

# VietSing: A High-quality Vietnamese Singing Voice Corpus

Minh Vu\*, Zhou Wei†, Binit Bhattacharai‡, Kah Kuan Teh§ and Tran Huy Dat¶

Institute for Infocomm Research (I2R), Singapore

\* E-mail: minhducvu1311@gmail.com

† E-mail: zhouwei\_2001@hotmail.com

‡ E-mail: binit.bhattacharai10@gmail.com

§ E-mail: teh\_kah\_kuan@i2r.a-star.edu.sg

¶ E-mail: hdtran@i2r.a-star.edu.sg

**Abstract**—This paper introduces a comprehensive Vietnamese dataset designed for singing voice synthesis (SVS). While there are extensive datasets available for widely spoken languages such as English and Chinese, resources for less common languages like Vietnamese are still scarce. The dataset, VietSing, comprises high-quality audio recordings and corresponding phonetic annotations, meticulously curated to support the development and evaluation of SVS systems. Detailed phonetic transcriptions and alignment with musical scores are provided to facilitate precise modeling of Vietnamese phonetics and prosody in song. We outline the data collection process, annotation methodology, and the challenges faced in ensuring linguistic and musical accuracy. Initial experiments using popular SVS model demonstrate the potential of VietSing to enhance the naturalness and intelligibility of synthesized Vietnamese singing voices. The dataset is not publicly available due to licensing restrictions. Interested researchers can request access to the dataset by contacting the corresponding authors. Access will be granted subject to the approval of the appropriate licensing agreements. Some audio samples and the code for our baseline evaluation method can be found at this link<sup>1</sup>.

## I. INTRODUCTION

Recently, singing voice synthesis (SVS) using deep learning model [1]–[6] has drawn a lot of attention from both industry and academic communities. The progress in deep learning, especially in neural network based text-to-speech (TTS) systems [7]–[9], has made the production of emotional speech [10] and singing voices more attractive.

In addition to improving deep learning architectures, one of the biggest challenges in the SVS task is obtaining an appropriate singing database. Unlike TTS task, which only requires audio and corresponding textual transcriptions, SVS models must also include the musical information of the audio. The process of collecting the singing voice data tends to be costly and typically necessitates self-recording to ensure the desired quality. Although numerous SVS datasets have been released for widely spoken languages such as English [11], [12], Chinese [1], [13], [14], Japanese [15], [16], there are no available Vietnamese singing corpora. As shown in Table I, despite several singing datasets in different languages having more hours and singers, not all of them include alignment and

TABLE I  
COMPARISON OF DIFFERENT SINGING DATASETS

Corpus	Language	#Hours	#Singers	Alignment	Score
NUS-48E [11]	English	1.91	12	✓	✗
NHSS [12]	English	7	10	✗	✗
JVS-MuSiC [15]	Japanese	2.28	100	✗	✗
Tohoku Kiritan [16]	Japanese	1	1	✗	✗
PopCS [1]	Chinese	5.89	1	✗	✗
M4Singer [14]	Chinese	29.77	20	✓	✓
Openpop [13]	Chinese	5.25	1	✓	✗
VietSing (Our)	Vietnamese	2.05	1	✓	✓

musical scores, which are crucial for SVS models. Different from many other tasks, Singing Voice Synthesis often requires fairly high-quality audio with clear voice, high sampling rate, etc. We find it impossible to obtain music data with the desired quality from open internet sources, and it is also very challenging to completely eliminate background music noise. Therefore, we decide to build the dataset from self-recorded songs.

This paper introduces VietSing, a high-quality Vietnamese singing dataset designed for the SVS task. In this corpus, we select 60 songs to be recorded by a professional female singer. The audio quality is guaranteed since all songs are recorded in a studio using modern devices. Further statistics about the corpus will be discussed in the following sections. Additionally, we present experimental results from testing the dataset with DiffSinger [1] model.

## II. VIET SING DATASET

In this section, we introduce the construction pipeline and the statistical aspects of VietSing. This dataset is specifically designed for the SVS task and created using an optimized strategy that minimizes the need for manual labeling. Fig. 1 shows the steps of creating the corpus. The details of each step are described in the following subsections.

### A. Songs and singer

All the songs chosen for the corpus are Vietnamese songs from the late 1990s and early 2000s, as songs from this period typically contain fewer special sound effects, making

<sup>1</sup><https://github.com/rosotron/VietSing>

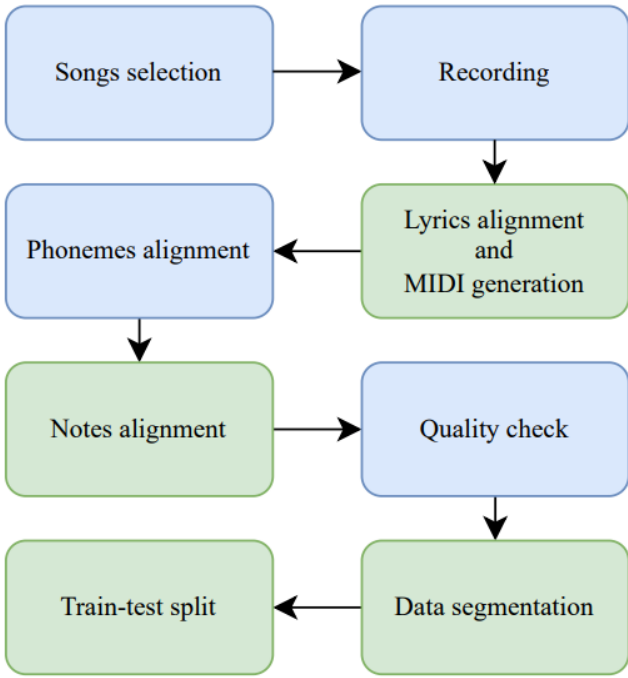


Fig. 1. Pipeline for creating VietSing. Blue boxes show manual steps and green boxes show code-based steps.

it easier for our singer to sing them. These songs must meet the following criteria:

- Lyrics are clearly sung
- Songs with code-switching lyrics are not chosen

### B. Recording

All the songs are recorded in a professional studio environment with minimal reverb. During the recording, the singer wears headphones to control the singing voice. The recordings are made using a Neumann U87 condenser microphone, capturing only the singing voice without any background music. The entire corpus is recorded at a sampling rate of 88.2 kHz with a bit depth of 24 bits.

### C. Data labeling

To minimize manual labeling effort for VietSing, we employ various code-based techniques (detailed in Fig.2). Subsequently, we implement a random quality assurance check on the automatically annotated data, which we will describe in this section.

1) *MIDI annotation*: In the context of Singing Voice Synthesis, detailed musical scores, including elements like note pitch, duration, and tempo, are crucial for training. Annotating such scores typically necessitates expertise in music theory. However, the initial absence of a qualified annotator presents a significant challenge. To address this, we construct a semi-automated annotation process that leverages publicly available, code-driven solutions to handle the musical score annotation steps.

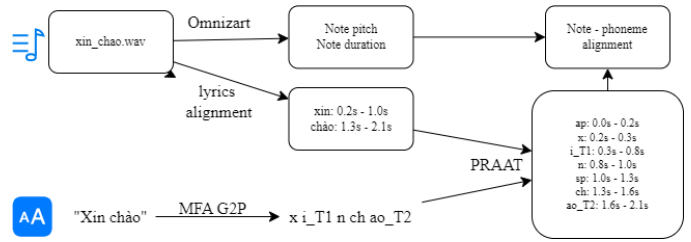


Fig. 2. Pipeline for extraction and alignment of note information and phoneme.

After recording, we create MIDI file from singing voice audio file using Omnizart [17], a Python library and a streamlined solution for automatic music transcription. Omnizart [17] provides options to extract note pitch and note duration from singing vocal. The quality of the generated MIDI files is subsequently assessed through subjective listening tests. The pitch values of the notes are translated into their respective note names and octaves, with A4 being equivalent to 69 in MIDI notation and corresponding to 440 Hz. While note duration were initially set, they are adjusted to align with manually annotated phoneme duration. The process of phoneme annotation will be explained in the following subsections.

2) *Phonemes annotation*: In Vietnamese, phonemic tone is crucial for both spoken and sung communication. While Mandarin, another tonal language, possesses a large number of Singing Voice Synthesis databases, its tonal simplicity allows listeners to comprehend lyrics independent of tone information. In Vietnamese songs, however, tone realization is important due to the interplay between semantic meaning and musical pitch. Vietnamese has a complex sound system with distinct features depending on the dialect. The standard Vietnamese phonology includes 6 tones, 14 vowel phonemes, and 19 consonant phonemes, but the Southern and Northern dialects have additional consonants unique to each region.

Due to the prevalence of the Northern dialect in Vietnamese music, particularly songs sung in the Hanoi accent, we use the Vietnamese (Hanoi) Montreal Forced Alignment (MFA) [18] G2P model v1.0.0. This model encompasses a comprehensive phonological set of 223 elements, incorporating both tones and phonemes. After splitting the lyrics into corresponding phonological elements, we perform manual phoneme alignment. We label the timestamp of each phoneme in the lyrics using PRAAT [19], this process also includes labeling special phonemes: 'sp' - spoken noise, 'sil' - silence, 'ap' - aspirate. Before manual annotation, we create a preliminary alignment using the Vietnamese song lyric alignment framework from Zalo AI Challenge 2022<sup>2</sup>, this has significantly improved the annotating pace and accuracy. Fig. 2 illustrates the steps of phoneme alignment.

### D. Note - Phoneme alignment

As described in previous sections, we extract the note pitch and duration using a deep-learning-based tool and we label

<sup>2</sup><https://github.com/nguyenvulebinh/lyric-alignment>

TABLE II  
THE NUMBER OF OCCURRENCE OF SEVERAL PHONEMES BEFORE AND AFTER DATA FILLING AND AUGMENTATION.

Phoneme	Before	After
s	1548	942
x	108	714
ch	2202	1575
tr	75	702
ã	0	196

the phoneme information manually. Consequently, the initial timestamps for phonemes and notes do not match. We assume that each note corresponds to a single phoneme, while a phoneme may extend across multiple slurred notes. To address this, we duplicate each phoneme according to the number of notes occurring within the same timestamp. These notes are designated as slurred, with the first slur note marked as 0 and the subsequent slur notes marked as 1. Conversely, for word boundaries, the initial phoneme of a word is marked with 1, while the remaining phonemes in that word are marked with 0. Special phonemes such as 'ap', 'sp', and 'sil' are automatically aligned with a note pitch of 0, and their duration are calculated based on the labels.

#### E. Phonemes filling and augmentation

Following the above steps, we analyze the database and observe differences in the frequency of phoneme occurrences, with some rare phonemes being absent from the recorded songs. To resolve this issue, we augment the dataset by adding and balancing the missing phonemes. Table II illustrates changes in occurrence number of several phonemes before and after data filling and augmentation.

1) *Phoneme augmentation*: We augment the data by leveraging the unique pronunciation characteristics of the Vietnamese Hanoi accent. Specifically, in speaking and singing with the Hanoi accent, certain phonemes are pronounced identically, such as 'ch' and 'tr', 'd', 'r', and 'gi', and 's' and 'x'. To balance the corpus, we replace the less frequent phonemes with their pronounced equivalents until we achieve a more balanced ratio between these phonemes.

2) *Missing phoneme filling*: After the phoneme augmentation step, 26 phonemes are still missing from the corpus due to their rarity and distinct pronunciation from other phonemes. To ensure the dataset fully covered the MFA [18] phone set, we record these missing phonemes. However, due to many reasons, the original singer of the corpus is unavailable for additional recordings. Therefore, we record these new phonemes with a different singer and convert the new singer's voice to the original singer's voice using the Singing Voice Conversion (SVC) technique. The singer is asked to sing each of the missing phonemes individually in random melodies, and we save each phoneme in a separate audio file. We utilize SoVITS [20] as our SVC system to clone the original singer's voice, training the model with the full set of audio data from the original singer over 200 epochs. The cloned voice audio files are then evaluated through listening tests to ensure they closely resemble the original singer's voice. After that, we run

the process of annotating notes and phonemes on the newly recorded data.

A representative data sample is presented in Table III, comprising raw audio, lyrics, phonemes, notes, note and phoneme durations, slur notes and the word boundary. This example illustrates the annotation process and the resulting transcription of an audio clip.

#### F. Post processing

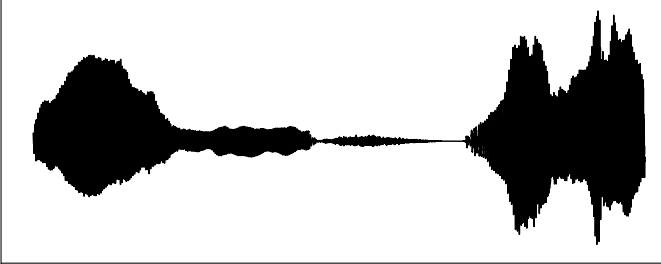
1) *Audio segmentation*: After the labeling steps, we segment each song into small parts, with each part not exceeding 11 seconds. The segmentation points are determined based on the end times of words in the lyrics. This segmentation is necessary for the SVS system to train effectively. The final corpus consists of 729 utterances and 2.05 hours of singing voice audio.

2) *Train test split*: We select 3 songs out of the 60-song corpus for the testing process. The chosen songs represent a range of average pitch values: the second lowest, the second highest, and a medium average pitch value. This selection provides a comprehensive evaluation of the SVS model. The chosen songs are song 2021 (57.35 average pitch value), song 2026 (60.20 average pitch value), and song 2034 (64.63 average pitch value). Song 2026 presents an additional challenge due to its numerous vocal ornaments and pronunciation adjustments by the singer to fit the notes, which happens sometimes in Vietnamese singing.

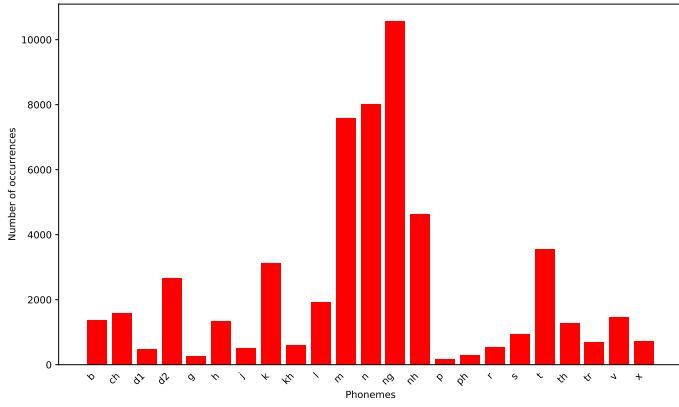
#### G. Statistics

As mentioned in Section II-C2, our phoneme set is composed of tonal phonemes, resulting in a relatively large number of distinct elements. Despite the natural rarity of certain phonemes in Vietnamese, our VietSing corpus covers all possible tonal phonemes. However, the frequency of some phonemes is higher than others. Fig. 3a and Fig. 3b display the

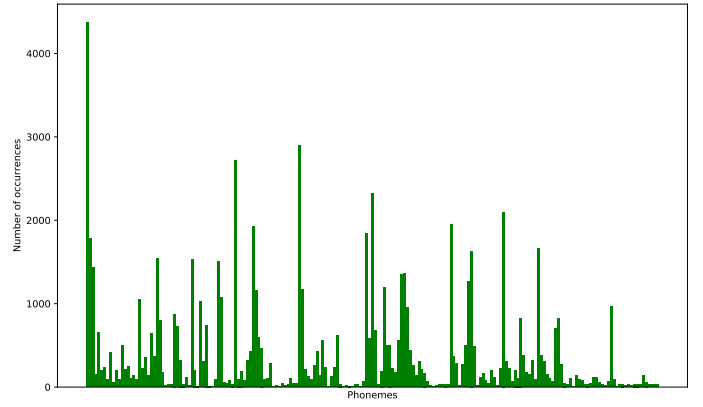
TABLE III  
TRANSCRIPTION OF A SAMPLE AUDIO CLIP. FROM TOP TO BOTTOM: WAVEFORM, LYRICS, SYLLABLE, PHONEME, NOTES, DURATION, SLUR NOTES, WORD BOUNDARY.



anh yêu anh												
anh				SP	yêu				anh			
a1_T1	nh			SP	i_T1	e2_T1	u1_T1	a1_T1	nh			
C#4	C#4	E4	D#4	Rest	D#4	A#3	A#3	A#3	D4	D4	D4	D#4
0.224	0.078	0.139	0.137	0.29	0.083	0.037	0.073	0.029	0.067	0.103	0.016	0.111
0	0	1	1	0	0	1	0	0	1	0	0	1
1	0	0	0	1	1	0	0	0	0	1	0	0



(a) Consonant phonemes



(b) Vowel phonemes

Fig. 3. The statistical distribution of phonemes

distribution of consonant and vowel phonemes in our dataset, respectively.

VietSing comprises 60 songs with varying BPMs, ranging from 80 to 220. The distribution for BPM is illustrated in Fig. 4. This diversity in tempo is beneficial for the SVS system, enabling it to recreate singing voices across a wide range of BPMs. Fig. 5 represents the statistical distribution of note pitch in VietSing. The note pitch values range from 47 (B2, 123.5 Hz) to 76 (E5, 659.3 Hz) with 60 (C4, 261.6 Hz) occurs the most.

### III. BENCHMARKS

We perform several benchmark tests to verify the quality of VietSing for singing voice synthesis task.

#### A. Method

Singing Voice Synthesis has experienced notable progress due to the emergence of deep learning methodologies. These advancements encompass innovations in non-autoregressive models [2], [21], [22], diffusion models [1], and end-to-end models [5]. Among the diverse systems designed for SVS, DiffSinger [1] and ViSinger2 [5] stands out as a prominent framework for producing high-quality singing voices. Despite

its achievements, ongoing efforts persist in enhancing the richness and expressiveness of synthesized voices, including those for Vietnamese songs.

In our experimental setup for our baseline, we employ two approaches to SVS. We use DiffSinger [1] as the acoustic model and HifiGAN [2] as the vocoder for the two stage approach and ViSinger2 [5] for the end to end approach for the singing voice synthesis task. DiffSinger [1] is grounded in the diffusion probabilistic model, while Hifi-GAN [2] serves as the vocoder. The generation process in this model determined by the intersection of diffusion trajectories between the ground-truth mel-spectrogram and the prediction from a straightforward mel-spectrogram decoder. Unlike conventional acoustic models that rely solely on basic L1 or L2 loss functions, which may lead to excessively smooth outputs. DiffSinger [1] adopts a hybrid approach, it computes the average of the L1 loss and the structural similarity (SSIM) index, resulting in mel-spectrograms that closely approximate the ground truth. To generate audio waveforms from the synthesized mel-spectrograms, we leverage a pretrained Hifi-GAN [2] vocoder.

On the other hand, ViSinger2 [5] is an end-to-end SVS model based on conditional variational autoencoder models

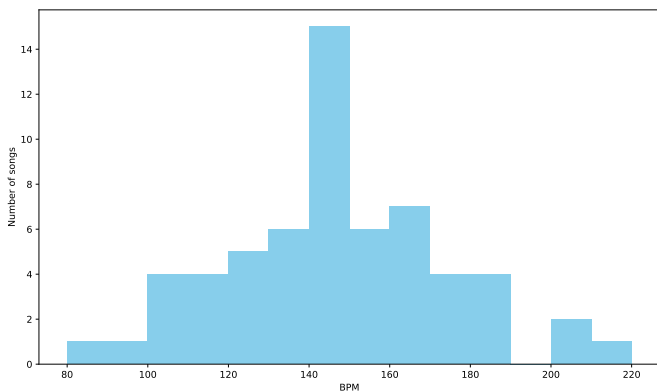


Fig. 4. The statistical distribution of BPM.

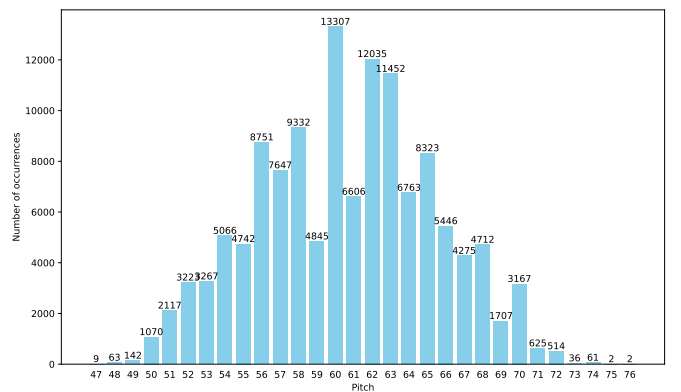


Fig. 5. The statistical distribution of note pitch with note A4 = 69 = 440 Hz.

with two separate encoders and a decoder. The posterior encoder uses 1-D convolution layers to extract latent representations from mel-spectrograms, while the decoder generates waveforms from these representations, incorporating DSP synthesizers to produce periodic and aperiodic components. These components are used to enhance HiFi-GAN [2]’s performance, mitigating phase and glitch issues. The prior encoder, based on FastSpeech [4], provides constraints by predicting fundamental frequency and mel-spectrograms. The model’s loss functions include GAN loss, KL divergence between the two encoders, DSP synthesizer loss, duration loss, and auxiliary feature loss, aiming to improve synthesis quality.

### B. Experimental setup

To perform the baseline test using DiffSinger, the audio is down-sampled to 24,000 Hz and the mel-spectrogram is extracted using 80 Mel bins and linearly scaled from -1 to 1. A pretrained vocoder, trained on 70 hours of singing data, is used as a universal vocoder. The model is trained for 1000 epochs with a learning rate of 0.001. The training process takes approximately 2 days using an NVIDIA Quadro RTX5000 16G. For training VISinger2[5], the model is trained up to 720K steps with a batch size of 4 using 3 NVIDIA Quadro RTX5000 16G.

### C. Results

To evaluate the SVS performance, we use duration accuracy (duracc) and mean opinion score (MOS). The score of MOS test ranges from 1 to 5, in which 1 means very bad and 5 means excellent. Each audio is rated by 15 qualified listeners. Within this evaluation, MOS-Q indicates the quality of the audio, while MOS-P measures the coherence and naturalness of the prosody. Table IV presents the results. The model excels in accurately reproducing duration length. The MOS scores range between 3 (average) and 4 (good), demonstrating the reliability of our corpus. Overall, the dataset produces stable audio quality (MOS-Q). For voice coherence and naturalness

TABLE IV  
OBJECTIVE AND SUBJECTIVE EVALUATION FOR VIETSing DATASET. FOR ALL THE METRICS, HIGHER VALUE MEANS BETTER PERFORMANCE, WITH "DURACC" MEANS DURATION ACCURACY OF GENERATED AUDIO COMPARED TO THE ORIGINAL AUDIO, "MOS-Q" INDICATES AUDIO QUALITY AND "MOS-P" REPRESENTS COHERENCE AND NATURALNESS OF PROSODY.

Test Set	Method	duracc	MOS-Q	MOS-P
low-pitch	Ground Truth	-	4.79±0.24	4.41±0.09
	DiffSinger[1]	0.996	3.80±0.20	3.79±0.21
	VISinger2[5]	0.994	3.65±0.05	3.70±0.10
mid-pitch	Ground Truth	-	4.70±0.20	4.57±0.07
	DiffSinger[1]	0.992	3.94±0.06	3.85±0.05
	VISinger2[5]	0.996	3.75±0.05	3.75±0.15
high-pitch	Ground Truth	-	4.65±0.05	4.62±0.02
	DiffSinger[1]	0.989	3.71±0.19	3.69±0.21
	VISinger2[5]	0.992	3.60±0.10	3.69±0.11

TABLE V  
THE AUDIO PERFORMANCE BEFORE AND AFTER DATA AUGMENTATION BY DIALECT CHARACTERISTICS.

Method	duracc	MOS-Q	MOS-P
Ground Truth	-	4.32±0.18	4.41±0.09
DiffSinger[1] - before augment	0.966	3.70±0.10	3.44±0.06
DiffSinger[1] - after augment	0.987	3.73±0.17	3.59±0.11

TABLE VI  
THE AUDIO PERFORMANCE OF VOICE-CLONED PHONEMES BY DIFFSINGER[1].

Metric	Score
MOS-Q	3.15±0.10
MOS-P	3.26±0.04

(MOS-P), the corpus receives slightly lower scores, but they remain at an acceptable level.

Moreover, to assess the effectiveness of phoneme augmentation and filling as discussed in Section II-C2, we prepare a separate test song for the model trained with data both before and after augmentation. To evaluate the missing phonemes filled by the SVC [20] technique, we modify the lyrics and phonemes of the test song to include these voice-cloned phonemes. The performance of these tests are measured using MOS-Q and MOS-P and is presented in Tables V and VI. As shown, data augmentation generally improves the synthesis quality. For the phonemes filled by So-VITS [20], the scores remain at acceptable levels, though it is true that the tonality of these phonemes is not very well covered.

## IV. CONCLUSION

In this paper, we present the VietSing dataset, a Vietnamese singing voice corpus specifically designed for SVS systems. All audio songs in our corpus are self-recorded using professional equipment. The annotation process combines manual and automatic methods, and we augment the data by leveraging unique features of the Vietnamese dialect.

We acknowledge several limitations of VietSing. First, despite random checks for annotation quality, there are inevitable errors in musical labeling when using deep learning methods. Second, to achieve the best synthesized results with our corpus, we recommend users to add simulated slur notes, as we have marked many notes as slurred to align with the notes detected by Omnizart [17]. Finally, we acknowledge that several tonal phonemes appear less frequently than others due to their natural rarity in music lyrics, which may result in occasional lack of clarity for these phonemes.

## REFERENCES

- [1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diff-singer: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 11 020–11 028.

- [2] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.
- [3] 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.
- [4] Y. Ren, C. Hu, X. Tan, *et al.*, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *International Conference on Learning Representations*, 2021.
- [5] Y. Zhang, H. Xue, H. Li, *et al.*, “VISinger2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer,” in *Proc. INTERSPEECH 2023*, 2023, pp. 4444–4448. DOI: 10.21437/Interspeech.2023-391.
- [6] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on convolutional neural networks,” *arXiv preprint arXiv:1904.06868*, 2019.
- [7] X. Tan, J. Chen, H. Liu, *et al.*, “Naturalspeech: End-to-end text-to-speech synthesis with human-level quality,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [8] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-tts: A generative flow for text-to-speech via monotonic alignment search,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
- [9] A. Łańcucki, “Fastpitch: Parallel text-to-speech with pitch prediction,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592.
- [10] “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022, ISSN: 0167-6393.
- [11] 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.
- [12] 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.
- [13] Y. Wang, X. Wang, P. Zhu, *et al.*, “Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis,” in *Proc. Interspeech 2022*, 2022, pp. 4242–4246. DOI: 10.21437/Interspeech.2022-48.
- [14] L. Zhang, R. Li, S. Wang, *et al.*, “M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 6914–6926, 2022.
- [15] H. Tamaru, S. Takamichi, N. Tanji, and H. Saruwatari, “Jvs-music: Japanese multispeaker singing-voice corpus. arxiv 2020,” *arXiv preprint arXiv:2001.07044*,
- [16] I. Ogawa and M. Morise, “Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs,” *Acoustical Science and Technology*, vol. 42, no. 3, pp. 140–145, 2021.
- [17] Y.-T. Wu, Y.-J. Luo, T.-P. Chen, *et al.*, “Omnizart: A general toolbox for automatic music transcription,” *Journal of Open Source Software*, vol. 6, no. 68, p. 3391, 2021. DOI: 10.21105/joss.03391. [Online]. Available: <https://doi.org/10.21105/joss.03391>.
- [18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi.,” in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [19] P. Boersma and V. Van Heuven, “Speak and unspeak with praat,” *Glott International*, vol. 5, no. 9/10, pp. 341–347, 2001.
- [20] Z. Ning, Y. Jiang, Z. Wang, B. Zhang, and L. Xie, “Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.
- [21] 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.
- [22] C. Wang, C. Zeng, and X. He, “Xiaoicesing 2: A high-fidelity singing voice synthesizer based on generative adversarial network,” *arXiv preprint arXiv:2210.14666*, 2022.