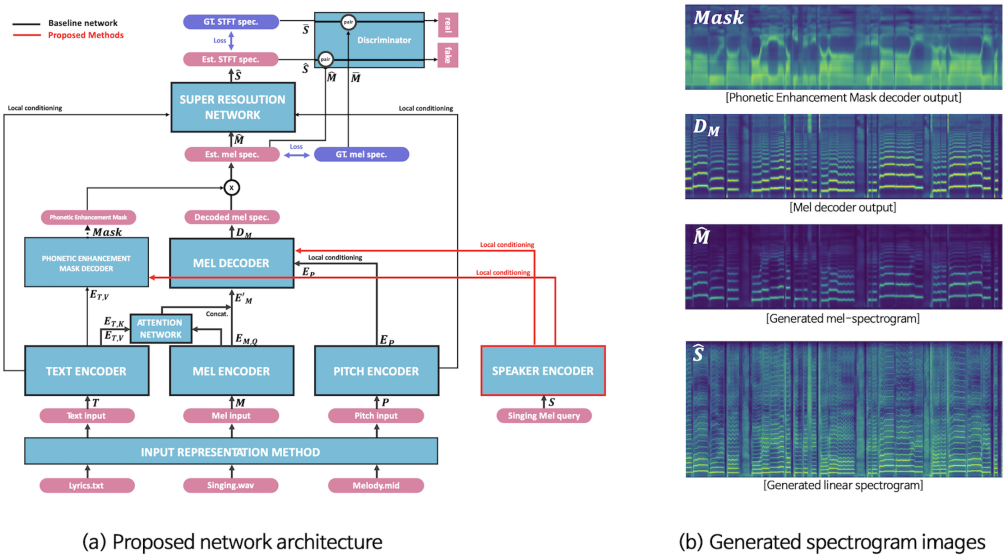
MULTI-SINGER SINGING VOICE SYNTHESIS SYSTEM

Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, Kyogu Lee
Music & Audio Research Group, Seoul National Universiy
{juheon2, kekepa15, vinyne, dg22302, kglee}@snu.ac.kr

EXTENDED ABSTRACT

In this study, we propose an end-to-end multi-singer singing voice synthesis (SVS) system. We conducted the study in a way to develop the pre-proposed single-speaker singing synthesis system [4] into a multi-singer system.

Our proposed system architecture is shown in Figure 1-(a). The input of the system consist of pitch, lyric, and speaker singing queries. Based on the given inputs, a mel-spectrogram is generated in an auto-regressive way. The generated mel-spectrogram is converted to a linear spectrogram via the super-resolution stage, and finally to waveform via griffin-lim algorithm [3].



(a) Proposed network architecture (b) Generated spectrogram images

Figure 1. Proposed network architecture and generated spectrogram images.

Our baseline SVS system differs from the existing singing synthesis studies in three respects. First, our network uses lyric information and pitch information independently when creating a mel-spectrogram. We assumed that implementing the principles of vocal organs proposed by the source-filter model [1] in a network structure would allow training data to be used more effectively for training, so we designed a phonetic enhancement mask decoder that only decode lyric information, and a mel decoder that decode pitch information. As a result, our network was able to train both information independently, as shown in figure 1-(b). Second, our network uses encoded text and pitch information conditioned to the super-resolution stage. We assumed that the mel-spectrogram generated by our network at the intermediate stage may be incomplete, which might cause a drop in the performance of the SR process. In order to solve this problem, we used a method to recycle not sufficiently conditioned pitch and lyric encoding information into the SR process as

local conditioning methods. As a result, our network improves pronunciation accuracy and sound quality of the generated result. Third, we applied the adversarial loss to the super-resolution stage. We used the GAN framework [2, 5] to make the results more realistic, and as a result, we were able to improve sound quality.

We have expanded the network to a multi-singer model by adding a query-based speaker encoder to the baseline network. The speaker encoder passes the given query mel-spectrogram through the two 1-d convolutional layers, and then maps it to embedding vector through an average time pooling layer and a fully-connected layer. Then, the speaker embedding vector is conditioned on two decoders and is involved in the generation process of the mel-spectrogram.

As we expanded the baseline network into a multi-singer model, we found the following points. First, the different speakers' timbre information affects the structure of the phonetic enhancement mask decoder output. As shown in figure 2, the mask decoder output for the same pronunciation generally has a similar structure, but we observed that it has a difference in the formant frequency as the speaker changes. Second, each speaker's singing style affects the generation of a mel-decoder output. Although the output of the mel-decoder generated for different singers is the same pitch structure, it can be confirmed that there are differences in terms of singing style, such as bending, vibrato, and breathing. We also confirmed that the singing voice was produced with the appropriate timbre & style for the given singing query.

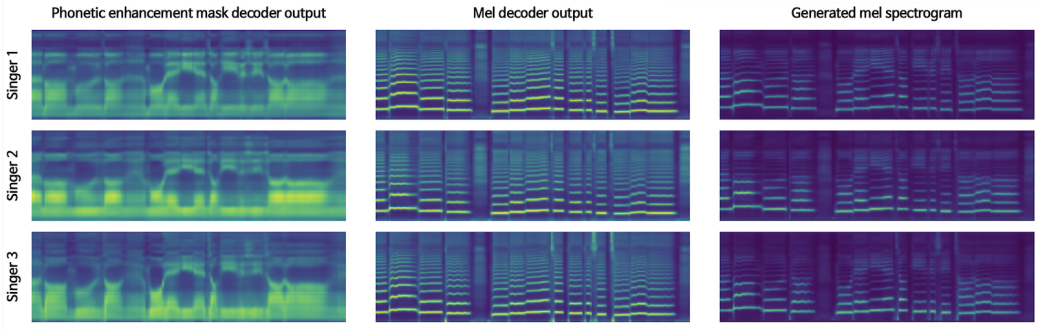


Figure 2. Generated spectrogram images of different singers.

The future direction of our research is as follows. First, we will collect a broad set of singing voice data, including a wide range of timbre and singing styles, and expand it to a vocal synthesizer that can control not only pitch, lyrics, but also styles and tones. Second, we plan to design a SVS system that extracts the timbre and singing styles from a given query and efficiently cloning them.

ACKNOWLEDGMENTS

This work has partly supported by National Research Foundation of Korea (NRF) funded by the Korea government (NRF-2017R1E1A1A01076284), and partly supported by SK T-brain.

REFERENCES

[1] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[4] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee. Adversarially Trained End-to-End Korean Singing Voice Synthesis System. In *Proc. Interspeech 2019*, pages 2588–2592, 2019.

[5] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.