

Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis

Yu Wang^{1,2*}, Xinsheng Wang^{3,4*}, Pengcheng Zhu^{3,4†}, Jie Wu^{1†}, Hanzhao Li³, Heyang Xue^{3,4},
Yongmao Zhang³, Lei Xie³, Mengxiao Bi⁴

¹ College of Design and Innovation, Tongji University, China

² School of Pop Music & Dance, Shanghai Institute of Visual Arts, China

³ Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China

⁴ Fuxi AI Lab, NetEase Inc., Hangzhou, China

asterwangyu@126.com, w.xinshawn@gmail.com, zhupengcheng@corp.netease.com,
lihanzhao.mail@gmail.com, xueheyang@corp.netease.com, zym@mail.nwpu.edu.cn,
wujie@tongji.edu.cn, lxie@nwpu.edu.cn, bimengxiao@corp.netease.com

Abstract

This paper introduces Opencpop, a publicly available high-quality Mandarin singing corpus designed for singing voice synthesis (SVS). The corpus consists of 100 popular Mandarin songs performed by a female professional singer. Audio files are recorded with studio quality at a sampling rate of 44,100 Hz and the corresponding lyrics and musical scores are provided. All singing recordings have been phonetically annotated with phoneme boundaries and syllable (note) boundaries. To demonstrate the reliability of the released data and to provide a baseline for future research, we built baseline deep neural network-based SVS models and evaluated them with both objective metrics and subjective mean opinion score (MOS) measure. Experimental results show that the best SVS model trained on our database achieves 3.70 MOS, indicating the reliability of the provided corpus. Opencpop is released to the open-source community WeNet¹, and the corpus, as well as synthesized demos, can be found on the project homepage².

Index Terms: Singing voice synthesis, corpus, text-to-speech, open source, benchmark

1. Introduction

Singing voice synthesis (SVS), which aims to synthesize singing voices from text and musical score information (e.g., note and tempo), has great potential for many creative applications, e.g., virtual avatars and artistic creation. To promote the development of Mandarin SVS, a high-quality Mandarin SAS corpus, referred to as Opencpop, with fine annotated textual information and musical scores, is released in this paper.

The development of neural end-to-end text-to-speech (TTS) models [1, 2, 3, 4, 5, 6] has greatly promoted speech synthesis. Generally, with a well-trained neural acoustic model [2, 5, 6, 7] and a neural vocoder [8, 9, 10, 11], or alternatively using fully end-to-end models [12, 13, 14] which directly construct wave signals from text input, it is able to synthesize high-quality neutral speech. Recently, much attention has been attracted to synthesizing expressive speech, such as stylized speech [15,

16], emotional speech [17, 18, 19, 20, 21, 22], and also singing voice [23, 24].

Apart from technological innovating, a proper database is crucial for evaluation and comparison. The end-to-end TTS technology allows to train a model only with audio files and paired textual transcriptions, such as corpora LJ-Speech [25], VCTK [26], Aishell-3 [27], and DiDiSpeech [28]. However, the case in the SVS task is different, in which the musical score is generally needed. While several open-source SVS databases designed for English [29, 30] or other languages [31] have been released and also lots of research has been conducted towards Mandarin singing voice synthesis [24, 32, 33], no accessible public Mandarin SVS corpus with high quality exists, thereby limiting the development of Mandarin SVS. Note that although Chinese singing voice databases OpenSinger [33] and PopCS [34] were released recently, accurate musical score information was missing in both corpora, with which an automatic fundamental frequency detection algorithm is necessary to calculate the rough pitch information. These corpora, thus, are hard to satisfy an SVS system with high quality.

A major cause of this data shortage is the challenge of annotating a high-quality SVS corpus. Different from the corpora designed for speech synthesis, in which only the recordings and corresponding transcriptions are necessary, extra musical score information, such as note and note boundaries, is also necessary for training an SVS model. While the song is generally sung based on the existed musical score, the singing voice is hard to be exactly aligned with the musical score even for a professional singer. Thereby, manual post-processing, such as sound shaping or musical score relabeling, is necessary, which depends on heavy labor from professionals, making it much harder to build such a corpus.

In this paper, we introduce the accessible public Opencpop, which is a high-quality SVS corpus with the phonetically manual-annotated musical score. In this corpus, 100 Mandarin Pop songs are recorded by a professional female singer. All audio is recorded in a professional recording studio with a sampling rate of 44,100 Hz. The total duration of the recording is around 5.2 hours, and the distributions of phoneme, syllable, and pitch will be introduced in the paper, respectively. Besides, based on the new corpus, baseline SVS models are built to test the reliability of the collected database, and also provide the baseline performance for future research.

*These authors contributed equally to this work.

†Corresponding authors.

¹<https://github.com/wenet-e2e/>

²<https://wenet.org.cn/opencpop/>

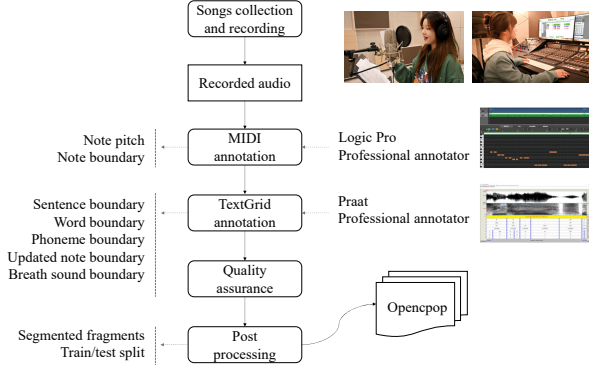


Figure 1: Flowchart for creating the Opencpop database. For clear presentation, here, quality assurance is only illustrated after the TextGrid annotation. In practice, the human double-check exists in every process.

The rest of this paper is organized as follows: Section 2 details the corpus creation pipeline from collecting songs to the final audio files. The corpus’s overall statistics, including both audio and textual transcripts, are also presented in this section. We introduce the details of training different models based on Opencpop as well as the evaluation and the results in Section 3. Finally, we conclude the paper in Section 4.

2. The creation of Opencpop

In this section, we will introduce Opencpop, a publicly available Mandarin singing corpus designed for the SVS task, which has fine manually annotated textual information and musical scores. The flowchart of Opencpop creation is shown in Fig. 1, and the details of each step for creating this database will be described in this section.

2.1. Songs and the singer

All songs were selected according to several music charts for China Pop songs. The originally collected songbook consisted of 300 songs, based on which the final 100 songs were selected with the following processes: 1) any song would be removed from the songbook if its lyric contains non-Chinese characters; 2) the remained songs’ beats per minute (BPM) were counted, and songs with the BPM that occurs in relatively lower frequency were chosen to the final song list preferentially; 3) similar to the previous BPM-based rule, remained songs with infrequent phonemes were selected to the final list; 4) the final selected song list was completed by choosing songs that contain as many phonemes as possible. After the above steps, 100 Mandarin Pop songs with sufficient phoneme and BPM coverage were chosen for the recording.

The singer who would perform these songs is a professional young female singer. Her vocal range can perfectly perform all the selected songs.

2.2. Recording

The recording was conducted in a professional studio with sufficiently little reverb, and the singing voice was recorded as an individual track, which means each recorded audio contains pure vocal only, without any background music. During the recording, the singer was equipped with a headphone, which provided the accompanying audio, to ensure the steadiness and accuracy of

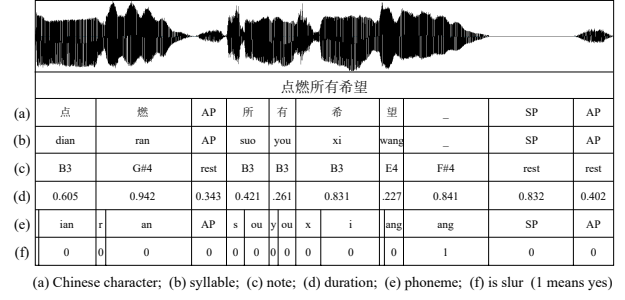


Figure 2: Illustration of a labeled sample. The phoneme “d” of syllable “dian” and the phoneme “w” of syllable “wang” are omitted due to the limited space. The “rest” in the note track corresponds to the silence (SP) or aspirate (AP).

pitch. All the data were recorded in 44,100 Hz sampling rate with 24 bits per sample in wav format.

2.3. Labeling

In the Opencpop database, manually labeled information includes note pitch, note boundary, phoneme boundary, syllable boundary, word boundary, utterance boundary, and the indicator to indicate whether a note is a slur or not. An illustration of a labeled sample is shown in Fig. 2. Details to obtain those annotations will be described in this part.

2.3.1. MIDI annotation

The musical score, as a written form of a musical composition, normally includes the note pitch, note duration, the tempo, the key signature, etc. The music score that the singer accords to during the recording can be easily transcribed to MIDI format. However, even though our singer is professional in singing, it is non-trivial to ensure the sing voice aligned with the musical score perfectly. Therefore, this music score-based method is hard to create adequate results. Instead, a semi-automatic method was adopted to create the MIDI score based on the recorded audio rather than the original music score.

To be specific, with the input of audio, Logic Pro³, which is a digital audio workstation and MIDI sequencer software, was utilized to automatically create the preliminary musical score. Then this preliminary musical score was manually adjusted by professional annotators to obtain the more accurate but not final musical score. The musical score obtained in this stage would be further tuned during the TextGrid annotation process, which will be introduced in the next subsection.

2.3.2. TextGrid annotation

The TextGrid annotation is to label all information illustrated in Fig. 2, in which process the note-related annotations, i.e., note pitch and note duration, were taken from the MIDI annotation results, and the note duration (boundary) was fine-tuned to align with the newly labeled syllable boundary. To efficiently label boundaries of pronunciation in various levels, a preliminary alignment between phonemes and the corresponding audio was obtained with Montreal Forced Aligner⁴ for automatic songs’ phoneme annotation. The achieved rough phoneme boundaries are very friendly for annotators to annotate the final accurate

³<https://www.apple.com/logic-pro/>

⁴<https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner>



Figure 3: The violin plots of pitch numbers in different subsets. Pitch is presented as MIDI note number, where $A4=69=440$ Hz.

phoneme boundaries in Praat [35]. Then, the syllable boundaries can be easily obtained based on the phoneme boundary. Here, we argue that the boundary of each syllable is aligned with the boundary of a pitch, which means the pitch boundary achieved in the MIDI annotation process would be adjusted to align with the syllable boundary.

A case to show different levels' boundaries is presented in Fig. 2. As can be seen, different from TTS-oriented textual phoneme sequences that can be directly converted from sentences, in Opencpop database, some phonemes could be repeated due to the slur in the song, for instance, the phoneme "ang" of syllable "wang" is repeated one time. The indicator of the slur was also annotated in Opencpop database. Besides, the aspirate was also labeled.

2.4. Post processing

2.4.1. Audio segmentation

After the annotation, we segmented the recording into smaller fragments for the training of the SVS system. All song audio was first segmented into sentence-level according to the transcription and the annotated boundary information. Then, to avoid the existence of too-long utterances in the final database, any utterances longer than 8 seconds would be further segmented into shorter utterances. During this segmentation process, the end boundary of a *rest* within the utterance (*SP* in Fig. 2) would be the priority segmentation point. If no *SP* exists within an utterance that is longer than 8 seconds, the prosodic boundary achieved by prosody analysis would be set as the segmentation point. After dropping silence utterances, we obtained 3,756 utterances finally with a total duration of around 5.2 hours.

2.4.2. Training set split

For the training and evaluation of an SVS system, 5 songs out of the total 100 songs were handpicked as the test set, which contains 206 utterances, and the remaining 95 songs with 3,550 utterances were set as the training set. The violin plots of pitch numbers in different test songs and the training set are compared in Fig. 3. As can be seen, the selected 5 test songs cover several typical cases, i.e., with relatively high (Test-2100), low (Test-2086 and Test-2092), and medium pitch values (Test-2100), and thus can support a comprehensive evaluation for an SVS system. Details of the Opencpop database's statistics will be introduced in Section 2.5.

2.5. Statistics of Opencpop

As introduced in Section 2.1, to make the Opencpop cover a wide range of BPM, a song with the infrequent BPM would take the priority to be added to the song list of Opencpop. The final BPM distribution is shown in Fig. 4. As can be seen, Opencpop contains songs with various BPM that ranges from 40 to 130.

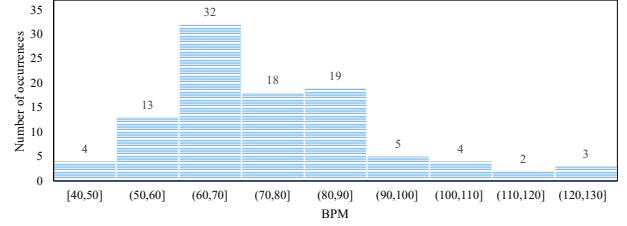


Figure 4: The statistical distribution of BPM.

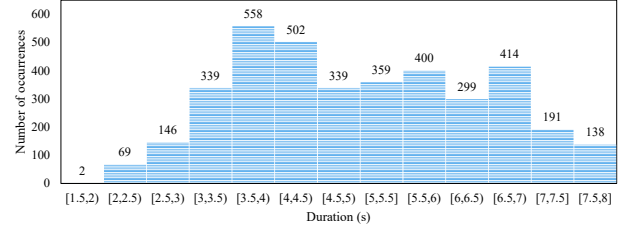


Figure 5: The statistical distribution of utterance duration. The shortest and longest duration is 1.8s and 8.0s, respectively.

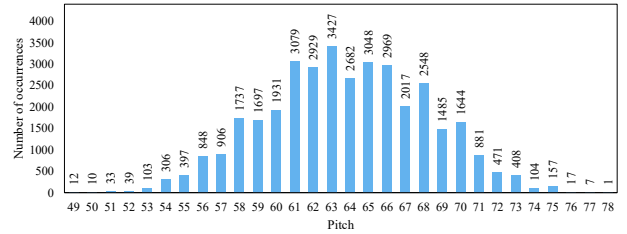


Figure 6: The statistical distribution of pitch. Pitch is presented as MIDI note number, where $A4=69=440$ Hz.

Most songs' BPM falls within the range of 58 to 94, which is a popular BPM range of Mandarin Pop music. This wide coverage of BPM allows a system trained with Opencpop to deal with various BPM conditions.

The duration distribution of segmented utterances is shown in Fig. 5. The duration of most utterances ranges from 3 seconds to 7 seconds. The note pitch distribution and phoneme distribution are presented in Fig. 6 and Fig. 7, respectively. The notes are mainly distributed between MIDI note number 56 (G#3, 196 Hz) to 70 (B4, 494 Hz). As for the phoneme distribution, Opencpop covers all Mandarin phonemes, with the lowest frequency of phoneme "iong" which occurs 25 times in total, which can qualify an SVS system to handle all phonemes.

3. Baseline systems and experiments

To verify the quality of Opencpop in the SVS task, and also to give a performance baseline in this task, several experiments towards the SVS task on the Opencpop database are conducted in this section.

3.1. Methods

A typical two-stage singing voice synthesis framework, which consists of an acoustic model and a vocoder is adopted in the experiments. In practice, FastSpeech2 [5] and HiFi-GAN [11], which are two popular models for spectrogram synthesis and waveform reconstruction respectively, are utilized in this work.

FastSpeech2, which is a fully-feedforward architecture based on Transformer blocks [3], is taken as the baseline in this paper. Different from the FastSpeech2-based TTS task, here, the prediction of energy and pitch is not considered. Instead, the input of the encoder is the concatenation of phoneme embedding

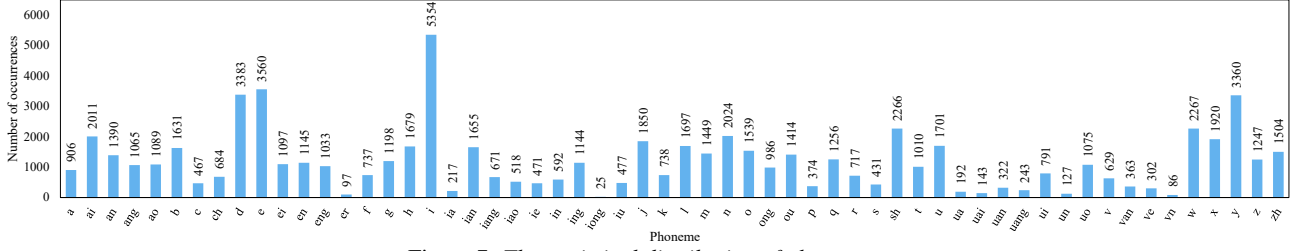


Figure 7: The statistical distribution of phonemes.

and note embedding. As for the duration prediction, in addition to the phoneme level loss, an extra note level duration loss is added to provide the note-level duration constraint during the training process. The input of duration predictors is not only the phoneme embedding as in Fastspeech2 but also concatenates with other information, including note duration, slur indicator, phoneme position, and the phoneme number within a note.

While the Transformer block is good at modeling long-range global context, it has limited capability in capturing fine-grained local context. To face this limitation, Conformer thus came into being and has shown promising performance in the TTS task [36]. In this work, we replace the Transformer blocks in Fastspeech2 with Conformer blocks, leading to a Conformer-based SVS model so-called *CpopSing*.

The Mel-prediction in Fastspeech2 is trained with L1 loss. However, it is well known that acoustic models trained with L1 or L2 loss suffer from over-smoothness of synthesized acoustic features, which could lead to singing voices with poor quality. Therefore, in addition to L1 loss, an adversarial training method is used during the training of *CpopSing*. This adversarial training method is similar to the sub-frequency adversarial loss in HifiSinger [23] but with an extra multi-length adversarial loss on the spectrogram.

In addition to *CpopSing*, with only L1 loss as the Mel reconstruction loss, a Transformed-based model and a Conformed-based model are also compared in the experiments. For convenience, these two models are referred to as *TFSing* and *CFSing* respectively, hereafter.

3.2. Implementation details

As a baseline for the Opencpop-based SVS task, here, we did not take the raw sampling rate 44,100 Hz as the target sampling rate. Instead, following the recent SVS research, the audio was down-sampled to 22,050 Hz, and then was represented by 80-band Mel-scale spectrograms with the frame-shift of 12.5 ms. During the waveform reconstruction, the vocoder was trained with the real audio-spectrogram pairs, and then was fine-tuned with synthesized spectrograms paired with real audio.

3.3. Results

Following [37], evaluation metrics, i.e., F0 Root Mean Square Error (F0-RMSE), F0 Pearson Correlation Coefficient (F0-PCC), and duration accuracy (duracc) are used to evaluate the synthesized results objectively. To match the length difference between the ground-truth singing voice and the generated voice, the calculation of F0-RMSE and F0-PCC is conducted on generated singing voices that were created based on the ground-truth phoneme duration. Besides, to verify the quality of the synthesized singing voices subjectively, a Mean Opinion Score (MOS) test is also performed. In practice, all synthesized samples of the test set are used in the objective evaluation, while 30 samples are randomly chosen from the test set in the MOS test. The score of MOS test ranges from 1 to 5, in which 1 means very bad and 5

Table 1: Objective and subjective performance of different models on Opencpop. Lower F0-RMSE value means better performance. As for other metrics, higher value means better. “gt dur” means the ground-truth phoneme duration is used in the synthesizing, while the synthesizing in “predict dur” is based on the phoneme duration predictor.

Method	F0-RMSE	F0-PCC	duracc	MOS	
				predict dur	gt dur
TFSing	29.99	0.893	0.885	3.45±0.05	3.51±0.05
CFSing	26.72	0.903	0.884	3.67±0.04	3.73±0.05
CpopSing	27.58	0.904	0.879	3.70±0.05	3.76±0.05
Ground Truth	—	—	—	4.51±0.03	

means excellent. Each audio is rated by 20 listeners.

The results are shown in Table 1. As shown, training on Opencpop, *CpopSing* achieves 3.70 MOS, falling in the range that near to *good*, indicating the reliability of the proposed database Opencpop. On all evaluation metrics, except for duracc, *CpopSing* and *CFSing* achieve better performance than *TFSing*. In terms of objective metrics, *CpopSing* and *CFSing* show similar performance, while *CpopSing* outperforms *CFSing* on the subjective evaluation. These results indicate the effectiveness of Conformer blocks and Mel-based adversarial training method on the singing voice synthesis. From the generated results, baseline models showed limited capability in handling the pitch that falls in a long tail distribution in terms of pitch frequency in the training set⁵. How to deal with this issue raised by the long tail distribution and even to perform a song with note pitch that is beyond the training set pitch distribution should be considered in future work.

4. Conclusions

In this paper, we introduced the Opencpop corpus designed primarily for SVS systems. All recordings are phonetically labeled with phoneme boundaries and note boundaries manually. To our knowledge, this is the first open accessible high-quality Mandarin singing corpus with manual annotation. Opencpop can open up a number of possibilities for further research in the areas of SVS.

5. Acknowledgements

The authors thank contributors of the WeNet Open Source Community for maintaining this awesome community, especially to the primary maintainer Binbin Zhang for helping the release of Opencpop on the WeNet community. The authors also thank Mingming Fu, Tianyao Bai, Chen Li, and their colleagues in the audio department of NetEase Leihuo for their guidance and support in music basics.

⁵Please refer to the demo page for audio examples with spectrum: <https://wenet.org.cn/opencpop/resources/testset/>

6. References

- [1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang *et al.*, “Tacotron: Towards end-to-end speech synthesis,” in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.
- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” *arXiv preprint arXiv:1712.05884*, 2017.
- [3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.
- [4] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2wav: End-to-end speech synthesis,” in *Proc. ICLR workshop*, 2017.
- [5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.
- [6] 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.
- [7] C. Miao, L. Shuang, Z. Liu, C. Minchuan, J. Ma, S. Wang, and J. Xiao, “Efficienttts: An efficient and high-quality text-to-speech architecture,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 7700–7709.
- [8] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [9] S. Kim, S.-g. Lee, J. Song, J. Kim, and S. Yoon, “Flowwavenet: A generative flow for raw audio,” *arXiv preprint arXiv:1811.02155*, 2018.
- [10] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.
- [11] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *arXiv preprint arXiv:2010.05646*, 2020.
- [12] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, “End-to-end adversarial text-to-speech,” in *International Conference on Learning Representations*, 2020.
- [13] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, “Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5679–5683.
- [14] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” *arXiv preprint arXiv:2106.06103*, 2021.
- [15] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.
- [16] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” *arXiv preprint arXiv:1803.09017*, 2018.
- [17] Y. Lee, A. Rabiee, and S.-Y. Lee, “Emotional end-to-end neural speech synthesizer,” *arXiv preprint arXiv:1711.05447*, 2017.
- [18] A. Rabiee, T.-H. Kim, and S.-Y. Lee, “Adjusting pleasure-arousal-dominance for continuous emotional text-to-speech synthesizer,” *arXiv preprint arXiv:1906.05507*, 2019.
- [19] X. Zhu, S. Yang, G. Yang, and L. Xie, “Controlling emotion strength with relative attribute for end-to-end speech synthesis,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 192–199.
- [20] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, “Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition,” *arXiv preprint arXiv:2010.13350*, 2020.
- [21] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, “Emotional speech synthesis with rich and granularized control,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.
- [22] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, “Controllable cross-speaker emotion transfer for end-to-end speech synthesis,” *arXiv preprint arXiv:2109.06733*, 2021.
- [23] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, “Hifisinger: Towards high-fidelity neural singing voice synthesis,” *arXiv preprint arXiv:2009.01776*, 2020.
- [24] Z. Liu, C. Miao, Q. Zhu, M. Chen, J. Ma, S. Wang, and J. Xiao, “Efficientsing: A chinese singing voice synthesis system using duration-free acoustic model and hifi-gan vocoder,” *Proc. Interspeech 2021*, pp. 1609–1613, 2021.
- [25] K. Ito and L. Johnson, “The lj speech dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [26] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, “Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2016.
- [27] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “Aishell-3: A multi-speaker mandarin tts corpus and the baselines,” *arXiv preprint arXiv:2010.11567*, 2020.
- [28] T. Guo, C. Wen, D. Jiang, N. Luo, R. Zhang, S. Zhao, W. Li, C. Gong, W. Zou, K. Han *et al.*, “Disispeech: A large scale mandarin speech corpus,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6968–6972.
- [29] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, “The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–9.
- [30] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, “Nhss: A speech and singing parallel database,” *Speech Communication*, vol. 133, pp. 9–22, 2021.
- [31] H. Tamaru, S. Takamichi, N. Nanji, and H. Saruwatari, “Jvs-music: Japanese multispeaker singing-voice corpus,” *arXiv preprint arXiv:2001.07044*, 2020.
- [32] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, “Xiaoicesing: A high-quality and integrated singing voice synthesis system,” *arXiv preprint arXiv:2006.06261*, 2020.
- [33] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, “Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3945–3954.
- [34] J. Liu, C. Li, Y. Ren, F. Chen, P. Liu, and Z. Zhao, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism,” *arXiv preprint arXiv:2105.02446*, vol. 2, 2021.
- [35] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glott. Int.*, vol. 5, no. 9, pp. 341–345, 2001.
- [36] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, “Recent developments on espnet toolkit boosted by conformer,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.
- [37] H. Xue, S. Yang, Y. Lei, L. Xie, and X. Li, “Learn2sing: Target speaker singing voice synthesis by learning from a singing teacher,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 522–529.