HIGHLY EXPRESSIVE PEKING OPERA SYNTHESIS WITH DURIAN SYSTEM

Yusong Wu, Shengchen Li
Beijing University of Posts and Telecommunications
wuyusong@bupt.edu.cn
shengchen.li@bupt.edu.cn

Chengzhu Yu, Heng Lu
Tencent AI Lab
czyu@tencent.com
bearlu@tencent.com

Chao Weng, Dong Yu
Tencent AI Lab
cweng@tencent.com
dyu@tencent.com

EXTENDED ABSTRACT

Jingju, also known as Peking or Beijing opera, is the most dominant form of Chinese opera which combines music, vocal performance, mime, dance and acrobatics. It arose in Beijing in the mid-Qing dynasty (1636–1912) and became fully developed and recognized by the mid-19th century. There are mainly four role types in Jingju including sheng (gentlemen), dan (women), jing (rough men), and chou (clowns). Each of the four roles has its own singing and performing style, which makes the music and vocal performance in Jingju highly expressive. In order to preserve this beautiful traditional performing art, this paper aims to model and synthesize Jingju singing from its lyrics and song score. Distinctly different from pop music, current challenges in Jingju synthesis involves:

- Modeling and synthesizing highly expressive rhythm and intonation. There is no strong fixed tempo in Jingju singing, which means the duration of notes can vary a lot from score. Moreover, the intonation in Jingju singing is consist of the combination of the complex transitory results from the abundant use of grace notes and the vibratos that could sustain several seconds and has variable rates.
- Generating phoneme sequences given lyrics can be hard. The dialect and special expressive pronunciation in Jingju singing has different rules in grapheme-phoneme translation compared with normal Mandarin Speaking.
- Training on weakly labeled data. Although some of the data is already labeled at the phoneme level, most of the data available are annotated in word level, bringing challenges to for automatic data labeling and modeling with the weakly labeled data.
- Multiple factors include intonation, rhythm, timbre, loudness, and phonation all play crucial role in Jingju singing. All of them needs to be well handled to build a fair Jingju singing synthesis system.

In this paper, present a prototype Jingju synthesis methods based on our recently proposed DurIAN [1] (DurIAN Informed Attention Network) framework. DurIAN is an autoregressive model in which the alignments between the input text and the output acoustic features are inferred from a duration model. While DurIAN was originally proposed for speech synthesis, we find it very compelling to apply this framework for Jingju synthesis tasks. Specifically, the decoder network of DurIAN is used for predicting acoustic features frame by frame in autoregressive manner. As the autoregressive model does not suffer from over-smoothing problem, it can be very helpful for generating natural Jinju singing voice. At the same time, the duration model inside the DurIAN can be used for predicting the phoneme and not-pitch duration. The model architecture of DurIAN is shown in Fig. 1.

We use a series of Jingju a cappella singing dataset [2–6] and applied an automatic note alignment using melodic transcription with a genre-specific musicology model similar to one used in [7]. In our generation network, the input sequence which consists of a concatenated phoneme, note-pitch, singer and role type

embedding, is sent to a decoder network of DurIAN which outputs the mel-spectrogram. The audio is then generated using WaveRNN by taking the predicted mel-spectrogram as feature. We demonstrate that our current model is capable of generating singing with expressive intonation using only note information. Our currently finished portion with respect to the whole picture is shown inside the red box in Fig. 1.

Future works include developing a duration model that jointly generate phoneme and note duration, and a phoneme prediction model that could predict Jingju-specific singing phonemes given arbitrary Chinese characters.

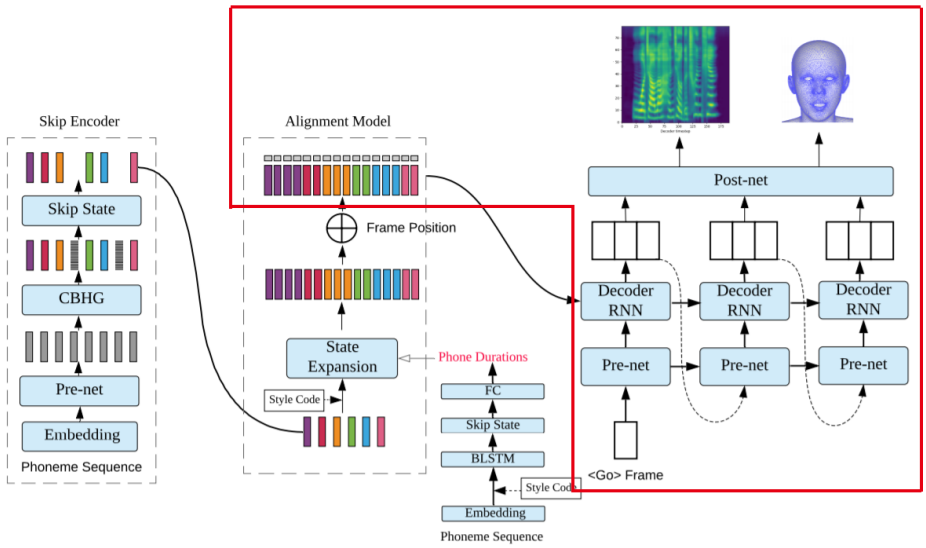

Figure 1. The model architecture of DurIAN, our currently finished works are indicated inside the red line.

ACKNOWLEDGMENTS

Yusong Wu is in placement in Tencent AI Lab.

REFERENCES

[1] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700*, 2019.

[2] Rong Gong, Rafael Caro, and Tiange Zhu. Jingju a cappella recordings collection, 2019.

[3] Rong Gong, Rafael Caro Repetto, Yile Yang, and Xavier Serra. Jingju a cappella singing dataset part1, 2018.

[4] Rong Gong, Rafael Caro Repetto, and Xavier Serra. Jingju a cappella singing dataset part2, 2018.

[5] Rong Gong and Xavier Serra. Jingju a cappella singing dataset part3, 2018.

[6] R. Caro Repetto, Rong Gong, Tiange Zhu, and X. Serra. Jingju music scores collection, 2019.

[7] Rong Gong, Yile Yang, and Xavier Serra. Pitch contour segmentation for computer-aided jinju singing training. In *Großmann R, Hajdu G, editors. SMC 2016. 13th Sound & Music Computing Conference; 2016 Aug 31-Sep 3; Hamburg, Germany. Hamburg: Hochschule für Musik und Theater Hamburg; 2016. p. 172-8.* Zentrum für Mikrotonale Musik und Multimediale Komposition (ZM4) Hochschule . . . , 2016.