Comparative Evaluation of Neural Vocoders for Speech Synthesis of Operatic Singing

Shimizu, Sota ; Matsubara, Keisuke ; Adachi, Yuji ; Tai, Kiyoto ; Takashima, Ryouichi ; Takiguchi, Tetsuya

(Citation)
2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech):28-29

(Issue Date)
2022-04-14

(Resource Type)
conference paper

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/0100481910

Comparative Evaluation of Neural Vocoders for Speech Synthesis of Operatic Singing

Sota Shimizu
Faculty of Engineering
Kobe University
Japan
shimizu_s@stu.kobe-u.ac.jp

Keisuke Matsubara
Graduate School of System Informatics
Kobe University
Japan

Yuji Adachi
MEC Company Ltd.
Japan

Kiyoto Tai
MEC Company Ltd.
Japan

Ryoichi Takashima
Graduate School of System Informatics
Kobe University
Japan

Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Japan

***Abstract*—The voice of someone who is singing opera has different characteristics compared with someone singing nursery rhymes or pop songs, which are treated by conventional singing voice synthesis. In this study, we compared and evaluated the speech synthesis performance of recently proposed neural vocoders for an operatic singing voice. From the results of our objective evaluation experiments, it was found that the mel-cepstrum estimation error for Parallel WaveGAN and PeriodNet is small. Furthermore, PeriodNet has a smaller estimation error of fundamental frequency.**

***Index Terms*—speech synthesis, operatic singing voice, vocoder, neural network**

I. Introduction

Speech synthesis for singing voices is finding wide use and attracting attention in the entertainment field. With the development of speech synthesis using neural networks, it has become possible to synthesize high-quality speech in the field of singing voice synthesis. In recent years, various models of neural vocoders have been proposed, and neural vocoders, such as WaveNet [1], have greatly surpassed the quality of the conventional source-filter vocoder [2], contributing greatly to the development of singing voice synthesis technology.

Although WaveNet provides high-quality speech synthesis, the synthesis speed is slow. To solve this problem, neural vocoders, such as Parallel WaveGAN [3] which can synthesize in real time, have been studied in recent years. PeriodNet [4] has also been proposed as a neural vocoder focusing on singing voices.

Operatic singing voices, the subject of this study, have different vibrato strength, pitch, vowel characteristics, etc., compared to nursery rhyme singing voices or pop song singing voices, which are handled using conventional singing voice synthesis. There are few studies on neural vocoders for singing voices that are different from normal singing voices. Therefore, in this study, we compare and evaluate the performance of recent neural vocoders for operatic singing voices. We conduct experiments of analysis synthesis task using a cappella opera songs sung by a singer, and we compare the performances of the neural vocoders described above using objective metrics.

II. Neural vocoders

A vocoder is a module that generates a speech waveform using acoustic features generated by speech synthesis as the input. In this study, three neural vocoders (WaveNet, Parallel WaveGAN, and PeriodNet) were evaluated.

A. WaveNet

WaveNet is a convolutional neural network with an auto-regressive structure. WaveNet directly estimates the speech waveform as a classification problem by predicting the next sample conditioned on past speech samples. In WaveNet, the speech signal is quantized from 16-bit to 8-bit by using μ-law transformation to come up with a simple prediction.

WaveNet is capable of producing high-quality speech, but its autoregressive structure and large model structure result in its synthesizing speed being slow, and the noise component caused by overtraining and prediction errors is also a problem. In this paper, we apply the time-invariant noise shaping method [5] to suppress the noise component.

B. Parallel WaveGAN

Parallel WaveGAN is a Generative Adversarial Network (GAN) based on Parallel WaveNet [6]. The input is white noise and acoustic features, and the generator based on a WaveNet generates all samples simultaneously. In addition to the adversarial loss used in conventional GANs, Short-Term Fourier transform (STFT)-loss is introduced to assist the training of the generator.

Since Parallel WaveGAN does not have an auto-regressive structure, it can be synthesized quickly by generating multiple samples at the same time, and it can be generated in real time.

TABLE I
MEL-CEPSTRAL DISTORTION FOR EACH METHOD. A, B, C, D, AND E MEAN THE EVALUATED FIVE SONGS.

Model	Evaluation data					Average
	A	B	C	D	E	
WORLD	3.656	3.246	3.378	3.211	3.974	3.493
WaveNet	4.015	4.346	3.810	3.883	3.903	3.991
Parallel WaveGAN	3.421	3.489	3.406	3.276	3.382	**3.395**
PeriodNet	2.981	3.335	3.127	3.169	3.037	**3.130**

TABLE II
ROOT MEAN SQUARED ERROR OF F_0 FOR EACH METHOD. A, B, C, D, AND E MEAN THE EVALUATED FIVE SONGS.

Model	Evaluation data					Average
	A	B	C	D	E	
WORLD	6.926	23.21	13.70	7.613	19.70	14.23
WaveNet	16.84	13.08	10.94	17.72	7.795	13.28
Parallel WaveGAN	8.003	11.37	12.95	15.91	9.317	11.51
PeriodNet	8.966	13.12	11.72	10.76	11.90	**11.29**

C. PeriodNet

PeriodNet is a GAN-based, non-auto-regressive structure network proposed as a method for speech synthesis for singing voices. The excitation signal is input explicitly, and the periodic and aperiodic components are generated separately by two generators. Similar to Parallel WaveGAN, adversarial loss and STFT loss are used, and discriminators are trained separately for multiple sampling frequencies by downsampling the speech signal.

PeriodNet, like Parallel WaveGAN, does not have an autoregressive structure, so it can be synthesized quickly by generating multiple samples at the same time, and real-time generation is possible.

III. EXPERIMENTS

A. Experimental conditions

We evaluated the performance of the three neural vocoders described in Section II on an analysis synthesis task for operatic singing voice. In this task, we extract acoustic features from a speech waveform, and then, we re-synthesize the speech waveform from the extracted acoustic features using vocoders. We recorded 48 Japanese a cappella opera songs (about 93 minutes) sung by a female singer at a sampling frequency of 16 kHz. Of the 48 songs, 43 songs were used as training data and 5 songs as test data. Five songs out of the 43 songs were used as development data when training Parallel WaveGAN and PeriodNet. We used WORLD [2] to extract 53 dimensional acoustic features, containing the mel-cepstrum (50 dims), log-F_0(1 dim), aperiodic parameter (1 dim), and voice/unvoice flag (1 dim) from speech signals. We then synthesized speech signals from the acoustic features using a vocoder and evaluated the quality of the synthesized speech signals. In addition to the three neural vocoders, we also evaluated WORLD, which is a conventional source-filter vocoder.

B. Experimental results

We evaluated the synthesized speech signals using mel-cepstral distortion (MCD) [dB] and the root mean square error of F_0(F_0-RMSE) [Hz]. Table I and Table II show the results of each evaluation. Table I shows that the MCD of Parallel WaveGAN and PeriodNet are small. Since these two models introduce STFT loss, the estimation accuracy of the mel-cepstrum is considered to be high. Table II shows that all the models have small F_0-RMSE, but PeriodNet has especially small F_0-RMSE. PeriodNet is robust to the variation of F_0because the excitation signal is explicitly input to the model, and, as a result, it is considered to have high accuracy in reproducing F_0.

IV. CONCLUSION

In this study, we compared the synthesis performance of recently proposed neural vocoders for operatic singing voices. Future work includes the improvement of quality by adjusting hyper parameters, experiments with speech at a sampling frequency of 48 kHz, and experiments with subjective evaluation.

REFERENCES

[1] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[2] 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.

[3] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *arXiv:1910.11480*, 2019.

[4] Y. Hono *et al.*, "Periodnet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components," *arXiv:2102.07786v1*, 2019.

[5] K. Tachibana *et al.*, "An investigation of noise shaping with perceptual weighting for wavenet-based speech generation," in *Proc.ICASSP*, 2018, pp. 5664–5668.

[6] A. van den Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv:1711.10433*, 2017.