fréchet Deep Speech Distances (FDSD) (Gritsenko et al. 2020) to judge the quality of synthetic audio samples based on their distance to a reference set. These distances are conceptually similar to the FID (Fréchet Inception Distance). The energy distance can be combined with GAN-based learning, further improving on either individual technique.

**Implementation details** The generator in SingGAN takes both log mel-spectrogram and F0 as input. 80-band log-mel-spectrograms are extracted in params (FFT:512, hop size:128, window size:512). We extract F0 from the singing audio using the public tool Parselmouth [6] To match hop size,
the upsampling rate of the mel-spectrogram is set to 8, 4, and the upsampling rate of F0 is set to 128. In order to smooth the low-frequency harmonics, we use 8 harmonics in source-excitation. At the training stage, frequency-domain loss are computed by the sum of three different STFT losses and three different mel-spectrogram losses as described in Appendix B.

The adversarial loss is computed by the average of per-time step scalar predictions with Multi-band discriminators. Models are trained for 400K steps with an RAdam optimizer to stabilize training. We apply pre-training on the generator in the first 50K steps, and two models are jointly trained afterwards.

When training the baseline systems such as WaveRNN, FB-MelGAN Parallel WaveGAN and NSF the implementations from the respective GitHub repositories are used for reproducibility and our configurations follow their original papers.

**Singing Voice**

To implement MOS assessments, we prepare a test set from our singing datasets, including 40 utterances in total for four seen and four unseen speakers. Table 1 shows that SingGAN scores higher than most models on the test set.

These experimental results show the robustness of SingGAN and its better generalization capabilities. All SingGAN models achieve higher MOS than the baseline models. SingGAN has 1.59M parameters, which is comparable to the number of parameters of Parallel WaveGAN (1.28M), but SingGAN is 53 times faster than Parallel WaveGAN and WaveRNN, respectively.

In the MOS evaluation, we conclude that 1) for singing voice generation, noise exists in high-frequency waveforms generated by WaveRNN due to its poor high-frequency reconstruction; 2) Parallel WaveGAN with limited receptive field fails to smooth the continuous pronunciation in singing voice, so glitches in the spectrogram of synthesized voices appear; 3) the glitches in the spectrogram and high-frequency noise even more obvious in FB-MelGAN, possibly because it stacks fewer residual layers; and 4) SingGAN generates noise-suppressed audio both in the low and high-frequency band, where source-excitation makes great contributions.

**Ablation Study**

We conduct an ablation study to verify the effectiveness of several components in SingGAN, including: 1) source-excitation; 2) adaptive feature learning (AFL) filter; 3) multi-band discriminators (MBD); and 4) frequency-domain loss and sub-band feature matching loss (FDL and SFML). We mainly conduct MOS evaluations to compare two different settings, where each of the randomly chosen 20 evaluation pieces in the test set are evaluated by 10 judges.

![Figure 5: Spectrograms of singing voice generated by different training loss. (a) Ground-truth. (b) Multi-resolution STFT loss only. A lot brighter spectrogram in the high-frequency band than usual. (c) mel-spectrogram loss only. High-frequency reconstruction failure. (d) Our joint method of frequency-domain loss.](image-url)
Model Parameters Inference speed (↑) MOS (↑) FDSD (↓)
---
WaveRNN 4.48M 0.77 3.88±0.12 0.385
FB-MelGAN 5.31M 125 3.24±0.10 0.864
Parallel WaveGAN 1.44M 32 3.58±0.11 0.484
NSF 1.2M 29 3.12±0.08 –
SingGAN 1.59M 41 4.05±0.09 0.402

Table 1: The inference speed and the MOS results with 95% confidence intervals on singing voice. The evaluation is conducted on a server with a single NVIDIA 2080Ti. Note that the inference speed k means that the system was able to generate waveforms k times faster than real-time.

---

Model Inference Speed CMOS
---
SingGAN 41 0.00
- Source 45 -0.28
- AFL 24 -0.10
- MBD 41 -0.22
- FDL 41 -0.16
- SFML 41 -0.12

Table 2: Ablation study results. Comparison of the effect of each component on the synthesis quality.

a better singing voice quality. Removing the sub-band feature matching results in lower expressiveness of the singing voice as well as a slower convergence during training. This demonstrates that the frequency-domain loss and the sub-band feature matching loss enable the training process to be more effective and that audio quality is improved.

**Singing Voice Synthesis System**

In order to verify the effectiveness of SingGAN in singing voice synthesis, we build a SVS system and adopt FastSpeech 2 as an acoustic model to convert the words of songs into acoustic features. During training, we adopt the configuration from (Ren et al. 2020) and make some modification as follows: 1) we remove pitch and duration prediction, taking the real F0 and phoneme duration as input; and 2) we add speaker embedding in FastSpeech 2 as identification for multi-speaker singing voice generation.

During inference, FastSpeech 2 converts phoneme, F0, and duration into acoustic features, while SingGAN generates audio waveform from mel-spectrogram as well as F0. The MOS evaluation for synthetic singing voice is presented in Table 3. Experimental results show that the audio generated by FastSpeech 2 and SingGAN has a higher quality than the one generated by FastSpeech 2 and Parallel WaveGAN, while previous discussed problems of Parallel WaveGAN still exist in this SVS system.

**Speech**

To further verify the generality of SingGAN, we conduct experiments on the multi-speaker English dataset VCTK and the single-speaker Chinese dataset CSMSC. Our configuration follows prior singing voice synthesis systems. We perform MOS tests and present the results in Table 4. The results show that SingGAN achieves good results in speech synthesis as well. Although MOS differences in speech are not so much as that in the singing voice, it is obvious that SingGAN outperforms prior neural vocoders in audio synthesis. Consequently, our proposed SingGAN has the potential to be widely applied in TTS and SVS.

**Conclusion**

We proposed SingGAN, a singing voice vocoder with a generative adversarial network. To our knowledge, SingGAN is the first vocoder designed towards high-fidelity multi-speaker singing voice synthesis, which exceeds the limitation of speech vocoders applied in SVS. To reduce glitches in spectrograms, we introduced the source excitation and AFL filter in the generator, which efficiently alleviates glitches in singing voice synthesis. To settle the problem of poor high-frequency reconstruction, we introduced multi-band discriminators and attached frequency-domain loss and sub-band feature matching loss for stable training and high-frequency reconstruction. Experimental results show that SingGAN synthesized singing voices with much higher quality than previous models. In future work, we will continue to reduce the number of model parameters and speed up high-fidelity singing voice generation.
References

Blaauw, M.; and Bonada, J. 2020. Sequence-to-sequence singing synthesis using the feed-forward transformer. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7229–7233. IEEE.

Chen, J.; Tan, X.; Luan, J.; Qin, T.; and Liu, T.-Y. 2020. Hi-FiSinger: Towards High-Fidelity Neural Singing Voice Synthesis. arXiv preprint arXiv:2009.01776.

Choi, S.; Kim, W.; Park, S.; Yong, S.; and Nam, J. 2020. Korean singing voice synthesis based on auto-regressive boundary equilibrium gan. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7234–7238. IEEE.

Gritsenko, A. A.; Salimans, T.; van den Berg, R.; Snoek, J.; and Kalchbrenner, N. 2020. A Spectral Energy Distance for Parallel Speech Synthesis. arXiv:2008.01160.

Gu, Y.; Yin, X.; Rao, Y.; Wan, Y.; Tang, B.; Zhang, Y.; Chen, J.; Wang, Y.; and Ma, Z. 2020. ByteSinging: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoder. arXiv preprint arXiv:2004.11012.

Huang, R.; Chen, F.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2021. Multi-Singer: Fast Multi-Singer Singing Voice Vocoder With A Large Scale Corpus. In Proceedings of the 29th ACM International Conference on Multimedia.

Jang, W.; Lim, D.; and Yoon, J. 2020. Universal MelGAN: A Robust Neural Vocoder for High-Fidelity Waveform Generation in Multiple Domains. arXiv preprint arXiv:2011.09631.

Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A. v. d.; Dieleman, S.; and Kavukcuoglu, K. 2018. Efficient neural audio synthesis. arXiv preprint arXiv:1802.08435.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. arXiv:2005.11129.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv preprint arXiv:2010.05646.

Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; and Courville, A. C. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. In Advances in Neural Information Processing Systems, 14910–14921.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; Liu, P.; and Zhao, Z. 2021. Diffisinger: Diffusion acoustic model for singing voice synthesis. arXiv preprint arXiv:2105.02446.

Mustafa, A.; Pia, N.; and Fuchs, G. 2020. StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization. arXiv preprint arXiv:2011.01557.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
Appendix A: Dilated Convolutions in WaveNet

Figure 6: X represents random noise as initial waveforms in usual, and S denotes upsampled mel-spectrogram feature (i.e. auxiliary feature). Firstly, X passes through the dilated convolution layer, which is an key component to enhance sense of hearing. Next, X and S are divided into xa, xb and sa, sb, respectively. After sigmoid-tanh calculation, the processing feature passes through two fully-connected networks and output H as well as X, which will be fed into the next dilated convolution layer.

Appendix B: Loss Details

| Loss type | FFT size | Frame shift | Window size |
|-----------|----------|-------------|-------------|
| STFT      | 1024     | 600         | 120         |
|           | 2048     | 120         | 250         |
|           | 512      | 240         | 50          |
| Mel       | 4096     | 540         | 2160        |
|           | 2048     | 270         | 1080        |

Table 5: The details of the multi-resolution STFT loss and multi-resolution mel-spectrogram loss. A Hanning window was applied before the FFT process.

Appendix C: Singing Voice Dataset

We check for audio quality and conduct statistical evaluation, including sentence-level and phoneme-level duration distribution, as well as pitch distribution. Evaluation results have been presented below:

To aid singing voice research in the community, we will release our singing voice dataset after publication.