Peking Opera Synthesis via Duration Informed Attention Network

Yusong Wu[1], Shengchen Li[1], Chengzhu Yu[2], Heng Lu[2], Chao Weng[2], Liqiang Zhang[2], Dong Yu[2]*

[1]Beijing University of Posts and Telecommunications   [2]Tencent AI Lab

{wuyusong, shengchen.li}@bupt.edu.cn,
{czyu, bearlu, cweng, tatelqzhang, dyu}@tencent.com

Abstract

Peking Opera has been the most dominant form of Chinese performing art since around 200 years ago. A Peking Opera singer usually exhibits a very strong personal style via introducing improvisation and expressiveness on stage which leads the actual rhythm and pitch contour to deviate significantly from the original music score. This inconsistency poses a great challenge in Peking Opera singing voice synthesis from a music score. In this work, we propose to deal with this issue and synthesize expressive Peking Opera singing from the music score based on the Duration Informed Attention Network (DurIAN) framework. To tackle the rhythm mismatch, Lagrange multiplier is used to find the optimal output phoneme duration sequence with the constraint of the given note duration from music score. As for the pitch contour mismatch, instead of directly inferring from music score, we adopt a pseudo music score generated from the real singing and feed it as input during training. The experiments demonstrate that with the proposed system we can synthesize Peking Opera singing voice with high-quality timbre, pitch and expressiveness.

Index Terms: singing synthesis, expressive singing synthesis, machine learning, deep learning, Lagrange multiplier

1. Introduction

Peking Opera, also known as Beijing Opera or Jingju, is Chinese traditional performing art which combines music, vocal performance, mime, dance and acrobatics. Singing in Peking Opera has various styles, each widely different depending on different role type and music styles. Strong personal styles also make the actual singing can be different from the given music notes. Like a dialect to Mandarin, it even has its unique way of pronunciation. Moreover, melody in singing often consist of arias with variation of complex transitory and vibratos, which makes the singing very expressive and difficult to learn. Another difference from normal singing is the note length has a great variance, sometime very long note can appear (can be more than 10 seconds). All above factors makes it very challenging to modelling and generating Peking Opera singing comparing to normal singing.

Although there are few works focusing on the synthesis of Peking Opera, or more broadly, opera, the synthesis of singing voice has been researched since 1962 when Kelly and Lochbaum [1] used an acoustic tube model to synthesis singing voice with success. Recently, several works [2–7] use deep neural networks to synthesis singing voice which, known as parametric systems, process fundamental frequency (or pitch contour, f_0) and harmonics features (or timbre) separately. As a typical case among such systems, Neural Parametric Singing Synthesizer (NPSS) [2] using a phoneme timing model, a pitch model and a timbre model each consist a set of neural networks

to generate acoustic parameters of the singing. In NPSS, a Fitting Heuristic method is introduced to eliminate the mismatch between music note duration and the predicted phoneme duration. However, Fitting Heuristic method is totally rule based and it requires to locate the principal vowel before adjusting phoneme duration. This maybe acceptable in most English or Japanese singing cases, but can cause huge duration error when synthesizing Peking Opera. Different from normal speech or singing, in Peking Opera, one syllable can last very long time and contains a long sequence of phonemes, e.g. "l-j-E-a-a-N". More importantly, one can't simply tell which phoneme amongst all these phonemes is the principle phone. There could be multiple equally important phonemes in Peking Opera singing.

To better synthesize the expressive Peking Opera, this paper proposes a Peking Opera singing synthesis system based on Duration Informed Attention Network (DurIAN) [8]. The main contribution in this study lies in the two following points: 1) To tackle with rhythm mismatch between music note duration and the predicted phoneme duration, contextual based mixture density networks (MDN) [9] followed by a Lagrange Multiplier optimization is proposed and implemented for duration modelling. This method is completely data-driven, and more importantly, skips the step of locating the principle phoneme from the conventional Fitting Heuristic method. 2) To deal with the melody mismatch between original music score and the actual singing, and also to better model the expressive variations and vibratos in Peking Opera, a pseudo music score is generated from the real singing and fed as input during DurIAN [8] model training. Experimental Results show proposed duration modeling and prediction method outperforms the Fitting Heuristic method by a large margin. And the generated pitch contours also demonstrate our system's ability to synthesize the singing variations and vibratos in Peking Opera.

The following sections of this paper are organized as follows. Firstly, the proposed model architecture is introduced. Next, proposed Lagrange Multiplier-based duration prediction and pseudo score generation are introduced in Section 2. In section 3 experiments are conducted based on a unique Peking Opera database. Finally, a quick discussion and conclusion is given in Section 4.

2. Methods

In proposed system, three efforts made it possible to synthesize expressive Peking Opera singing from music score: 1) DurIAN [8] based auto-regressive framework is used to generate output spectrogram features, with generated pitch as condition. 2) With the constraint of Lagrange Multiplier, a contextual dependent mixture density network (MDN) based phoneme duration modeling and generation method is introduced for accurate duration modelling. And 3) a melody transcription method is used to obtain a pseudo score from training waveform sam-
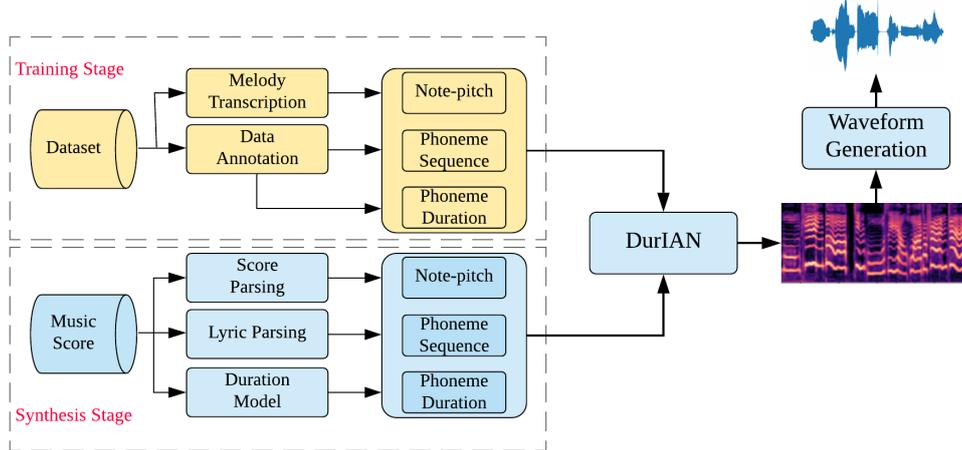

Figure 1: *The overview diagram of the proposed system. The proposed system is trained using data annotation and melody transcription result. In the synthesis stage, note-pitch and phoneme sequence are parsed from score, while the phoneme duration is predicted by the proposed duration model.*

ples in training process. This step is to ensure the extracted acoustic feature and the input music score consistent in training as well as in inference. Fig. 1 shows the training and synthesis diagram. In training stage, note-pitch of the singing is provided by melody transcription results while phoneme sequence and phoneme duration are obtained from data annotation. While in the synthesis, input note-pitch is from real music score and input phoneme sequence comes from lyric parsing module which also use music score as reference.

2.1. Spectrogram Generation

Duration Informed Attention Network (DurIAN) [8] structure is used here as main frame. Similar to DurIAN, in proposed system, the features of temporal dependency in phoneme sequences is first modelled by a **phoneme encoder** consist of a two-layer prenet followed by a CBHG module [10] (1-D convolution bank + highway network+ bidirectional GRU) for learning contextual dependencies of phonemes. The encoded phoneme sequence is aligned with output spectrogram in **alignment model** where singer identity and frame-wise generated pitch are further added. The final Mel-spectrogram frames are generated by a **decoder** with Gated Recurrent Unit (GRU) [11] in an auto-regressive manner where the temporal dependencies of the Mel-spectrogram are modelled, and a post-net is used for refining the output. Different from DurIAN, a CBHG module is used as a post-net instead of a fully convolution network to improve generation quality.

In the training stage, phoneme sequences are obtained from data annotation. Ground-truth phoneme duration is used in phoneme sequence expanding in alignment model. The input music note is generated from real singing by the melody transcription module which is introduced in section 2.3. While in the synthesis stage, the phoneme duration is predicted by the trained duration model introduced in section 2.2 and the input music note is from the real music score.

2.2. Phoneme Duration Modelling

In the music score, the duration of each note and also the lyric aligned with note is given. However, the duration of each phoneme needs to be predicted in order to generate syllable or word pronunciation during synthesis stage. Thus, a phoneme

duration model is trained in training stage and used to generate phoneme duration in generation process. Unlike in speech synthesis where no music score exists, in singing synthesis, the sum of predicted phonemes duration in a note should be always equal to the duration of the note. This is to make sure synthesized singing follows the rhythm on the score. To ensure the duration of the phoneme adds up to the note duration, Fitting Heuristic method is introduced in [2, 12]. However, Fitting Heuristic duration scaling relies heavily on the selection of primary vowel and r_0. In case of Peking Opera singing, there are prevailing use of very long notes. When scaling phoneme duration for a very long note, and sometimes the note can contain more than 5 equally important phonemes, the incorrect choice of primary vowel will lead to huge phoneme duration prediction errors. Instead, our system generates phoneme duration with Lagrange multiplier constraint which has been used in speech synthesis [13–15]. All the scaling or compensate factors are generated through data-driven methods other than based on rules.

Bi-directional Long Short Term Memory (Bi-LSTM) networks [16, 17] with mixture density output layer [9] is used as out phoneme duration model. With a total of M phonemes in the given note from music score, the duration distribution $\mathcal{L}(d_i, \lambda)$ for each phoneme can be represented as

$$\mathcal{L}(d_i, \lambda) = \sum_{k=1}^{K} \Pi_i^k N(\mu_i^k, (\sigma_i^k)^2) \qquad (1)$$

and

$$\sum_{i=1}^{M} d_i = T, \ \ i = 1, 2, ..., M \qquad (2)$$

where K is the total number of Gaussian Mixtures for each phone. Π_i^k, μ_i^k and σ_i^k are the weight, mean and standard deviation for the kth mixture of the ith phoneme respectively. λ indicates the total MDN parameter set mentioned above. In training stage, the parameters in proposed Bi-LSTM based mixture density network duration model are learned through back-propagating the negative log-likelihood error.

In synthesis stage, to maximize log MDN likelihood under the constraint of given total word duration, a Lagrange Multiplier is introduced and the task becomes to solve the following equation:

$$\nabla_{d_i,\alpha}[\sum_{i=1}^{M} log(\mathcal{L}(d_i,\lambda)) - \alpha(\sum_{i=1}^{M} d_i - T)] = 0 \qquad (3)$$

where α is the Lagrange Multiplier factor.

Because there is no close form solution for the Multi-Gaussian distribution maximum likelihood optimization problems, to simply the form of final solution, we always choose the Gaussian with the maximum weight instead of using all mixtures when calculating. Thus,

$$d_i \sim \mathcal{L}(d_i, \{\mu_i^k, \sigma_i^k\}) \quad \text{where} \quad k = \underset{k}{\mathrm{argmax}}\, \Pi_i^k \qquad (4)$$

Let $\hat{\mu}_i$, $\hat{\sigma}_i$ be the Gaussian means and standard deviations with the largest weight for the ith phone, and the solution for Eq. 3 is formulated as:

$$d_i = \hat{\mu}_i + \hat{\sigma}_i^2 \cdot \alpha \quad \text{where} \quad \alpha = \frac{T - \sum_{i=1}^{M} \hat{\mu}_i}{\sum_{i=1}^{M} \hat{\sigma}_i^2} \qquad (5)$$

As a comparison, if there's no Lagrange Multiplier constraint in Eq. 3, the result for the maximum likelihood will be only the means for the Gaussian $\hat{\mu}_i$. The equation above can be interpreted as scaling of each phoneme duration prediction $\hat{\mu}_i$ by a compensation term α which is decided by 1) the gap between music score note duration and sum of predicted phoneme duration, and 2) its standard deviation. And that indicates phonemes with large variance, such as some vowels, can be extend or compress more than the ones with smaller variance. This result is actually consistent with the duration scaling rules in Fitting Heuristic, but more generally formulated, and obtained through purely data-driven method.

2.3. Melody Transcription

Although published version of the Peking Opera score is available, the improvisation and expressiveness of the singer makes the actual singing inconsistent with the score. Efforts have been made to build Peking Opera singing synthesis system directly from original music scores, but the model turned out to be badly trained, and the generation results were with poor quality. Alternatively, a pseudo music score is automatically transcribed from actual Peking Opera singing to replace original music score in training stage. Same melody transcription method proposed in [18] is used here. The melody transcription method is based on a Hidden Markov Model [19] based pitch tracking module which uses probabilistic YIN [20] pitch estimation. First, frame-wise fundamental frequency of the singing with temporal smoothness is estimated by probabilistic YIN [20] pitch estimation where a Hidden Markov Model is applied to decode from multiple pitch candidates. Then, the estimated pitch track is fed into another Hidden Markov Model to render frame-wise discrete note pitch. The Hidden Markov Model used in transcription process contains 3-state consisting of Attack, Stable and Silent. When transitioning from current note to the next note, probabilities of note transition are calculated by a note transition probability function. A Peking Opera genre-specific note transition distribution are used for better note pitch decoding, achieving a note transcription F-score of 0.73. The output of the transcription algorithm is frame-wise discrete note pitch of the singing, where possible pitch value ranges from MIDI pitch 35 (B1) to MIDI pitch 85 (C#6) with 3 steps per semitone. Here, note pitch output are further quantified to integral MIDI

pitch, and pitch of the unvoiced frame is set to 0. An example of the transcribed melody is shown in Fig.2, where the orange line shows the f_0 of the singing pitch and the blue dotted line indicates the transcribed result.
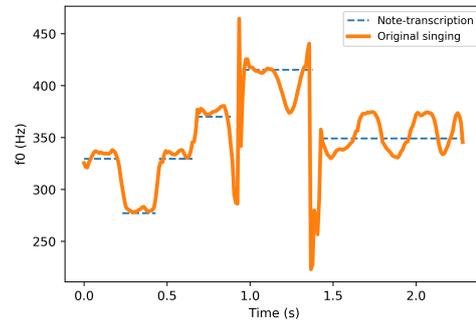

Figure 2: *A example of the melody transcription results on Peking Opera singing.*

3. Experiments and results

To demonstrate the effectiveness of proposed methods, two experiments are conducted. In phoneme duration prediction experiment, the objective predicted phoneme duration error is compared between proposed duration model and Fitting Heuristic method. In Peking Opera synthesis experiment, generated pitch contours are drawn and subjective Mean Opinion Score (MOS) test show our system is capable of synthesizing Peking Opera with fair quality.

3.1. Data

Although being popular for centuries, Peking Opera has received little effort in research field. To collect Peking Opera training data is not a trivial task. One of the available data set is the "Jingju a cappella singing" [21–24]. "Jingju a cappella singing" contains over 10 hours of a cappella Peking Opera recordings and annotations, but amongst which only 2 hours of recordings have been phoneme annotated. The 2 hours of recordings with phoneme annotation contains 71 Peking Opera singing fragments. And the annotation of phoneme adopts a modified X-SAMPA (Extended Speech Assessment Methods Phonetic Alphabet) phoneme set containing 51 phonemes. The singing data is further segmented according to line boundary, resulting a set of 606 singing phrases. In the experiment part, all experiments are conducted using the 2 hours of the annotated dataset.

3.2. Phoneme Duration Prediction

In order to compare our Lagrange Multiplier-based phoneme duration prediction method with Fitting Heuristic based methods, six phoneme duration prediction models are implemented and compared. First, the duration model used in NPSS system [2] is implemented (**CNN-Softmax-Fit**) which consist of 1d-CNNs and a softmax output predicting duration which are discretized to 50 log scale bins. To compare different duration scaling method, the softmax output in **CNN-Softmax-Fit** is then replaced with mixture density output with Lagrange Multiplier for duration optimization (**CNN-MDN-Lag**). Next, the duration model proposed in this paper is built (**BiLSTM-MDN-Lag**) which used a BiLSTM with mixture density out-

Table 1: *The mean phoneme duration prediction error in number of frames for different duration model.*

	all	notes $< 2s$
BiLSTM-MDN-Lag (proposed)	**8.91**	**3.24**
BiLSTM-MSE-Fit	13.63	4.23
BiLSTM-MDN-Fit	17.15	4.16
CNN-MDN-Lag	14.74	4.75
CNN-Softmax-Fit [2]	19.61	5.83
Mean-Fit [12]	18.69	5.52

put and Lagrange Multiplier optimization. For comparison, the output layer of the proposed duration model is replaced with a scalar prediction and Fitting Heuristic scaling while trained using mean square error (MSE) criterion (**BiLSTM-MSE-Fit**). To further compare the duration scaling method, the proposed system using Fitting Heuristic scaling instead of the Lagrange Multiplier optimization is built (**BiLSTM-MDN-Fit**) which use the mean of the predicted distribution as duration prediction. Last, the duration prediction method originally used in [12] is implemented (**Mean-Fit**) which use the mean duration of each phoneme as a look-up table for duration prediction and Fitting Heuristic for scaling.

In the experiment, the frame length is set to 10ms. 64 Peking Opera singing phrases are randomly chosen from database for test, while the rest of the data is used for training. The CNN blocks are built according to [2]. All Bidirectional-LSTMs use 2 hidden layers and has a 256-dimensional hidden layer size with training dropout rate of 0.5. And all Fitting Heuristic scaling set the second phoneme as the primary vowel. The number of Gaussian mixtures predict by model is set by 2 which we find yields best results.

The results are shown in Tab. 1 with average duration prediction error per phone, in terms of number frames. For in Peking Opera singing, there could be extremely long notes which may last for seconds. Thus when counting the results, two kind of results are counted: 1) the music score notes shorter than 2 seconds case, and 2) all notes cases. This objective average phoneme prediction error results reveal the proposed duration model outperforms other methods, achieving minimum prediction error. Probably because the dynamic range of phoneme duration in Peking Opera is much larger than normal speech or singing, 50 discrete log scale bins is too coarse as the prediction target and thus introduces largest prediction error in **CNN-Softmax-Fit**. From the results, we can see propose Lagrange Multiplier-based phoneme duration scaling consistently outperforms the Fitting Heuristic scaling by a large margin. And our proposed Mixture Density Network based Lagrange Multiplier phone duration generation method **BiLSTM-MDN-Lag** renders the best performance. It is worth noting that **BiLSTM-MDN-Fit** outperforms **BiLSTM-MSE-Fit** in predicting phoneme duration for shorter notes while perform worse when including long notes. Further analysis shows this happens when the music note is long, and when Fitting Heuristic scaling is employed, the generated phoneme duration is dominant by the stretching length. Moreover, when the principle vowel is locating wrong, the prediction error can be huge.

3.3. Peking Opera Synthesis

3.3.1. Subjective MOS score

In order to evaluate the subjective naturalness of the proposed Peking Opera synthesis system, Mean Opinion Score test is conducted where test participants are asked to score the singing samples from 1-5 according to the naturalness of the singing. 1 stands for most unnatural and unintelligible, 5 stands for most natural and almost the same as sung by real people. 2,3,4 are somewhere between. Peking Opera music scores in the format of MusicXML are used as score input which first parsed into notes and corresponding phonemes then input to the duration model to get each duration of the phonemes. Using the predicted phoneme duration along with notes and phonemes as input, singing samples are generated by the proposed system, and MOS test is conducted using the generated singing. In result, a 3.34 of MOS score is obtained, which indicates that the proposed system can generate Peking Opera singing with fair quality. Generated audio samples can be found on the web page: https://lukewys.github.io/files/Peking-Opera-Synthesis-2020.html.

3.3.2. Generated Pitch Contour

The generated f_0 contour from the synthesized singing is drawn with input music notes in Fig. 3. Different from the pitch contour generated by normal TTS or singing systems, generated pitch contours by our proposed system in Peking Opera show more up and downs and with larger variation. The vibratos and transitions curves in Fig. 3 is consistent with Peking Opera singing style and shows the ability of proposed system to generate the expressiveness in Peking Opera.
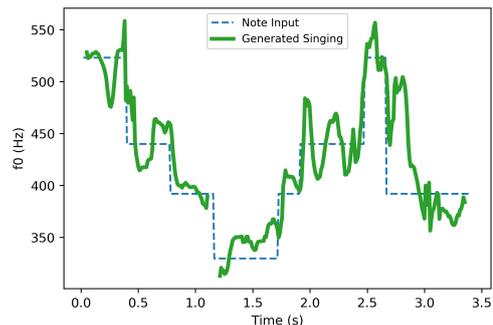

Figure 3: *generated Peking Opera singing f_0 contour by proposed system*

4. Conclusions

Improvisation and expressiveness in Peking Opera singing makes it extremely difficult to synthesize this classical performing art. With proposed MDN-based phoneme duration generation with Lagrange Multiplier optimization, our system can generate more accurate phoneme duration compared to the Fitting Heuristic phoneme duration scaling method. Pseudo music notes are generated through the melody transcription algorithm to solve the score inconsistency problem in training. Both the objective average predicted phoneme duration error and the generated pitch contour show our system performances well in generating Peking Opera singing. And as one can see from MOS and the generated samples that there is still a gap between the generated singing and the real performance in terms of naturalness. Our further work includes collecting and labelling more Peking Opera singing data, conducting MOS test in larger scale with subjects in musical background, and improving the quality and pitch accuracy of the generated singing.

5. References

[1] C. Lochbaum and J. Kelly, "Speech synthesis," in *Proceedings of the Speech Communication Seminar*, 1962, pp. 583–596.

[2] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[3] Y.-H. Yi, Y. Ai, Z.-H. Ling, and L.-R. Dai, "Singing voice synthesis using deep autoregressive neural networks for acoustic modeling," *arXiv preprint arXiv:1906.08977*, 2019.

[4] K. Kaewtip, F. Villavicencio, F.-Y. Kuo, M. Harvilla, I. Ouyang, and P. Lanchantin, "Enhanced virtual singers generation by incorporating singing dynamics to personalized text-to-speech-to-singing," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6960–6964.

[5] Y. Wada, R. Nishikimi, E. Nakamura, K. Itoyama, and K. Yoshii, "Sequential generation of singing f0 contours from musical note sequences based on wavenet," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 983–989.

[6] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, "Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan," *arXiv preprint arXiv:1903.10729*, 2019.

[7] Y. Hono, S. Murata, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Recent development of the DNN-based singing voice synthesis system — sinsy," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, nov 2018.

[8] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[9] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[12] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7229–7233.

[13] J. Yamagishi, "An introduction to hmm-based speech synthesis," *Technical Report*, 2006.

[14] H. Zen, T. Masuko, K. Tokuda, T. Yoshimura, T. Kobayasih, and T. Kitamura, "State duration modeling for hmm-based speech synthesis," *IEICE transactions on information and systems*, vol. 90, no. 3, pp. 692–693, 2007.

[15] H. Lu, Y.-J. Wu, K. Tokuda, L.-R. Dai, and R.-H. Wang, "Full covariance state duration modeling for hmm-based speech synthesis," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 4033–4036.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] R. Gong, Y. Yang, and X. Serra, "Pitch contour segmentation for computer-aided jinju singing training," in *2016 Sound & Music Computing Conference (SMC)*, 2016, pp. p. 172–8.

[19] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the tony software: Accuracy and efficiency," *2015 Sound & Music Computing Conference (SMC)*, 2015.

[20] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.

[21] R. Gong, R. Caro, and T. Zhu, "Jingju a cappella recordings collection," https://doi.org/10.5281/zenodo.3251760, 2019, accessed: 2020-05-01.

[22] R. Gong, R. C. Repetto, Yile Yang, and X. Serra, "Jingju a cappella singing dataset part1," https://doi.org/10.5281/zenodo.1323561, 2018, accessed: 2020-05-01.

[23] R. Gong, R. C. Repetto, and X. Serra, "Jingju a cappella singing dataset part2," https://doi.org/10.5281/zenodo.1421692, 2018, accessed: 2020-05-01.

[24] R. Gong and X. Serra, "Jingju a cappella singing dataset part3," https://doi.org/10.5281/zenodo.1286350, 2018, accessed: 2020-05-01.

[25] D. A. Black, M. Li, and M. Tian, "Automatic identification of emotional cues in chinese opera singing," *2014 Int. Conf. on Music Perception and Cognition (ICMPC)*, p. pp.250–255, 2014.

[26] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 55–73, 2015.