IMPROVING ADVERSARIAL WAVEFORM GENERATION BASED SINGING VOICE CONVERSION WITH HARMONIC SIGNALS

Haohan Guo[*], *Zhiping Zhou*[†‡], *Fanbo Meng*[†‡], *Kai Liu*[†‡]

[*]The Chinese University of Hong Kong, Hong Kong SAR, China
[†]Sogou Inc., Beijing, China, [‡]Tencent PCG AIBU

hguo@se.cuhk.edu.hk, {merryzhou, fanbomeng, kevinllliu}@tencent.com

ABSTRACT

Adversarial waveform generation has been a popular approach as the backend of singing voice conversion (SVC) to generate high-quality singing audio. However, the instability of GAN also leads to other problems, such as pitch jitters and U/V errors. It affects the smoothness and continuity of harmonics, hence degrades the conversion quality seriously. This paper proposes to feed harmonic signals to the SVC model in advance to enhance audio generation. We extract the sine excitation from the pitch, and filter it with a linear time-varying (LTV) filter estimated by a neural network. Both these two harmonic signals are adopted as the inputs to generate the singing waveform. In our experiments, two mainstream models, MelGAN and ParallelWaveGAN, are investigated to validate the effectiveness of the proposed approach. We conduct a MOS test on clean and noisy test sets. The result shows that both signals significantly improve SVC in fidelity and timbre similarity. Besides, the case analysis further validates that this method enhances the smoothness and continuity of harmonics in the generated audio, and the filtered excitation better matches the target audio.

Index Terms— singing voice conversion, harmonic signal, sine excitation, neural vocoder, GAN

1. INTRODUCTION

Singing Voice Conversion (SVC) aims to convert the timbre of a song to the target singer without changing the singing content including the melody and lyrics. With the promotion of various karaoke and social Apps, SVC has been paid more and more attention to provide a better and more diverse entertainment experience. To build a good SVC system, it is expected not only to imitate the target timbre well, but also to keep the original singing content and high audio quality. Many approaches, e.g. adversarial training [1, 2] and Phonetic PosteriorGrams (PPG) [3, 4, 5, 6, 7], have been proposed to extract disentangled features representing content, melody, timbre, respectively, to better convert the timbre. To produce a high-quality singing audio, the NN-based waveform generation module is usually adopted as the backend of SVC model.

Among different frameworks for audio generation, GAN-based approaches [8, 9, 10, 11, 12, 13] have been widely used in VC [14, 15, 16] and SVC [17, 18]. Different from other architectures, such as WaveNet [19] and WaveGlow [20], it can rapidly generate high-quality audio in parallel with a smaller-footprint model, which is helpful for the end-to-end training. However, its instability also affects the smoothness and continuity of the generated singing audio, hence causes some problems, such as pitch jitters and U/V errors in SVC. They are mainly reflected in the harmonic component which is more important in the singing voice, and affect the auditory experience of the whole song seriously. To tackle these issues, in this paper, we purpose to directly feed harmonic signals to the model to help generate the better singing voice with continuous and smooth harmonics.

This work is composed of two aspects, the generation of harmonic signals, and the approach applying them to SVC models. Firstly, we investigate the sine excitation, which can be directly computed from F0. It is widely used in source-filter vocoders, including the neural version [21]. Then, we enhance it by filtering it with a Linear Time-Varying (LTV) filter [13]. Its coefficients are estimated by a neural network according to the input features. This signal better matches the target audio, hence improves the effect of harmonic signals. To validate the effectiveness of these signals in SVC, we conduct experiments based on the end-to-end PPG-based SVC model [17]. And two mainstream architectures, MelGAN [8] and ParallelWaveGAN [11], are both investigated in our work, to show the universal usability of the proposed method to SVC based on adversarial audio generation.

In this paper, we will firstly introduce our approach extracting the harmonic signals in Sec.2, then illustrate the model architecture of the SVC model in Sec.3. Finally, a MOS test is conducted to evaluate the models in audio quality and timbre similarity. The result shows that harmonic signals enhance all SVC models significantly, especially when combining the raw sine excitation and the filtered one. In addition, the analysis of the spectrogram also reveals that the smoothness and continuity of the harmonics are improved obviously by our approach.

2. HARMONIC SIGNALS

To provide valuable help to SVC, we first attempt to find suitable harmonic signals that well-match the target singing audio. Since the sine excitation has shown its importance in signal-based and NN-based vocoders, we mainly investigate it in our work.

2.1. Sine Excitation

Sine excitation is a simple harmonic signal that can be directly converted from the input F0 via additive synthesis. The frame-level pitch sequence $f_0[m]$ is firstly interpolated linearly to the sample-level pitch sequence $f_0[n]$ according to the sample rate f_s. Then sine excitation can be generated using the below function:

$$p[n] = \begin{cases} \sum_{k=1}^{K} \sin(\phi_k[n]) & f_0[n] > 0 \\ 0 & f_0[n] = 0 \end{cases} \quad (1)$$

where K is the number of harmonic and $\phi_k[n]$ is the phase of the k-th harmonic at timestep n, they can be determined as follows:

$$K = \lfloor \frac{f_s}{2f_0[n]} \rfloor \quad (2)$$

$$\phi_k[n] = \phi_k[n-1] + 2\pi k \frac{f_0[n]}{fs} \quad (3)$$

2.2. Filtered Sine Excitation

The sine excitation can describe the spectral shape of the harmonics in the singing voice accurately. However, the energy uniformly distributed on each harmonic cannot match the target audio well. The signal does not include sufficient information about the lyrics, energy, and timbre. For better matching, an LTV filter is constructed to filter the raw sine excitation based on the input features. Its coefficients are time-varied, and are estimated by a neural network according to the frame-level input sequence, which is similar to NHV [13]. This operation dynamically adjusts the amplitudes of different harmonic components, then provides a more suitable harmonic signal for the following waveform generation.

3. MODEL ARCHITECTURE

In the autoencoder-based SVC framework, feature extraction and audio reconstruction are the two main modules. The disentangled features representing content, melody, and timbre need to be extracted for precise manipulation. This work is implemented based on Phonetic PosteriorGrams (PPGs), a robust content representation, to ensure high intelligibility. In addition, for better conversion quality, an end-to-end model based on adversarial waveform generation is constructed to directly map these features to the corresponding audio.

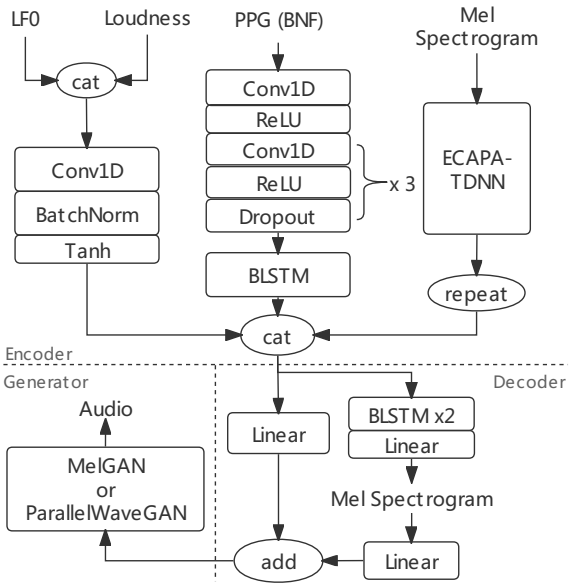


Fig. 1. The architecture of our SVC model

3.1. PPG-based Singing Voice Conversion Model

Fig.1 shows the architecture of the end-to-end SVC model. Compared with the conventional approach, which predicts acoustic features firstly, then generates audio using a vocoder trained independently, it avoids the cascade errors, and exploits richer and useful information from the input, hence generates better results [17]. This model is composed of three modules, an encoder, a decoder, and a waveform generation module (vocoder). They are trained jointly, and do not need pre-training.

The encoder encodes different features representing content, melody, timbre, respectively. The content representation, PPGs, is encoded through a stack of 1-D convolutional layers and a BLSTM layer, which is similar to the text encoder in Tacotron2 [22]. The LF0 and loudness non-linearly transformed are adopted as the melody representation. The ECAPA-TDNN [23] is used as the timbre encoder to extract timbre representation from the Mel-spectrogram of the input audio. Its advanced model structure and attentive statistical pooling can provide a more robust and effective speaker-dependent embedding, which has been validated in speaker verification.

Due to larger parameters and the more complicated model structure, the waveform generation module is usually unstable in end-to-end training. So we add a decoder to predict the Mel-spectrogram before it, then combine its outputs and the encoder outputs using two linear layers and an addition operation. We find that this approach can force the encoder to provide more robust acoustic-related information during the training process. In this way, the degradation of the generalization and stability due to the end-to-end training can be alleviated effectively.

3.2. Adversarial Waveform Generation

As shown in Fig.2, there are two mainstream adversarial frameworks, the MelGAN-derived and the ParallelWaveGAN-derived. MelGAN generates the waveform by upsampling the input features using a upsample network. Instead, ParallelWaveGAN (PWG) processes a Gaussian noise sequence to the target waveform with a WaveNet-like model. Most recently proposed adversarial vocoders are based on these two frameworks. So we apply our approach in both frameworks to validate its effectiveness and universal usability in adversarial waveform generation.

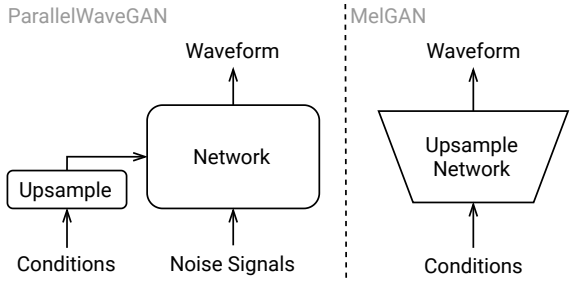


Fig. 2. The architectures of MelGAN and ParallelWaveGAN

The unfiltered and filtered sine excitation are both investigated in our experiments. They are inserted to MelGAN and PWG in the way shown in Fig.3. Before waveform generation, the features are firstly used to estimate the LTV filter coefficients to generate harmonic signals. The F0 extracted from the source audio can be used to calculate the sine excitation. For PWG, the harmonic signals can be directly concatenated with the noise signals together as the inputs. For MelGAN, the conditions are upsampled to the hidden features with different scales by several upsampling blocks. So the harmonic signals are also downsampled to these scales via a neural network, then concatenated with them for the following operation. Due to different model structures, the filtered sine excitation is also different, which are shown in 4.3.

4. EXPERIMENTS

4.1. Experimental Protocol

Our experiments are all implemented on a clean multi-singer singing corpus. It is composed of 200 different Mandarin pop songs (total 10 hours) sung by 10 singers (5 males and 5 females, 20 songs per singer). The F0, loudness, and 80-dim Mel-spectrograms are extracted from the audio with a sample rate of 16kHz. The frameshift of these features is 10ms. PPGs are extracted from a Transformer-CTC ASR model [24, 25] trained with a large-scale Mandarin speech corpus (over 10,000 hours). Its 384-dim encoder output with 40ms frameshift is adopted as the PPG feature.

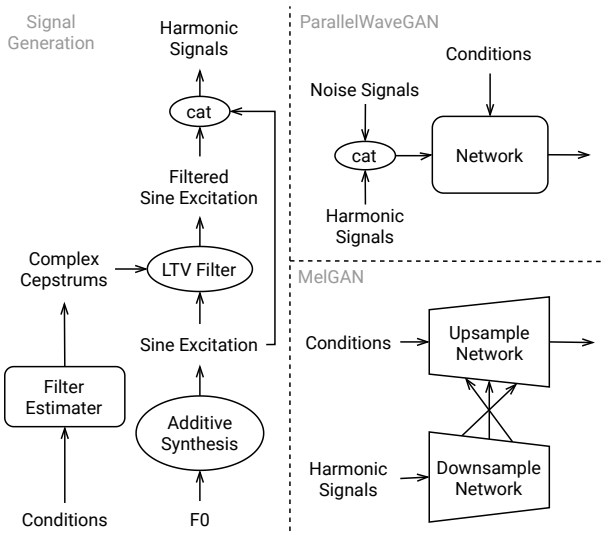


Fig. 3. Applying harmonic signals to MelGAN and PWG

In the content encoder, the first layer is a 1-D transposed convolutional layer to upsample PPGs to align other features. The output dimension of each encoder is set to 256. The structure of MelGAN is the same as [17], with the upsampling factors of 8, 6, 5. The model structure of the waveform discriminator [8] is adopted to downsample harmonic signals to the corresponding framerates. The ParallelWaveGAN is optimized with LVCNet [12] to accelerate the training and inference speed. For the LTV filter, we adopt the model structure proposed in [13] to estimate the FIR filter coefficients to process the sine excitation.

The SVC model and the discriminator are trained adversarially with the same algorithm described in [17]. Besides the adversarial loss function L_{adv} and the MR-STFT Loss L_{stft}, L_{dec} is set for the decoder to calculate MAE loss between the predicted Mel-spectrogram and the target one. The final loss function is:

$$L = \alpha * L_{dec} + \beta * L_{adv} + L_{stft} \tag{4}$$

which α and β are set to 200 and 4 respectively. All models are trained for 400,000 iterations with a batch size of 16x4, i.e. 16 audio segments with the length of 4 seconds.[1] [2]

4.2. Subjective Evaluation

The mean opinion score (MOS) test is conducted to evaluate the proposed method in two aspects: sound quality and singer similarity. Two test sets are used, a clean set with 30 high-quality clean singing segments, and a noisy set with 30 singing segments recorded with worse quality. The timbre in

[1]The details of LVCNet can be found at `https://github.com/zceng/LVCNet`.

[2]The training details can be found at `https://github.com/hhguo/EA-SVC`

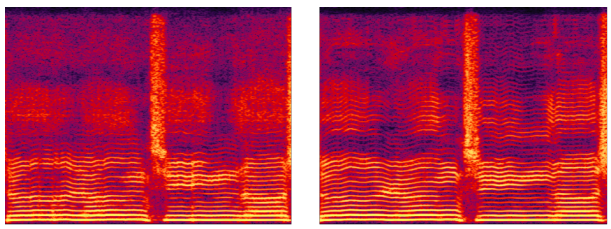
Table 1. The MOS test results ($\pm$ indicates 95% CI, MG: MelGAN, PWG: ParallelWaveGAN)

Model	Harmonic		Clean		Noisy		Overall	
	Raw	Filtered	Quality	Similarity	Quality	Similarity	Quality	Similarity
MG	×	×	3.05 ± 0.14	3.17 ± 0.14	2.85 ± 0.14	3.03 ± 0.16	2.96 ± 0.10	3.10 ± 0.11
MG	✓	✓	3.45 ± 0.14	3.45 ± 0.15	3.17 ± 0.14	3.21 ± 0.16	3.31 ± 0.10	3.33 ± 0.11
PWG	×	×	3.08 ± 0.15	3.22 ± 0.16	2.94 ± 0.13	3.11 ± 0.15	3.01 ± 0.10	3.17 ± 0.11
PWG	✓	×	3.25 ± 0.13	3.38 ± 0.14	3.07 ± 0.14	3.14 ± 0.15	3.16 ± 0.10	3.26 ± 0.10
PWG	✓	✓	3.41 ± 0.14	3.42 ± 0.15	3.15 ± 0.15	3.19 ± 0.15	3.31 ± 0.10	3.31 ± 0.11

each segment is converted to a random singer contained in the training set. Finally, 24 listeners are involved to rate the 60 test cases. Each of them only rates 20 cases to ensure an accurate result [3].

The final results are shown in Table.1. For the same model structure, the models with harmonic signals achieve the higher score on all test sets, especially for the clean set, which shows the best conversion performance. For the noisy set, although the extraction of PPG and F0 are easily contaminated by the noise, it still achieves significant improvement, which validates the effectiveness and robustness of our method. Moreover, compared with the similarity, the improvement is more obvious in sound quality, which gets an increase of more than 0.3 on the whole test set. It strongly verifies that it is important to present good harmonics in SVC.

To investigate the effect of the filtered sine excitation, we also evaluate the PWG with or without it. The MOS result shows that the model with only the raw sine excitation can already outperform the baseline PWG. But the filtered sine excitation can further improve the model significantly, which shows its importance and improvement to SVC.

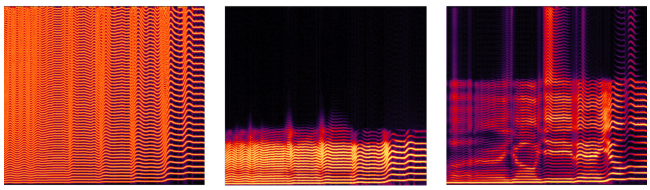


(a) w/o harmonic signals (b) w/ harmonic signals

Fig. 4. The spectrograms generated by the MelGAN models w/ or w/o harmonic signals

4.3. Analysis

Fig.4 shows the spectrograms synthesized by MelGAN with or without harmonic signals. The problems in (a), including pitch jitter and unsmoothed harmonics, are all solved in (b). Besides, different from the muffle spectrogram in the middle frequency in (a), (b) shows more clear harmonics, which makes the converted singing audio more smooth and clean.



(a) Raw (b) Filtered in PWG (c) Filtered in MelGAN

Fig. 5. The spectrograms of different sine excitations

We also compare different harmonic signals in MelGAN and PWG. As shown in Fig.5, (a) is the spectrogram of the raw sine excitation. It only has harmonics with evenly distributed energy. After filtered by the LTV filter in PWG, the high frequency is removed dynamically, whose range is different at each frame. In the reserved part, the energy is distributed differently to each curve, which better matches the target audio. This phenomenon becomes more eminent in MelGAN. In (c), the reserved frequency has a wider range and higher variety. The vowel is synthesized well that can be recognized clearly. Its timbre also becomes more distinguishable to imitate the target one.

5. CONCLUSION

This paper purposes to utilize harmonic signals to improve singing voice conversion based on adversarial waveform generation. We investigate two signals, the raw sine excitation and the sine excitation filtered by an estimated LTV filter, and apply them to two mainstream adversarial waveform generation models, MelGAN and ParallelWaveGAN. The experimental results show our method significantly improves the sound quality and timbre similarity on the clean and noisy test sets. And the filtered sine excitation is also validated as a powerful harmonic signal to SVC. The case analysis shows that the method improves the smoothness, continuity, and clarity of harmonic components. And the filtered excitation better matches the target audio.

[3] Samples are available at `https://hhguo.github.io/DemoHarSVC`

6. REFERENCES

[1] Eliya Nachmani and Lior Wolf, "Unsupervised singing voice conversion," in *INTERSPEECH*, 2019.

[2] C. Deng, C. Yu, H. Lu, C. Weng, and D. Yu, "Pitchnet: Unsupervised singing voice conversion with pitch adversarial network," in *ICASSP*, 2020, pp. 7749–7753.

[3] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*. IEEE, 2018, pp. 5274–5278.

[4] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*. IEEE, 2016, pp. 1–6.

[5] Xiaohai Tian, Eng Siong Chng, and Haizhou Li, "A speaker-dependent wavenet for voice conversion with non-parallel data.," in *INTERSPEECH*, 2019, pp. 201–205.

[6] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu, "Singing voice conversion with non-parallel data," in *MIPR*, 2019, pp. 292–296.

[7] Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Ling Xu, Chen Shen, and Zejun Ma, "Ppg-based singing voice conversion with adversarial representation learning," in *ICASSP*. IEEE, 2021, pp. 7073–7077.

[8] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14910–14921.

[9] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 492–498.

[10] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[11] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[12] Zhen Zeng, Jianzong Wang, Ning Cheng, and Jing Xiao, "Lvcnet: Efficient condition-dependent modeling network for waveform generation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6054–6058.

[13] Zhijun Liu, Kuan Chen, and Kai Yu, "Neural homomorphic vocoder.," in *INTERSPEECH*, 2020, pp. 240–244.

[14] Yaogen Yang, Haozhe Zhang, Xiaoyi Qin, Shanshan Liang, Huahua Cui, Mingyang Xu, and Ming Li, "Building bilingual and code-switched voice conversion with limited training data using embedding consistency loss," *arXiv preprint arXiv:2104.10832*, 2021.

[15] Shijun Wang and Damian Borth, "Noisevc: Towards high quality zero-shot voice conversion," *arXiv preprint arXiv:2104.06074*, 2021.

[16] Kang-wook Kim, Seung-won Park, and Myun-chul Joe, "Assem-vc: Realistic voice conversion by assembling modern speech synthesis techniques," *arXiv preprint arXiv:2104.00931*, 2021.

[17] Haohan Guo, Heng Lu, Na Hu, Chunlei Zhang, Shan Yang, Lei Xie, Dan Su, and Dong Yu, "Phonetic posteriorgrams based many-to-many singing voice conversion via adversarial training," *arXiv preprint arXiv:2012.01837*, 2020.

[18] Songxiang Liu, Yuewen Cao, Na Hu, Dan Su, and Helen Meng, "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[19] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio.," in *Speech Synthesis Workshop (SSW)*, 2016, p. 125.

[20] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. IEEE, 2019, pp. 3617–3621.

[21] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP*. IEEE, 2019, pp. 5916–5920.

[22] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *INTERSPEECH*, 2020, pp. 3830–3834.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeuralPS*, 2017, pp. 5998–6008.

[25] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*. IEEE, 2017, pp. 4835–4839.