LEARNING SINGING FROM SPEECH

Liqiang Zhang[*1], *Chengzhu Yu*[2], *Heng Lu*[2], *Chao Weng*[2], *Yusong Wu*[2],
Xiang Xie[1], *Zijin Li*[3], *Dong Yu*[2]

[1]Beijing Institute of Technology [2]Tencent AI Lab [3]China Conservatory of Music
{zhlq,xiexiang}@bit.edu.cn,{czyu,bearlu,cweng,ysw,dyu}@tencent.com,lizijin2019@hotmail.com

ABSTRACT

We propose an algorithm that is capable of synthesizing high quality target speaker's singing voice given only their normal speech samples. The proposed algorithm first integrate speech and singing synthesis into a unified framework, and learns universal speaker embeddings that are shareable between speech and singing synthesis tasks. Specifically, the speaker embeddings learned from normal speech via the speech synthesis objective are shared with those learned from singing samples via the singing synthesis objective in the unified training framework. This makes the learned speaker embedding a transferable representation for both speaking and singing. We evaluate the proposed algorithm on singing voice conversion task where the content of original singing is covered with the timbre of another speaker's voice learned purely from their normal speech samples. Our experiments indicate that the proposed algorithm generates high-quality singing voices that sound highly similar to target speaker's voice given only his or her normal speech samples. We believe that proposed algorithm will open up new opportunities for singing synthesis and conversion for broader users and applications.

Index Terms— Singing Synthesis, Singing Voice Conversion

1. INTRODUCTION

Singing is one of the most important music expression and the techniques of singing synthesis have many applications in entertainment industries. Over the past decades, many approaches have been proposed for singing synthesis. These include methods based on concatenative unit selection [1] as well as more recent approaches based on deep neural network (DNN) [2] and autoregressive generation models [3].

While existing singing synthesis algorithms are capable of producing natural singing, it normally requires a large amount of singing data for training new voices. Compared to normal speech data, singing data is much more difficult and expensive to collect. To address such limitation, more data efficient singing synthesis approaches [4] have been proposed recently, which adapts a multi-speaker trained singing synthesis model with a small amount of target speaker's singing data.

Alternatively, singing synthesis with new voices can be achieved through singing voice conversion. The task of singing voice conversion is to convert one's singing with the voice of another while keeping singing content the same. Traditional singing voice conversion [5, 6, 7] relies on parallel singing data to learn conversion function between different speakers. However, a recent study [8] on unsupervised singing voice conversion uses a WaveNet [9] based autoencoder architecture to achieve singing voice conversion without parallel singing data or even the transcribed lyrics or notes.

While data efficient singing synthesis approach [4] as well as unsupervised singing voice conversion method [8], could efficiently generate singing with new voices, it still requires a minimal amount of singing voice samples from target speakers. This has limited the applications of singing voice synthesis to relatively restricted scenarios where the target speaker's singing voice has to be available.

On the other hand, normal speech samples are much easier to collect than singing. However, there are only a few studies have investigated the use of speech samples for singing synthesis. The speech-to-singing synthesis method proposed in [10] attempts to convert a speaking voice to singing by directly modifying acoustic features such as f0 contour and phoneme duration in read speech. While speech-to-singing approaches could produce singing from read lyrics, it normally requires non-trivial amount of manual tuning of acoustic features for achieving high intelligibility and naturalness of singing voices.

In this paper, we propose an algorithm that directly synthesizes natural singing with target speakers' voice by learning their voice characteristics from speech samples[1]. The key part of proposed algorithm is to learn universal speaker embeddings, such that the speaker embeddings learned for the task of speech synthesis can be used for singing synthesis, and vice versa. For this purpose, we use our recently proposed

[*]Work performed while interning at Tencent AI Lab

[1]Sound demo of proposed algorithm can be found at https://tencent-ailab.github.io/learning_singing_from_speech

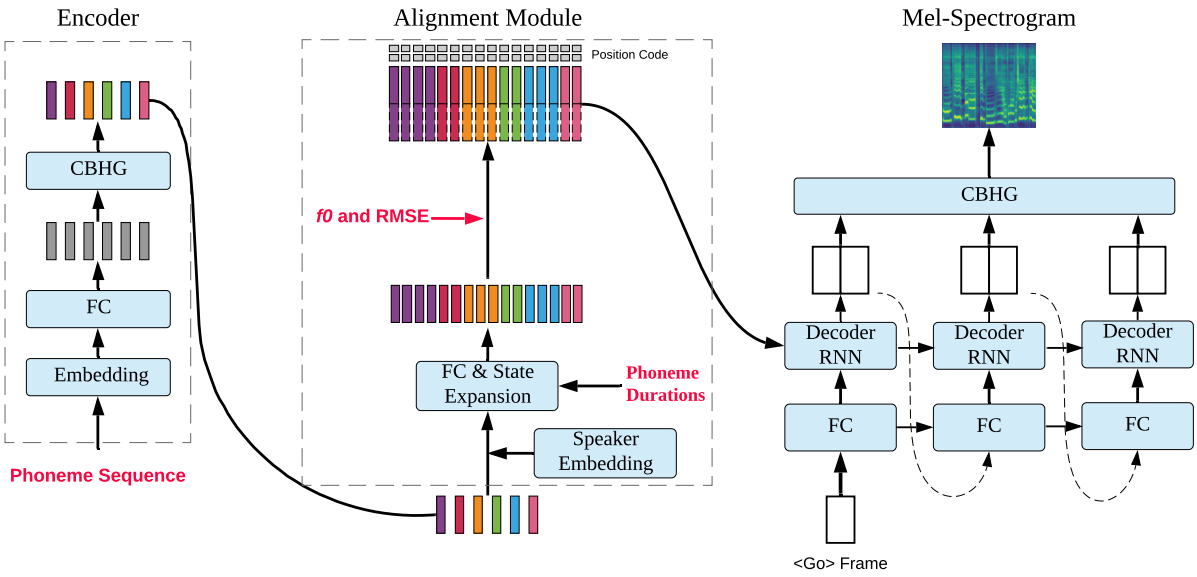

Fig. 1: Model architecture of DurIAN-4S.

autoregressive generation model, Duration Informed Attention Network (DurIAN) [11], for unifying text-to-speech and singing synthesis into a single framework. DurIAN, originally proposed for the task of multimodal synthesis, is essentially an autoregressive feature generation framework that could generates acoustic features (e.g., mel-spectrogram) from any audio source frame by frame. In proposed method, phoneme duration, fundamental frequency (F0) and root-mean-square energy (RMSE) are extracted from training data containing both singing or normal speech, and used as inputs for reconstructing target acoustic features. The entire model is trained jointly with learnable speaker embeddings as conditional input to the model. The trained model and speaker embeddings can be used to convert any singing into target speaker's voice by using his or her speaker embedding as conditional input.

The paper is organized as following. Section 2 introduces the architecture of our conversion model. Section 3 introduces the experiment. Section 4 and 5 are the conclusion and acknowledgements.

2. MODEL ARCHITECTURE

In this section, we first describe DurIAN based Speech and Singing Synthesis System (DurIAN-4S), a unified speech and singing synthesis system based on DurIAN. After that, we present singing voice conversion approach based on DurIAN-4S.

2.1. DurIAN-4S

While DurIAN was originally proposed for the task of multimodal speech synthesis, it is a general autoregressive framework that can be used for other synthesis tasks. The original DurIAN model is modified here to perform speech and singing synthesis at the same time. The major difference of DurIAN-4S compared to DurIAN is that it takes additional inputs. These additional inputs are attributes of singing that are useful for singing synthesis (music note, *f0*, etc.). As the focus of this study is singing voice conversion[2], we use frame level *f0* and root mean square energy (RMSE) extracted from original singing/speech as additional inputs[3] (Fig. 1).

The architecture of DurIAN-4S is illustrated in Fig. 1. It includes (1) an encoder that encodes the context of each phoneme, (2) an alignment model that aligns the input phoneme sequence and to target acoustic frames, (3) an autoregressive decoder network that generates target mel-spectrogram features frame by frame.

2.1.1. Encoder

We use phoneme sequence $\mathrm{x}_{1:\mathrm{N}}$ directly as input for both speech and singing synthesis. The output of the encoder $\mathrm{h}_{1:\mathrm{N}'}$ is a sequence of hidden states containing the sequential representation of the input phonemes as

$$\mathrm{h}_{1:\mathrm{N}} = \mathrm{encoder}(\mathrm{x}_{1:\mathrm{N}}), \quad (1)$$

where N is the length of input phoneme sequences[4].

[2]For the task of singing synthesis from note and lyrics, the note of music can be used as additional inputs.

[3]The *f0* and RMSE will not be available at inference time of speech synthesis to be used as additional inputs. But, our objective is singing voice conversion, and the model will not be used for speech synthesis inference.

[4]The state skipping structures in DurIAN [11] is not used here as it is not a necessary component for singing synthesis or conversion.

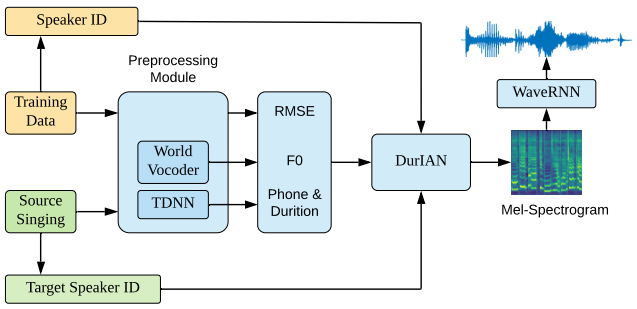

Fig. 2: The process diagram of training and converting. The yellow parts are used in training stage, the green parts are used in converting stage and the blue parts are used in both stages. The WaveRNN [12] model is trained separately.

2.1.2. Alignment model

The purpose of alignment model is to generate frame aligned hidden states that will be used as input for autoregressive generation. Here, the output hidden sequence from encoder $\mathrm{h}_{1:\mathrm{N}'}$ is first concatenated with speaker embedding $\mathrm{m_s}$ and followed by a fully connect layer used for dimension reduction as

$$\mathrm{h}^{'}_{1:\mathrm{N}} = \mathrm{fc}(\mathrm{h}_{1:\mathrm{N}} \vee \mathrm{m_s}) \tag{2}$$

where $\vee$ indicates concatenation and m_s indicates the embedding of speaker s. The output hidden states after dimension reduction layer will be expanded according to the duration of each phoneme as

$$\mathrm{e}_{1:\mathrm{T}} = \mathrm{state\_expand}(\mathrm{h}^{'}_{1:\mathrm{N}}, \mathrm{d}_{1:\mathrm{N}}), \tag{3}$$

where T is the total number of input audio frames. The state expansion is simply the replication of hidden states according to the provided phoneme duration. The duration of each phoneme is obtained from force alignments performed on input phonemes and acoustic features. The frame aligned hidden states $\mathrm{e}_{1:\mathrm{T}}$ is then concatenated with frame level *f0*, RMSE, and relative position of every frame inside each phone.

$$\mathrm{e}^{'}_{1:\mathrm{T}} = \mathrm{e}_{1:\mathrm{T}} \vee \mathrm{f}_{1:\mathrm{T}} \vee \mathrm{r}_{1:\mathrm{T}} \vee \mathrm{p}_{1:\mathrm{T}} \tag{4}$$

where $\mathrm{f}_{1:\mathrm{T}}$ and $\mathrm{r}_{1:\mathrm{T}}$ represents *f0* and RMSE for each frame respectively. And $\mathrm{p}_{1:\mathrm{T}}$ is the position code of each frame.

2.1.3. Decoder

The decoder is the same as in DurIAN, composed of two autoregressive RNN layers. Different from the attention mechanism used in the end-to-end systems, the attention context is computed from a small number of encoded hidden states that are aligned with the target frames, which reduce artifacts observed in the end-to-end system. We decode two frames per time step in this paper. The output from the decoder network $\mathrm{y}^{'}_{1:\mathrm{T}}$ is passed through a post-CBHG [13] to improve the quality of predicted mel-spectrogram as

$$\mathrm{y}^{'}_{1:\mathrm{T}} = \mathrm{decoder}(\mathrm{e}^{'}_{1:\mathrm{T}}) \tag{5}$$

$$\hat{\mathrm{y}}_{1:\mathrm{T}} = \mathrm{cbhg}(\mathrm{y}^{'}_{1:\mathrm{T}}) \tag{6}$$

The entire network is trained to minimize the mel-spectrogram prediction loss before and after post-CBHG as

$$\mathrm{L} = \sum_{\mathrm{i}=1}^{\mathrm{T}} |\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| + \sum_{\mathrm{i}=1}^{\mathrm{T}} |\mathrm{y_i} - \mathrm{y}^{'}_\mathrm{i}| + \mathrm{l2loss} \tag{7}$$

where $l2loss$ represents l2 regularization.

2.2. Singing Voice Conversion

The whole process of our method is illustrated in Fig. 2. The training dataset contains a multi-speaker speech and singing corpus. For singing voice conversion task, the target speaker or singer should be included in the training data, while the source singing or singer to be converted doesn't have to be seen in the training. The preprocess module mainly consists of two parts: the TDNN based phoneme alignment model [14] and the world vocoder [15]. The TDNN model is a component of a pre-trained general speech recognition model, which generates the phoneme sequence and its duration alignment from speech and singing data. The world vocoder is used to extract F0 which reflects the rhythm and melody of singing. Because the F0 envelope also determines the tone of each phone, so we use the non-tonal phones in our experiment. In addition, we also found that the RMSE can greatly improve the quality and stability of singing voice conversion. The input of DurIAN-4S is phoneme sequence, phoneme durations, *f0*, RMSE and speaker identity. The training target of DurIAN-4S is to reconstruct the mel-spectrogram. In the training stage, embeddings of speakers with speech samples and singing samples are all also optimized jointly.

After the model of DurIAN-4S is trained, it can be used to convert any singing to a target speaker's voice. The process of singing voice conversion is that, we first extract the f0, phoneme duration, and RMSE from the preprocess module, and use these as input for singing generation. By choosing different speaker embedding during singing generation, we could produce singing with different voice. The generated mel-spectrogram from DurIAN-4S after conversion will be used for WaveRNN [12] model for waveform generation.

When conversion between male and female, the input *F0* should be multiplied by a scalar ν as:

$$\nu = \frac{\sum_i^N mean(x_i^t)}{N \cdot mean(x^s)} \tag{8}$$

where x^s is the source singing, x_i^t is the target speaker t, $mean$ is the average F0 of vowel phone in the audio. However, the pitch of one's singing is usually higher than the pitch

of speech from the same person and it is common to adjust the key of songs within a certain range for different singers. We could control the scalar ν to get a flexible conversion performance.

3. EXPERIMENT

3.1. Dataset

The training set contains the Tencent multi-speaker speech corpus (TSP) and the Tencent singing corpus (TSG). In TSP corpus, we choose 3 male speakers and 4 female speakers, each with 1.5 hours of data. The TSG corpus contains a total of 28 hours singing data recorded by 3 female singers. For singing voice conversion task, we choose source singing from a separate singing corpus, which will not be used in training. All the data has a sampling rate of 24K.

3.2. Model Parameters

In our experiment, the dimensions of the phoneme embedding, speaker embedding, encoder CBHG module, attention layer are all 256. The decoder has 2 GRU layers with 256 dimension and the batch normalization is used in the encoder and post-net module. We use Adam optimizer and 0.001 initial learning rate with warm-up [16] schedule. There is a total of 250,000 steps with a batch size of 32 to converge the model. We found multi-speaker trained WaveRNN model will improve the synthesis stability in this singing voice conversion task.

3.3. Quality and Similarity Evaluation

Since we are not able to find any public benchmarks on speech based singing voice conversion, we compared it singing voice conversion based on singing samples. Both the quality of converted singing voice and similarity of converted singing voice and target speaker's voice is compared. Subjective evaluation with Mean Opinion Scores (MOS) is used. A total of 14 subjects have been participated in our listening tests.

We select 20 segments from two different songs from a separate singing corpus. Three male speakers and two female speakers from TSP corpus are selected as target speakers, and 1 singer from TSG corpus as target singer. We perform the experiments conduct an ablation study on the importance of using RMSE for singing voice conversion. For the timbre similarity evaluation, the subjects are asked to score a similarity of voice timbre between converted singing and target speaker's normal speech.

The similarity evaluation results are shown in Table 1. The scale of MOS is set between 1 to 5 with 5 being the highest score. We first compare the effects of RMSE as additional input for singing voice conversion. And the results show that using RMSE improves both the quality and similarity significantly. We found that the energy information of each frame concatenated with F0 could help the model learning the pronunciation of long vowels. The energy of each frame indicates the loudness of pronunciation, helping the model to determine when vowels should stop properly.

Table 1: MOS for Singing Conversion Quality and Similarity. Target speaker type indicates what types of samples we used for singing voice conversion from target speaker. 'singing' means the singing samples from target speaker are used for singing voice conversion, and 'speech' means speech samples from target speaker are used for singing voice conversion.

Method	Target Speaker	Naturalness	Similarity
f0	singing	3.23	3.00
f0	speech	2.77	2.84
f0 + RMSE	singing	3.80	3.65
f0 + RMSE	speech	3.42	3.49

We also compare the performance of singing voice conversion using target speaker's normal speech versus using their singing samples. Singing voice conversion using target speaker's singing samples receives better score than that of using their speech samples. This is expected performance, as it is much easier to learn speaker's singing voice from their singing samples than speech samples. However, we could see that the similarity of singing voice conversion using speech samples are not too far off, showing that proposed algorithm could synthesis target speaker's singing voice, in both high quality and similarity, with only speech samples. The samples used in our experiments can be found at https://tencent-ailab.github.io/learning_singing_from_speech.

4. CONCLUSION

In this paper, we proposed an algorithm that synthesizes natural singing in target speaker's voice given only their normal speech samples. We evaluate proposed algorithm on singing voice conversion task with speech samples, and obtained very promising results. In future work, we will focus on reducing the amount of target speech samples for both target singing synthesis and conversion tasks.

5. ACKNOWLEDGEMENTS

The authors would like to thanks Chunlei Zhang, Dongxiang Xu and other members in the Tencent AI Lab team for providing suggestions on model structure and optimization.

6. REFERENCES

[1] Jordi Bonada, Martí Umbert, and Merlijn Blaauw, "Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016.," in *INTERSPEECH*, 2016, pp. 1230–1234.

[2] Masanari Nishimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "Singing voice synthesis based on deep neural networks.," in *Interspeech*, 2016, pp. 2478–2482.

[3] Merlijn Blaauw and Jordi Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, pp. 1313, 2017.

[4] Merlijn Blaauw, Jordi Bonada, and Ryunosuke Daido, "Data efficient voice cloning for neural singing synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6840–6844.

[5] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Fernando Villavicencio and Jordi Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Eliya Nachmani and Lior Wolf, "Unsupervised singing voice conversion," *arXiv preprint arXiv:1904.06590*, 2019.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] Takeshi Saitou, Masataka Goto, Masashi Unoki, and Masato Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 215–218.

[11] 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.

[12] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," *CoRR*, vol. abs/1802.08435, 2018.

[13] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[14] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] Masanori Morise, Fumiya Yokomori, and Kenji 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.

[16] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.