HuangmeiSinger: A Dataset and A Branchformer-Diffusion Model for Huangmei Opera Synthesis

Yufeng Qiu
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
2022111052@mail.hfut.edu.cn

Guofu Zhang*
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
zgf@hfut.edu.cn

Zhaopin Su
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
szp@hfut.edu.cn

Yang Zhou
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
2023170616@mail.hfut.edu.cn

Xiaoyi Bian
College of Humanities and Social Sciences
Anhui Agricultural University
Hefei, Anhui, China
bianxiaoyi@ahau.edu.cn

Abstract

Singing voice synthesis has been extensively used in metaverse, music creation and entertainment, and cultural preservation and inheritance. However, the synthesis of traditional operas, such as Huangmei opera, has been limited due to the lack of professionally annotated high-quality datasets and appropriate deep learning models. In this work, we develop a singing voice dataset and propose an acoustic model tailored for the unique singing style for Huangmei opera. More specifically, we first propose a data annotation method, effectively addressing the challenges posed by the numerous arias in this art form. Next, we construct our Huangmei opera singing voice dataset with the detailed musical score information, where each singing recording is captured at a high sampling rate of 44.1 kHz. Subsequently, we incorporate the Branchformer encoder and the pitch diffusion module to handle the complex and diverse melodies characteristic of Huangmei opera. Finally, extensive subjective and objective experiments demonstrate the effectiveness of the proposed dataset and model. Audio samples, the dataset, and the codes are available at https://walkinginthelight.github.io/HuangmeiSinger.github.io/opera/.

CCS Concepts

• **Computing methodologies** → Artificial intelligence; Natural language processing; Natural language generation.

Keywords

Singing voice synthesis, Corpus, Text-to-speech, Diffusion model, Chinese operas

*Corresponding author

CAICE 2025, Heifei, China

ACM ISBN 979-8-4007-1264-7/2025/01
https://doi.org/10.1145/3727648.3727752

ACM Reference Format:
Yufeng Qiu, Guofu Zhang, Zhaopin Su, Yang Zhou, and Xiaoyi Bian. 2025. HuangmeiSinger: A Dataset and A Branchformer-Diffusion Model for Huangmei Opera Synthesis. In *The 4th International Conference on Computer, Artificial Intelligence and Control Engineering (CAICE 2025), January 10–12, 2025, Heifei, China*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3727648.3727752

1 Introduction

Huangmei opera is one of China's five major opera types and one of the Chinese intangible cultural heritages, which has been in existence for more than 200 years. However, the rapid development of modern entertainment activities and the strong appeal of popular music to young people have led to the gradual decline of traditional Chinese drama. Therefore, it is necessary and timely to study the synthesis of Huangmei opera.

Deep learning based singing voice synthesis (SVS) provides a new solution to the research and creation of Huangmei opera. However, the existing SVS datasets and models mainly focus on Chinese pop music rather than Chinese traditional opera. For Chinese Pop Music, some researchers [1-4] have constructed small-scale corpora with annotations or large-scale datasets with music score information. In addition, some typical vocal synthesis models [5-8] have been proposed for the synthesis task of popular music, mainly achieving accurate and natural synthesized sound, as well as the glitch artifacts in Mel spectrograms. For Chinese Traditional Opera, Wu et al. [9] presented a duration informed attention network framework to synthesize expressive Peking opera singing from the music score using Peking opera dataset, Zhang et al. [10] constructed a Gezi opera dataset for Minnan dialect with 1,938 audio clips and developed a fine-grained tune based on GAN framework on the Gezi opera synthesis task.

Note that, the singing style of Huangmei opera is mainly based on the variation of the board style: floral, colorful, and tonal. Each style has multiple different forms, making Huangmei opera more complex and varied in tone compared to general popular music. Specifically, the duration of many phonemes in Huangmei opera varies greatly, with some phonemes having a short duration and

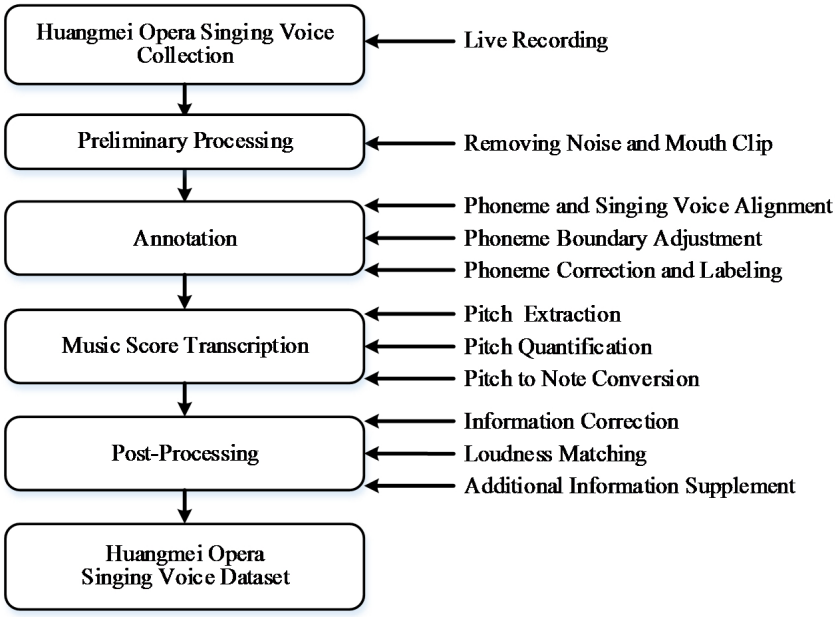

Figure 1: The flowchart of creating the dataset of Huangmei opera.

others having a long duration. In the long-term pronunciation of the same phoneme, the corresponding rhythm and tone will undergo different changes. That is, the same phoneme will correspond to multiple pitches in continuous time. This is significantly different from the popular music, which brings great difficulties to the synthesis of Huangmei opera singing style.

Against this background, we discuss the synthesis of Huangmei opera singing voice and make the following contributions: (1) We construct Huangmei opera singing voice dataset based on annotation strategy and a pitch quantification strategy, which contains high-quality recordings, manually annotated music score, and phoneme information. (2) We develop a Huangmei opera singing voice synthesis model based on Branchformer encoder and pitch diffusion module. The experimental results demonstrate that our model outperforms baseline models in terms of subjective and objective evaluation metrics.

2 Dataset Construction

The construction process of our Huangmei opera dataset is shown in Figure 1, which will be elaborated.

2.1 Huangmei Opera Singing Voice Collection

All the Huangmei opera singing voice recordings are recorded live. We invite six professional Huangmei opera actors/actresses (two male and four female (in order to meet the reality and cover more vocal ranges)) to participate in the recording. Under qualified recording conditions, the actor/actress of Huangmei opera is required to perform with deep affection. The lyrics for Huangmei opera are carefully selected. These lyrics include only the solo and singing parts and basically cover all the tonal ranges of Huangmei opera. During the recording process, the actor/actress wears headphones to ensure that the rhythm and lyrics of the recording can be synchronized, and only pure vocals are recorded. The recording equipment meets professional studio standards, uses a filter to prevent interference from air noise, and can effectively prevent explosive sounds. Each recording is sampled at 44.1 kHz, with a 24-bit depth, and saved in WAV format. Table 1 shows more detailed information for each actor/actress who has an integral and similar pitch range. This can help us to train a model with enough notes for each voice.

2.2 Annotation

Annotations are used to build more fine-grained manual music scores. For phoneme and singing voice alignment, we first convert the lyrics corresponding to the recording into syllables. Then, we split the obtained syllables into phonemes according to the rules of initial and final sounds. We use the Montreal forced aligner model [11] to preliminarily align the phoneme lyrics with the recording and save the results into a TextGrid file. To further improve the accuracy of the alignment results, we manually adjust the phoneme and syllable boundaries using the Praat annotation software [12] for finer labeling of the TextGrid files. Based on the annotation experience from the Opencpop dataset [2] for popular music, we also use a marker “silence pronunciation” (SP) to indicate that the Huangmei opera singing voice segment corresponding to SP is in a period of silence.

For artistic expression, Huangmei opera singing voice includes many legato and melisma, where the same phoneme corresponds to multiple different pitches due to its long duration. To facilitate subsequent processing, we add boundaries at special legato positions while adjusting the boundaries, performing phoneme segmentation, and marking with specific symbols. To reduce the marking time

Table 1: The information of the six singers.

Gender	SingerID	Pitch Range	Hours
Female	Singer1	54-78(D#3,185.00Hz - F#5, 739.99Hz)	1.00
	Singer2	53-76(F3,174.61Hz - E5, 659.25Hz)	0.36
	Singer3	53-77(F3,174.61Hz - F5, 698.46Hz)	0.24
	Singer4	49-73(C#3,138.59Hz – C#5, 554.37Hz)	0.62
Male	Singer5	49-70(C#3,138.59Hz – A#4, 466.16Hz)	0.24
	Singer6	43-69(G2,98.00Hz – A4, 440.00Hz)	1.26

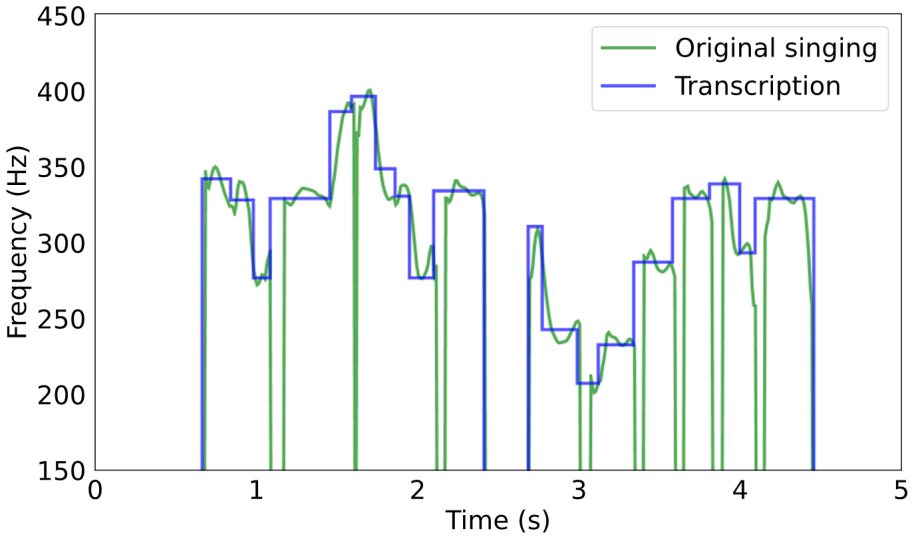

Figure 2: A example of the pitch quantization results on our dataset.

and cost, the manual marking does not involve segmenting and duplicating phonemes. Instead, repeated phonemes are automatically marked in the final dataset using a script. For example, in the lyrics "shu shang di niao er cheng shuang dui", the actual pronunciation of "shu" and "dui" lasts for several seconds. After processing, the phoneme representation is "SP sh u u sh ang d i n iao er er ch eng sh uang d ui ui ui".

2.3 Music Score Transcription

Generally, the Musical Instrument Digital Interface (MIDI) format is used to represent the musical score. Note that the actual singing of an actor/actress may not necessarily perfectly align with the musical score. In this work, we use the recorded Huangmei opera singing voice to construct the musical score.

We first apply the vocal pitch estimation model [13] to extract the pitch information from the singing recordings. By quantizing the pitch, we can attain the notes and MIDI numbers. To make the quantized pitch closer to the actual singing rhythm, we improve the quantization method by dynamically expanding the duration of each phoneme in the recording. Assuming the original duration of a phoneme is t, we extend its duration forward and backward by half a beat, denoted as tf and tb, respectively. Then, the total quantized duration is $T = t + tf + tb$. During this period, we calculate the average pitch, excluding any abrupt pitch changes in the process, and ultimately obtain the pitch value for that phoneme using formula (1), where $F0$ represents the pitch of the extracted recording within T.

$$Pt = \frac{\sum_{i=1}^{N} F0_i}{T} \tag{1}$$

Based on the above operations, each phoneme can be accurately mapped to its corresponding pitch value. The processed result is shown in the Figure 2. Since there is a mapping relationship between pitch, notes, and MIDI numbers, the pitch can be directly converted into notes and MIDI numbers. The annotated files include information such as the lyrics of the song, syllables, phonemes, notes, durations, and slur, as shown in our demo link.

2.4 Post-Processing

To facilitate the subsequent use of the dataset and for training tasks, we first trim the long recording into small segments ranging from 5 to 15 seconds and correspondingly segmented the annotation files (e.g., TextGrid files). Next, we use scripts to clean the data and ensure that each singing voice segment in the dataset is within an appropriate length range by removing segments that were too

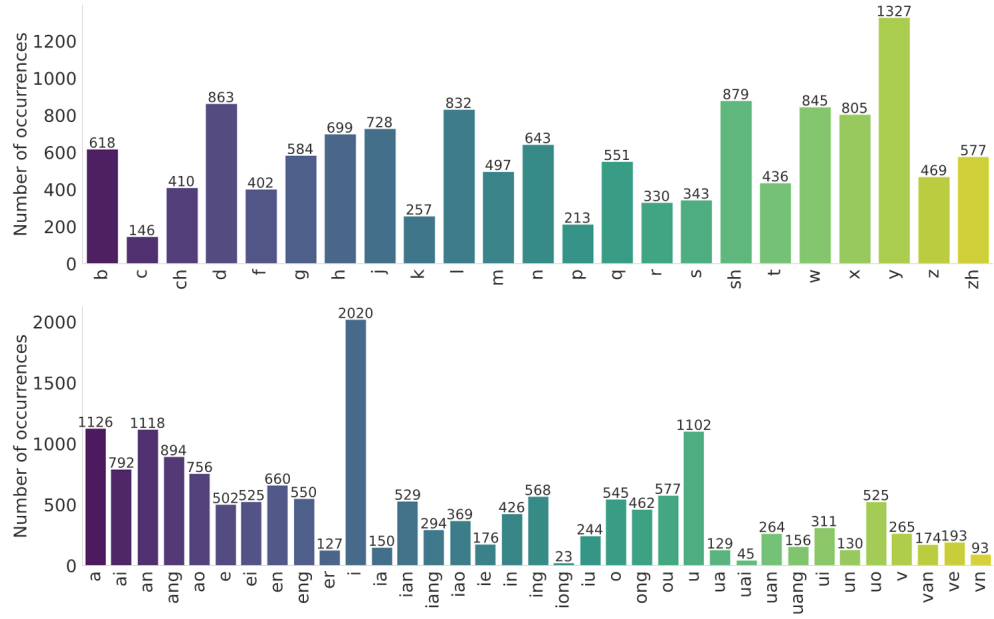

Figure 3: The distribution of phonemes, divided into Shengmu (top) and Yunmu (bottom).

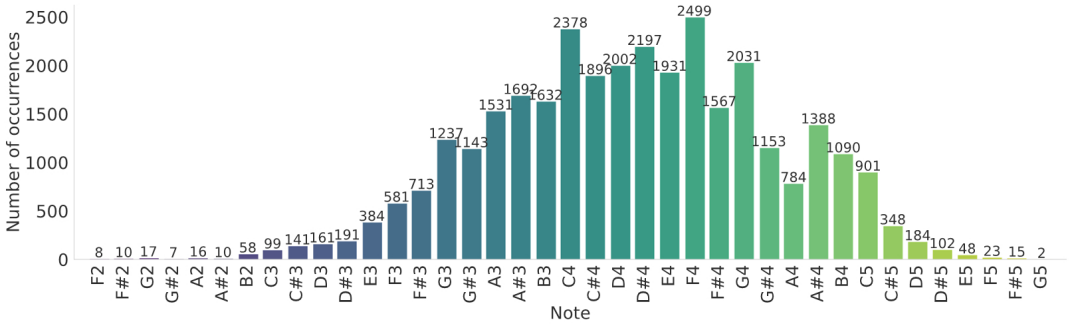

Figure 4: The statistical distribution of pitch in our dataset. Pitch is presented as note.

long or too short. Then, we check the accuracy of the lyrics and phonemes and eliminate any potential errors or duplicate lyrics. In addition, we use "Adobe Audition" to match the loudness of the singing voice. Finally, the segmented and cleaned the singing voice segments are saved as individual singing voice files, and their corresponding annotation information is saved in text files, just constructing our Huangmei opera dataset.

2.5 Dataset Analysis

Our Huangmei opera singing voice dataset is divided into 1,830 recordings, each of which is between 5s and 15s, with a total duration of 3.7 hours. The training set contains 1,780 recordings and the testing set includes 50 recordings. The statistical distribution of phonemes and pitches in our Huangmei opera dataset is illustrated in Figure 3 and Figure 4. It can be observed from the figure that our dataset covers all the phonemes, and the pitch mainly ranges from F2 to G5 (40 to 79) that the recordings in our dataset basically cover both high and low frequencies.

3 Model for Huangmei Opera Synthesis

The architecture of the proposed model is shown in Figure 5, where the Branchformer model [14] is used as the encoder to better process rhythm and phoneme information in Huangmei opera recordings, multiple Transformer blocks based FastSpeech2 model is for decoding and generating mel-spectrogram, and pre-trained HiFi-GAN model from Liu et al. [17] (https://github.com/MoonInTheRiver/DiffSinger/blob/master/docs/README-SVS-popcs.md)is adopted for the vocoder to ensure the naturalness and realism of the synthesized singing voice.

To capture pitch variations and tonal details in Huangmei opera, we adopt a conditional diffusion model based on WaveNet [15]. In the forward training process, we input the acoustic information

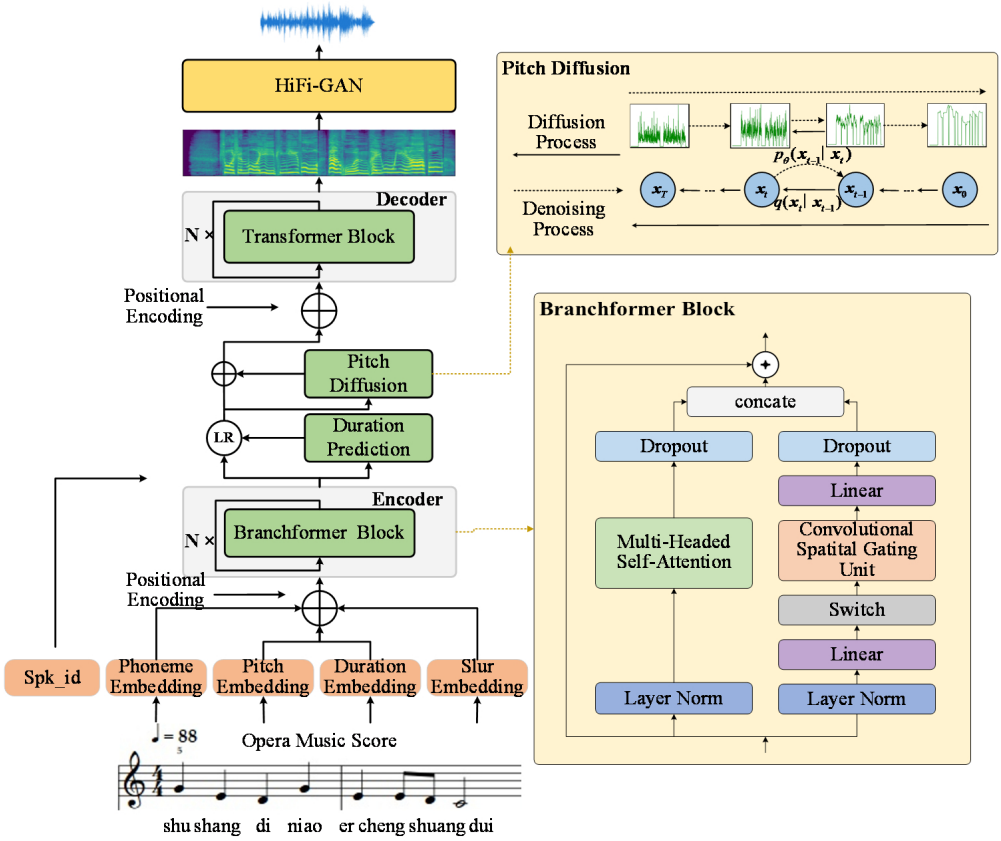

Figure 5: The overall architecture of our branchformer-diffusion model.

generated by the encoder as the prediction condition c into the network. Then, we train this denoising network θ using the formula (2) [16], where $t \in [1, T]$ is the index for diffusion step; T denotes the maximal iteration number; $x_0$represents a pitch; $\bar{\alpha}_t$ represents the product of the empirical constant α_t used to adjust the variance in the previous t steps; ϵ denotes the added original noise, following a Gaussian distribution; and ϵ_θ is a WaveNet based approximator to predict the previously added ϵ.

When the diffusion is finished, we can achieve a noise x_T that also follows a Gaussian distribution. In the process of reverse inference, for the given condition c, the noise x_T is iteratively sampled according formula (3), Where $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}$; and z represents the added original noise and follows a Gaussian distribution. After sampling in T steps, the pitch x_0 can be predicted.

$$\nabla_\theta \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)^2 \tag{2}$$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, c, t\right)\right] + \sigma_t z \tag{3}$$

It should be noted that, in the previous study [16], the condition for the pitch diffusion module is often frozen or a basic $F0$ derived from the data preprocessing is used for training. Differently, in this work, we directly use the output of the encoder as the condition for the pitch conditional diffusion module, and take the logarithm of the original $F0$ as the training target to simplify the training process.

To make the duration prediction more accurate, we not only train at the phoneme level, but also constrain at the sentence level and the word level. We use the L2 loss function for constraint. The total duration prediction loss is defined as formula (4), where $\lambda_i \in [0, 1]$ is the weight coefficient, satisfying $\lambda_1+\lambda_2+\lambda_3 = 1$. In the pitch prediction, the loss of the diffusion module L_{diff} is defined as formula (5). In the mel-spectrogram prediction, The total loss function is defined the sum of the L1 loss, the structural similarity index (SSIM) loss and the mel loss.

$$L_{duration} = \lambda_1 L2_{phoneme} + \lambda_2 L2_{word} + \lambda_3 L2_{sentence} \tag{4}$$

$$L_{diff}(\theta) = \mathbb{E}_{t,x_0,\epsilon} \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)^2 \tag{5}$$

$$L_{total} = L_{duration} + L_{diff} + L_{mel} \tag{6}$$

4 Experiments

4.1 Experimental Setup

We adopt FastSpeech2 [6], VISinger2 [7] and DiffSinger [17] for our comparative experiments using their official implementation. In our model, the hidden layer units and mel bins are set to 256 and 80, respectively. In the Branchformer module, the number of blocks is 12; each block is configured with 4 attention heads. The decoder

Table 2: Comparative experimental results obtained by the four models in terms of the used evaluation metrics.

Model	MCD↓	F0RMSE↓	DurAcc↑	F0Pcc↑	MOS↑
GT	–	–	–	–	4.36±0.05
FastSpeech2	7.41	24.76	0.838	0.9778	3.54±0.04
DiffSinger	**7.15**	24.51	0.832	0.9784	3.67±0.05
VISinger2	7.69	25.57	0.890	0.9133	3.86±0.04
Ours	7.28	**21.23**	**0.895**	**0.9832**	**3.93±0.05**

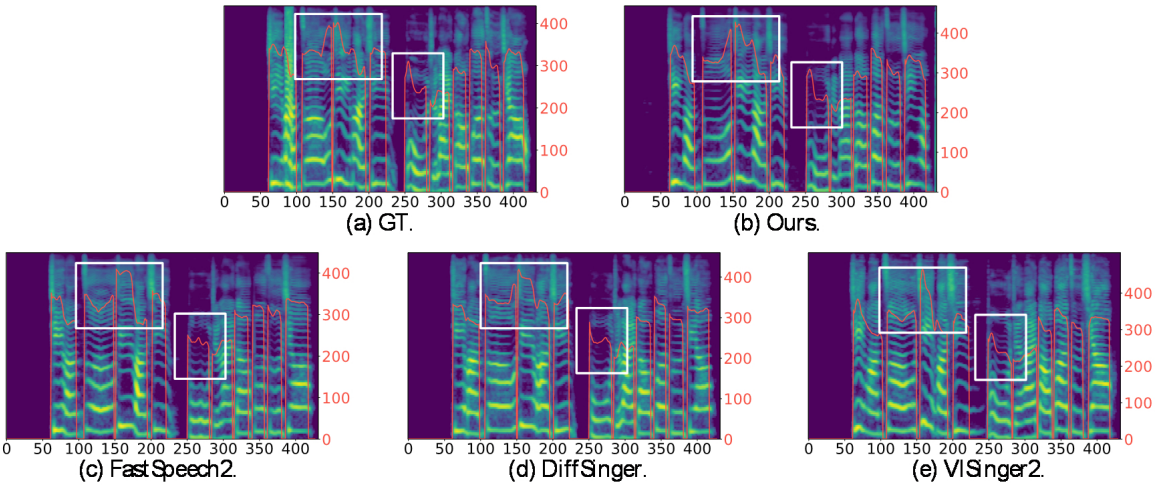

Figure 6: Visualization of the pitch contour and mel-spectrogram of GT and different models (one case study).

contains 4 layers and 2 attention heads. For the diffusion model, the number of time steps is 100; the linear α_t ranges from 0.94 to 0.9999. For the training, the initial learning rate is set to 0.001; the AdamW optimizer is used with a decay rate of 0.1. More details can be found in our code. The entire model is trained for 190,000 iterations on an Nvidia GeForce RTX 3090 GPU.

4.2 Results

To assess the quality of the synthesized Huangmei opera singing voice, we adopt the mean opinion score (MOS) [9] metric where higher MOS metrics indicate more authentic, natural, high-quality, and distortion free audio. And four objective metrics, such as mel-cepstral distortion (MCD) [4], $F0$ root mean square error (F0RMSE) [4], duration accuracy (DurAcc) [18], and Pearson correlation coefficient for $F0$ (F0Pcc) [5].

Table 2 shows the experimental results between the ground-truth (GT) and different models. Please note that GT is the recorded audio and represents the best performance. It can be observed that our model outperforms the baseline model on four metrics. Particularly, our model performs significantly better than other models on F0RMSE and DurAcc. Next, the pitch prediction ability has been significantly improved due to the fact that F0RMSE has decreased by 3. In addition, the pitch correlation has also been improved. The above results indicate that the Huangmei opera singing voice synthesized by our model is closer to the original voice.

In terms of MOS, our model is slightly better than VISinger2. This indicates that the Huangmei opera singing voice synthesized by our model is more realistic, natural, and expressive. For further visual illustration, Figure 6 depicts the pitch contour and mel-spectrogram generated by different models. It can be found that our model can generate more natural pitch contours especially in the general part (white box region). This also indicates that our method can achieve better pitch prediction.

4.3 Ablation Experiments

To validate the impact of the modified modules on the synthesis quality, we conducted two ablation experiments. In the first experiment, we replaced the Branchformer module with the original transformer module. In the second ablation experiment, we removed the pitch diffusion model and replaced it with a transformer-based $F0$ predictor. The experiments are evaluated in terms of comparative MOS (CMOS). The results are shown in Table 3.

As can be seen, on one hand, the absence of the Branchformer encoder significantly degraded the model's ability to predict durations, with a reduction of about 0.06 in duration accuracy. The pitch prediction ability also slightly decreased with a drop of 0.28 in MOS. On the other hand, the absence of the pitch diffusion model led to a noticeable decline in pitch prediction performance. Particularly, the ability to predict duration was also weakened, resulting in a 0.22 decrease in MOS. The above results demonstrate the effectiveness

Table 3: Ablation experimental results obtained by our model.

Model	F0RMSE↓	DurAcc↑	CMOS
Full model	21.23	0.895	0.00
w/o branchformer	22.09	0.834	-0.28
w/o pitch diffusion	23.43	0.891	-0.22

of the Branchformer module and the pitch diffusion predictor in our Huangmei opera singing voice synthesis.

5 Conclusion

To better inherit and protect our traditional opera, in this work, we created a Huangmei opera singing voice dataset with accurate alignment of voice and text through meticulous labeling. Furthermore, based on this dataset, we developed a Huangmei opera singing voice synthesis model tailored to the characteristics of Huangmei opera. The experimental results demonstrate the effectiveness of the proposed dataset and model, stemming from the fact that the synthesized Huangmei opera singing voice exhibits a certain level of expressiveness and naturalness.

To our best knowledge, this work can be seen as the first attempt to investigate the dataset and model of Huangmei opera synthesis, which may be helpful for researchers or practitioners of the synthesis of Chinese traditional opera. As the coverage of the synthesized pitch is still limited due to the relative scarcity of high-quality recordings in the dataset, our future work focuses on how to record more high-quality Huangmei opera singing recordings to expand our dataset.

Acknowledgments

This work was supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant 24YJA870011.

References

[1] Rongjie Huang, *et al.* Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In Proceedings of the 29th ACM International Conference on Multimedia. 2021. p. 3945-3954.
[2] Wang, Yu, *et al.* Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. arXiv:2201.07429, 2022.
[3] Zhang, Lichao, *et al.* M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. Advances in Neural Information Processing Systems, 2022, 35: 6914-6926.
[4] Chu, Chan-Chuan, *et al.* MPop600: A Mandarin popular song database with aligned audio, lyrics, and musical scores for singing voice synthesis. In proceedings of APSIPA ASC. IEEE, 2020. p. 1647-1652.
[5] Lu, Peiling, *et al.* Xiaoicesing: A high-quality and integrated singing voice synthesis system. arXiv:2006.06261, 2020.
[6] Ren, Yi, *et al.* Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv:2006.04558, 2020.
[7] Zhang, Yongmao, *et al.* Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. arXiv:2211.02903, 2022.
[8] Zhang, Zewang, *et al.* Wesinger: Data-augmented singing voice synthesis with auxiliary losses. arXiv:2203.10750, 2022.
[9] Wu, Yusong, *et al.* Peking opera synthesis via duration informed attention network. arXiv:2008.03029, 2020.
[10] Zhang, Meizhen, *et al.* FT-GAN: Fine-Grained Tune Modeling for Chinese Opera Synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence. 2024. p. 19697-19705.
[11] McAuliffe, Michael, *et al.* Montreal Forced Aligner: An accurate and trainable aligner using Kaldi. In: Poster presented at the Linguistic Society of America 91st Annual Meeting, Austin. 2017.
[12] Boersma, Paul. Praat, a system for doing phonetics by computer. Glot. Int., 2001, 5.9: 341-345.
[13] Wei Haojie, *et al.* RMVPE: A robust model for vocal pitch estimation in polyphonic music. arXiv:2306.15412, 2023.
[14] Peng, Yifan, *et al.* Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In proceedings of PMLR, 2022. p. 17627-17643.
[15] Van Den Oord, Aaron, *et al.* Wavenet: A generative model for raw audio. arXiv: 1609.03499, 2016, 12.
[16] Li, Xiang, *et al.* Diverse and expressive speech prosody prediction with denoising diffusion probabilistic model. arXiv preprint arXiv:2305.16749, 2023.
[17] Liu, Jinglin, *et al.* Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In: Proceedings of the AAAI conference on artificial intelligence. 2022. p. 11020-11028.
[18] Xue, Heyang, *et al.* Learn2sing: Target speaker singing voice synthesis by learning from a singing teacher. In proceedings of 2021 IEEE SLT. IEEE, 2021. p. 522-529.