GTSinger: A Global Multi-Technique Singing Corpus with Realistic Music Scores for All Singing Tasks

Yu Zhang[†] Changhao Pan[†] Wenxiang Guo[†] Ruiqi Li Zhiyuan Zhu Jialei Wang
Wenhao Xu Jingyu Lu Zhiqing Hong Chuxin Wang LiChao Zhang Jinzheng He
Ziyue Jiang Yuxin Chen Chen Yang Jiecheng Zhou Xinyu Cheng Zhou Zhao[*]
Zhejiang University
{yuzhang34,panch,guowx314,zhaozhou}@zju.edu.cn

Abstract

The scarcity of high-quality and multi-task singing datasets significantly hinders the development of diverse controllable and personalized singing tasks, as existing singing datasets suffer from low quality, limited diversity of languages and singers, absence of multi-technique information and realistic music scores, and poor task suitability. To tackle these problems, we present **GTSinger**, a large **G**lobal, multi-**T**echnique, free-to-use, high-quality singing corpus with realistic music scores, designed for all singing tasks, along with its benchmarks. Particularly, (1) we collect 80.59 hours of high-quality singing voices, forming the largest recorded singing dataset; (2) 20 professional singers across nine widely spoken languages offer diverse timbres and styles; (3) we provide controlled comparison and phoneme-level annotations of six commonly used singing techniques, helping technique modeling and control; (4) GTSinger offers realistic music scores, assisting real-world musical composition; (5) singing voices are accompanied by manual phoneme-to-audio alignments, global style labels, and 16.16 hours of paired speech for various singing tasks. Moreover, to facilitate the use of GTSinger, we conduct four benchmark experiments: technique-controllable singing voice synthesis, technique recognition, style transfer, and speech-to-singing conversion. The corpus and demos can be found at http://gtsinger.github.io. We provide the dataset and the code for processing data and conducting benchmarks at https://huggingface.co/datasets/GTSinger/GTSinger and https://github.com/GTSinger/GTSinger.

1 Introduction

Traditional singing tasks, typically singing voice synthesis (SVS) [11, 29], aim to generate high-quality singing voices using lyrics and musical notations, attracting broad interest in the industry and academic communities. As deep learning technology advances, there is a growing demand for more controllable and personalized singing experiences. This burgeoning demand has catalyzed the emergence of various new singing tasks like technique-controllable SVS, technique recognition, style transfer, and speech-to-singing (STS) conversion [1, 13]. These tasks have been progressively developed and applied in real life, like short videos and professional composition [28].

Despite the significant progress made in multiple singing tasks, the scarcity of publicly available high-quality and multi-task singing datasets has become a major bottleneck in their development due to the high cost of recording songs and manual annotations. The primary limitations of existing open-source singing datasets are as follows: 1) The **low quality** [19] may lead to singing models

---

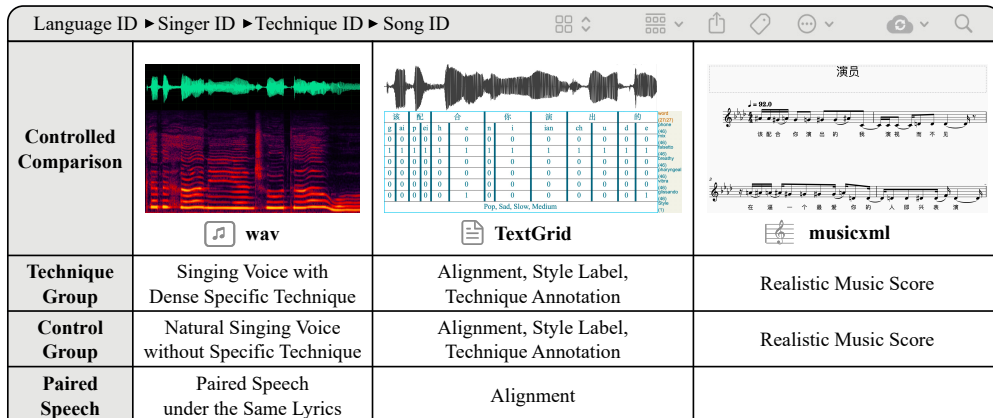[†]Equal contribution
[*]Corresponding Author

Figure 1: The composition of each song in GTSinger. There are 1,366 songs recorded by 20 singers using six singing techniques across nine languages. Each song contains a technique group and a control group for the controlled comparison, along with a paired speech for STS tasks. Alignments, style labels, technique annotations, and realistic music scores are manually created for each group.

producing off-pitch, unpleasant, or noisy results. 2) A limited variety in **languages** [25] and **singers** [23] restricts personalized singing models to learn diverse timbres and styles. 3) The absence of the controlled comparison and annotations for multiple **singing techniques** (like falsetto) [4], constrains the technique modeling and control for singing models. 4) The lack of **realistic music scores** [8] hinders human composers from using singing models in real-world musical composition. 5) Poor **task suitability** [26] forces multiple emerging singing tasks to customize new datasets with high cost.

To address these challenges, we introduce **GTSinger**, a large **G**lobal, multi-**T**echnique, free-to-use, high-quality **Sing**ing corpus with realistic music scores, designed for all singing tasks. Our dataset contains 80.59 hours of high-quality singing voices without accompaniment, delivered by 20 professional singers covering nine widely spoken languages, including **Chinese, English, Japanese, Korean, Russian, Spanish, French, German, and Italian**. Moreover, GTSinger integrates phoneme-level annotations for six commonly used singing techniques, namely **mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando**. As shown in Figure 1, each song includes a control group for the natural singing voice and a technique group that intensively applies specific techniques under the same lyrics. Additionally, GTSinger furnishes realistic music scores for real-world musical composition. We also incorporate manual phoneme-to-audio alignments, global style labels (including singing method, emotion, range, and pace), and 16.16 hours of paired speech for various singing tasks. Overall, GTSinger boasts several advantages for multiple singing tasks over existing singing datasets:

- **80.59 hours of singing voices** in GTSinger are recorded in professional studios by skilled singers, ensuring **high quality and clarity**, forming the largest recorded singing dataset.
- Contributed by **20 singers** across **nine widely spoken languages** and all four vocal ranges, GTSinger enables zero-shot SVS and style transfer models to learn diverse timbres and styles.
- GTSinger provides **controlled comparison** and **phoneme-level annotations** of **six singing techniques** for songs, thereby facilitating singing technique modeling, recognition, and control.
- Unlike fine-grained music scores, GTSinger features **realistic music scores** with regular note duration, assisting singing models in learning and adapting to real-world musical composition.
- The dataset includes **manual phoneme-to-audio alignments**, **global style labels**, and **16.16 hours of paired speech**, ensuring comprehensive annotations and broad task suitability.

The rest of the paper is organized as follows. In Section 2, we briefly review and compare with current singing datasets. In Section 3, we provide the construction details and data statistics of GTSinger. In Section 4, to demonstrate the use and validate the quality of GTSinger, we conduct extensive experiments and establish benchmarks for four different singing tasks, including **technique-controllable singing voice synthesis, technique recognition, style transfer, and STS conversion**, employing recently published state-of-the-art methods for each task. In Section 5, we make the conclusion and discuss some potential risks along with the limitations of GTSinger.

2 Related Work

The advancement of deep learning has enabled singing models, like singing voice synthesis (SVS) models, to achieve remarkably high-quality vocal results [11, 29, 19, 27, 15, 7]. Fueled by the growing demand for controllable and personalized singing experiences, diverse new singing tasks have emerged, like technique-controllable SVS, technique recognition, style transfer, and speech-to-singing (STS) conversion [9, 21, 1, 13, 16, 28]. Unlike traditional datasets [19, 6, 22] designed for a singular task, a high-quality and multi-task singing dataset has higher demands. A high-quality dataset not only requires recordings by professional singers with manual alignments but also needs to include multiple singers and languages to broaden timbres and styles. Furthermore, a multi-task dataset also needs to feature controlled comparisons and phoneme-level annotations of singing techniques for technique modeling, realistic music scores for real-world musical composition, global style labels for global control, and paired speech for STS tasks. These requirements significantly elevate the recording and annotation costs, explaining the scarcity of high-quality and multi-task datasets.

Table 1: The information table of existing open-source singing datasets. Align and RMS mean manual phoneme-to-audio alignment and realistic music scores. Style denotes global style labels.

| Corpus | Language | Singer | Hours | | Manual Annotations | | | | Controlled Comparison |
			Singing	Speech	Align	RMS	Tech	Style	
VocalSet [25]	1	20	10.1	0	✗	✗	✗	✗	✓
CSD [4]	2	1	4.86	0	✗	✗	✗	✗	✗
KVT [10]	1	114	18.85	0	✗	✗	✗	✓	✗
PopBuTFy [16]	2	34	50.8	0	✗	✗	✗	✗	✗
OpenSinger [8]	1	66	50	0	✗	✗	✗	✗	✗
NHSS [20]	1	10	4.75	2.25	✗	✗	✗	✗	✗
Tohoku Kiritan [18]	1	1	1	0	✓	✗	✗	✗	✗
OpenCpop [23]	1	1	5.25	0	✓	✗	✗	✗	✗
M4Singer [26]	1	20	29.77	0	✓	✗	✗	✗	✗
GTSinger (Ours)	**9**	**20**	**80.59**	**16.16**	✓	✓	✓	✓	✓

As delineated in Table 1, several datasets endeavor to mitigate specific challenges to cater to designated tasks. VocalSet [25] provides the controlled comparison of various singing techniques, albeit constrained by its reliance solely on a range of vowels rather than linguistic content in singing voices. CSD [4] features recordings in both English and Korean, yet a limited number of vocalists constrains its diversity. KVT [10] annotates some types of global style labels in K-pop songs but uses existing songs and does not separate vocals from accompaniments. PopBuTFy [16] provides singing voices in both English and Chinese, but without annotations. OpenSinger [8] encompasses a substantial volume of vocal recordings across numerous singers, yet it does not contain any annotation. NHSS [20] introduces paired speech for STS tasks but falls short in providing manual phoneme-level alignments and other annotations. Tohoku Kiritan [18] provides manual alignments but is limited by its small scale. Opencpop [23] and M4Singer [26] mark significant advancements with their manual alignments and music scores. However, they only provide fine-grained music scores, which disrupt the regularity of note duration and thus, hinder the application to real-world musical composition. Moreover, they lack other annotations and paired speech for more tasks. In this paper, we construct a large multi-lingual, multi-singer, free-to-use, high-quality singing corpus with controlled comparison and phoneme-level annotations of multiple techniques, along with manual phoneme-to-audio alignments, realistic music scores, global style labels, and paired speech. We seek to comprehensively address the limitations in previous singing datasets and cater to all current singing tasks.

3 Dataset Description

In this section, we formally introduce GTSinger, a large global, multi-technique, free-to-use, high-quality singing corpus with realistic music scores, which aims to support all current singing tasks and can be used under license CC BY-NC-SA 4.0. Figure 2 depicts the pipeline of the creation of GTSinger, with detailed explanations of each step provided in the following subsections. Then, we present the necessary dataset statistics to enhance the understanding of our GTSinger. At last, we also provide the instructions to use our dataset and codes.
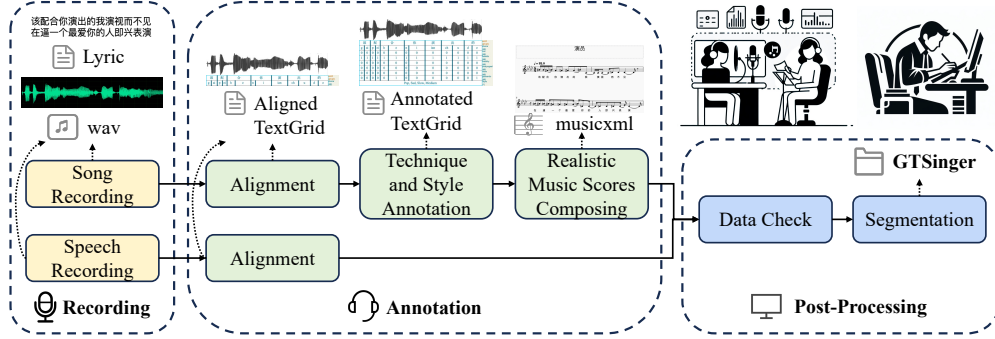
Figure 2: The pipeline of data collection of GTSinger. Human double-checks exist in each process.

3.1 Songs and Singers

To construct GTSinger, we first select nine widely spoken languages: Chinese, English, Japanese, Korean, Russian, Spanish, French, German, and Italian. Then, we also choose six commonly used singing techniques: mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando. After rigorous auditions, we select 20 professional singers, covering all four vocal ranges (alto, soprano, tenor, bass), and each singer is proficient in all six techniques and one of the widely spoken languages. Before recording, all singers agree to make their vocal performances open-source for academic research. We carefully select songs based on the representativeness of each language, the vocal range of each singer, and the suitability of singing each technique densely. As shown in Table 2, all singers are listed and anonymized by their languages along with vocal ranges.

Table 2: The information table of languages, singers, techniques, and duration. Singing hours for technique ID count time for singing voices in both control groups and technique groups.

Language ID	Singer ID	Total Hours		Singing Hours of Technique ID				
		Singing	Speech	Mixed Voice and Falsetto	Breathy	Pharyngeal	Vibrato	Glissando
Chinese (ZH)	ZH-Tenor-1	8.45	1.82	3.6	1.26	1.18	1.18	1.23
	ZH-Alto-1	8.14	1.49	3.7	1.13	1.06	1.13	1.12
English (EN)	EN-Tenor-1	4.76	0.87	2.06	0.69	0.65	0.7	0.66
	EN-Alto-1	3.47	0.67	1.6	0.52	0.51	0.28	0.56
	EN-Alto-2	4.9	1.04	2.05	0.74	0.67	0.73	0.71
Japanese (JA)	JA-Tenor-1	2.13	0.29	1.01	0.33	0.34	0.15	0.3
	JA-Soprano-1	4.32	0.87	2.24	0.56	0.41	0.53	0.58
Korean (KO)	KO-Tenor-1	4.61	1.32	1.19	0.87	0.88	0.83	0.84
	KO-Soprano-1	0.95	0.24	0.19	0.16	0.2	0.21	0.19
	KO-Soprano-2	2.72	0.61	1.12	0.37	0.42	0.42	0.39
Russian (RU)	RU-Alto-1	4.32	0.76	1.81	0.63	0.55	0.7	0.63
Spanish (ES)	ES-Bass-1	4.45	0.9	2.01	0.61	0.61	0.61	0.61
	ES-Soprano-1	3.48	0.82	1.4	0.59	0.4	0.53	0.56
French (FR)	FR-Tenor-1	4.58	0.58	1.27	0.9	0.84	0.66	0.91
	FR-Soprano-1	3.96	0.59	1.75	0.58	0.58	0.57	0.48
German (DE)	DE-Tenor-1	4.54	0.9	2.19	0.56	0.59	0.59	0.61
	DE-Soprano-1	4.54	0.82	1.9	0.64	0.63	0.67	0.7
Italian (IT)	IT-Bass-1	3.21	0.82	0.86	0.76	0.17	0.68	0.74
	IT-Bass-2	1.61	0.4	0.32	0.32	0.3	0.33	0.34
	IT-Soprano-1	1.45	0.35	0.98	0.11	0.1	0.05	0.21
All	**All**	**80.59**	**16.16**	**33.25**	**12.33**	**11.09**	**11.55**	**12.37**

3.2 Recording

Singers perform a multitude of songs, each selected to highlight a specific singing technique (like falsetto). For each song, they maintain a consistent rhythm, lyrics, and key, recording twice: once

4

densely applying the specific technique (technique group) and once for the natural singing voice without the specific technique(control group). We especially manage falsetto and mixed voice techniques due to their correlations. They form a distinct group, recording a natural singing voice (control group), and two technique groups, for both falsetto and mixed voice. Furthermore, each song includes an additional spoken lyric sentence recorded by the same singer, providing paired speech for STS tasks. All recordings are carried out in a professional studio, with singers listening to the song's accompaniment through headphones, ensuring clean vocal tracks devoid of accompaniment yet preserving rhythm and timing. Each audio is recorded at a 48kHz sampling rate with 24 bits in WAV format, ensuring high-quality data for further statistics and research. Table 2 presents the duration of 1,366 final recorded songs. For more details, please refer to Appendix A.1.

3.3 Annotation

Alignment: We initially use the Montreal Forced Aligner (MFA) [17] for a coarse alignment of the original lyrics and audio and store the results in TextGrid format. Chinese phonemes are extracted using pypinyin, English phonemes follow the ARPA standard, Italian phonemes follow the Epitran standard, and others follow the MFA standard. These are the most effective and suitable phoneme standards for these languages. Next, annotators with a musical background use Praat [2] to correct the rough annotation results, focusing on the following areas: (1) Boundary correction: Annotators correct the boundaries of words and phonemes by listening to the audio and observing the mel-spectrogram, which forms the bulk of this step. (2) Word and phoneme correction: In cases of missing or incorrect lyrics, annotators are required to correct the words and corresponding phonemes based on their auditory perception. This is because singers may mispronounce words, or there may be homophones in Chinese that cause phoneme errors. (3) Unvoiced labeling: The unvoiced region, including breathing and silent sections, is marked by annotators who identify the boundaries respectively. In this step, we perform alignment for both the singing voice and paired speech.

Technique and Style Annotation: Following the alignment process, we instruct our annotators to perform phoneme-level annotations of six singing techniques on the TextGrid, including mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando. Annotators continue to use Praat [2] for annotations based on their auditory perception, indicating the presence or absence of each technique for every phoneme. Notably, in technique groups, singing voices employ densely specific techniques but not exclusively, as other techniques may also be used. In control groups, specific techniques are excluded but other techniques can be present as singers are asked to sing naturally. Next, annotators also label the singing method (pop and bel canto), emotion (happy and sad), pace (slow, moderate, and fast), and range (low, medium, and high) as global style labels for each group.

(a) Difference between F0 and music scores

(b) Fine-gained music scores
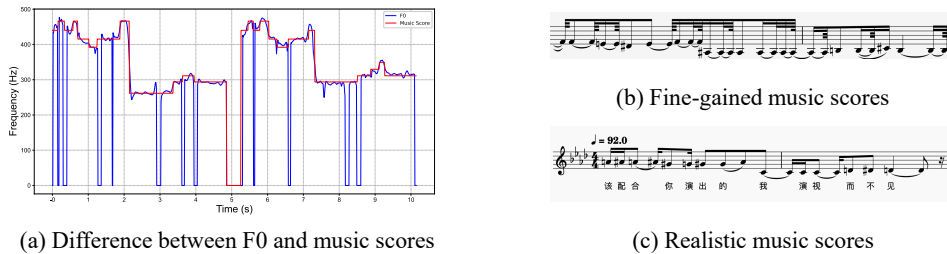
(c) Realistic music scores

Figure 3: The comparison between F0, fine-grained music scores, and realistic music scores. Score pitches are converted to frequencies and are very different from F0. Fine-grained music scores disrupt the regularity of note duration, resulting in fragmented notes that are unsuitable for composing.

Realistic Music Score Composing: The difference in F0, fine-grained music scores, and realistic music scores are depicted in Figure 3. To compose realistic music scores, we initially employ RMVPE [24] to extract F0 for each singing voice. Then, we use ROSVOT [14] to derive the MIDI form of the music scores. The MIDI is obtained by referring to the F0 curve for determining the note pitch and duration. Subsequently, we engage music experts to listen to the recorded songs, refer to original accompaniments, and carry out the following steps: 1) Determine the actual tempo, clef, and key. 2) Adjust the music scores to match the true note pitch. 3) Modify the note duration following regular realistic music score rules. 4) Annotate the note type to be rest, lyric, or slur. The outcome is realistic music scores in the muxicxml format. More annotation details can be found in Appendix A.2.

3.4 Post-Processing

Data Check: For each language with fully annotated data, we employ an additional music expert proficient in that language to randomly inspect 25% of the annotations. Their primary tasks include: (1) Checking alignment, including word and phoneme boundaries, incorrect characters, polyphonic phonemes in Chinese data, and annotations of unvoiced sections. (2) Examining technique and style annotations, focusing on annotations of techniques outside the specific group. (3) Reviewing realistic music scores, paying attention to key, tempo, and clef, and correcting note pitch and duration.

Segmentation: After completing the data annotation and inspection, we segment the audio into smaller fragments to facilitate training for singing tasks. For the same song, the control group, technique group, and paired speech are synchronously segmented into sentence-level segments, with their alignments, annotations, and scores correspondingly segmented. By leveraging the manual alignment results, we set a threshold for the unvoiced region and established maximum and minimum lengths for the voiced region as the conditions for performing the segmentation process. As shown in Figure 4 (a), we ensure more than 95% sentences are between 5 and 20 seconds in duration. Finally, we get 29,261 singing utterances and 12,373 speech utterances.

3.5 Statistics:

(a) The distribution of segment duration   (b) The distribution of techniques   (c) The distribution of BPM
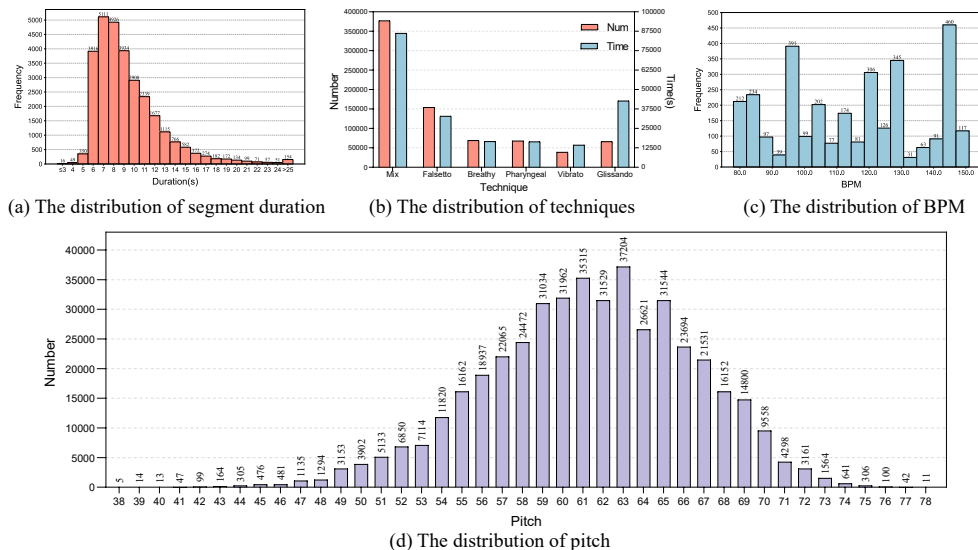

(d) The distribution of pitch

Figure 4: The statistical distribution of segment duration, techniques, BPM, and pitch.

To illustrate the diverse range of recorded singing voices, Figure 4 (b) presents the distribution of six singing techniques. The prevalence of mixed voice can be attributed to its widespread use in both pop and bel canto singing methods. Conversely, vibrato is the least frequent because it can only be used on sustained notes, resulting in a very low distribution density. Figure 4 (c) presents the distribution of beats per minute (BPM) across 3,145 groups (each song contains technique and control groups). The majority of groups fall within the 80 to 150 BPM range, encompassing the primary range for both pop and bel canto singing methods. Figure 4 (d) illustrates that note pitches of realistic music scores are primarily distributed between MIDI note numbers 50 (D3, 146.83 Hz) and 70 (B4, 494 Hz), covering four vocal ranges. This broad spectrum of languages, singers, techniques, BPM, and pitch suggests the potential of models trained on GTSinger to effectively handle a diverse range of singing styles. Refer to Appendix A for further detailed statistics.

3.6 Instructions for Use

The dataset can be freely downloaded at https://huggingface.co/datasets/GTSinger/GTSinger and noncommercially used under license CC BY-NC-SA 4.0. Users can define additional tasks under this license. We also provide the code for processing data at https://github.com/GTSinger/GTSinger. For suggestions, please contact us via email. Regular updates will be provided on our GitHub repository.

6

4   Benchmarks

In this section, to assess the quality and versatility of GTSinger, we conduct a comprehensive evaluation across four singing tasks: technique-controllable SVS, technique recognition, style transfer, and speech-to-singing (STS) conversion. We evaluate GTSinger with recently published state-of-the-art methods, utilizing four NVIDIA 2080Ti GPUs for all experimental protocols. Hifi-GAN [12] is employed as the vocoder for audio synthesis. We conduct the MOS (mean opinion score), FFE, MCD, and Cos for evaluation of these tasks on the test set. For more details about the evaluation, please refer to Appendix B.1 and B.2. For more detailed results, please refer to Appendix B.

4.1   Technique-Controllable Singing Voice Synthesis

Previous SVS models are limited by datasets and cannot achieve technique-controllable SVS. Technique-controllable SVS can provide more controllable and personalized singing experiences for real-world applications, allowing even those who cannot sing to customize a variety of professional techniques. We randomly choose five songs from each singer, totaling 100 songs as the test set, with the remainder as the training set. We conduct experiments on the following systems: (1) GT: The ground truth singing voice; (2) GT (vocoder): The audio generated by the pre-trained HiFi-GAN; (3) DiffSinger [15]: A popular SVS system based on the diffusion model; (4) RMSSinger [7]: An outstanding SVS system using the diffusion model to predict F0. (5) StyleSinger [28]: the current state-of-the-art SVS system using the residual style encoder to model both global and detailed styles. We enhance both models with phoneme-level technique embedding to achieve technique control.

We employ both objective and subjective evaluation metrics for evaluation. For subjective evaluation, MOS-Q indicates the quality, naturalness, and clarity of the synthesized audio, while MOS-C reflects the expressiveness and accuracy of technique controllability. Both metrics are rated on a scale from 1 to 5 and reported with 95% confidence intervals. For objective evaluation, we use F0 Frame Error (FFE) to measure the accuracy of F0 and UV prediction, and Mean Cepstral Distortion (MCD) for audio quality measurement. We conduct both parallel and non-parallel experiments according to the target technique sequence. In the parallel experiments, we use the GT technique sequence as the target. In the non-parallel experiments, six techniques are randomly yet appropriately assigned to each target phoneme (zero, one or more techniques on each phoneme).

Table 3: Technique-controllable SVS performance in both parallel and non-parallel experiments. We use FFE, MCD, MOS-Q, and MOS-C for comparisons.

Method	Parallel				Non-Parallel	
	FFE $\downarrow$	MCD $\downarrow$	MOS-Q $\uparrow$	MOS-C $\uparrow$	MOS-Q $\uparrow$	MOS-C $\uparrow$
GT	-	-	4.54 ± 0.06	-	-	-
GT (vocoder)	0.05	1.33	4.21 ± 0.07	4.42 ± 0.03	-	-
DiffSinger [15]	0.29	3.58	3.81 ± 0.06	3.83 ± 0.07	3.77 ± 0.05	3.78 ± 0.07
RMSSinger [7]	0.27	3.43	3.94 ± 0.07	3.95 ± 0.05	3.86 ± 0.06	3.89 ± 0.06
StyleSinger [28]	0.25	3.27	4.01 ± 0.09	4.15 ± 0.06	3.95 ± 0.08	4.10 ± 0.05

As shown in Table 3, we can observe the following: (1) Leveraging the diffusion decoder, DiffSinger achieves reasonable sound quality (MOS-Q). However, it struggles to model and control techniques effectively (MOS-C). Additionally, the high FFE and VDE indicate that the generated styles deviate significantly from the actual curve. (2) By using a diffusion model to model F0, RMSSinger significantly improves its ability to handle techniques (MOS-C) and achieves better synthesis quality (MOS-Q). This highlights the impact of pitch modeling on technique representation. (3) StyleSinger integrates multi-level style information from the reference mel-spectrograms, and renders techniques more naturally and expressively, outperforming all other baseline models across all metrics. This demonstrates the complexity involved in modeling techniques across different singing styles. However, these results indicate that there is still a significant gap between the model's performance and GT (vocoder), highlighting ample room for improvement in the technique controllability of singing tasks. Future work can explore using more advanced generation models to incorporate realistic music scores and leverage phoneme-level technique information for better F0 and mel-spectrogram generation, as well as higher technique controllability. For more detailed and visualized results about technique-controllable SVS, please refer to Appendix B.3.

4.2 Technique Recognition

Technique recognition aims to predict the techniques in unseen audio samples, facilitating the augmentation of existing singing datasets with technique annotations and aiding the real-world learning of singing techniques. We design a technique recognition model based on ROSVOT [14] and change the loss to the cross entropy loss of each technique label. The inputs of the technique recognition model include the mel-spectrogram, pitch, and phoneme boundaries, with the output being the predicted probabilities of six techniques in each phoneme. We conduct both overall and cross-lingual experiments to evaluate our model's performance and the annotation quality of GTSinger. We categorize the languages into two groups: Asian (Chinese, Japanese, and Korean) and European (Italian, Spanish, English, French, German, and Russian). In overall experiments, We reuse the rule in Section 4.1 to split training and test sets. In cross-lingual experiments, models are trained on one group of languages (like Asian) and tested on the other language group (like European) to assess their generalization capabilities. For evaluation, we provide F1 and Accuracy.

Table 4: F1 and Accuracy of each technique in overall and cross-lingual technique recognition.

Experiment	Metric	Technique Recognition Accuracy					
		mixed voice	falsetto	breathy	pharyngeal	vibrato	glissando
Overall	F1	0.78	0.96	0.99	0.85	0.70	0.70
	Accuracy	0.78	0.84	0.78	0.80	0.89	0.85
Cross-Lingual	F1	0.72	0.94	0.96	0.84	0.66	0.64
	Accuracy	0.75	0.79	0.72	0.77	0.84	0.78

As shown in Table 4, our model demonstrates good F1 and Accuracy of all six techniques in both overall and cross-lingual experiments, highlighting the quality of our technique recognition model, as well as the merit of designing controlled comparison and phoneme-level annotation of six techniques in GTSinger. However, it is evident that the overall performance still surpasses that of cross-lingual cases. This indicates ample room for improvement in the technique recognition model, suggesting the potential to enhance generalization, thus handling out-of-domain technique recognition better. For more details about the model and results, please refer to Appendix B.4.

4.3 Style Transfer

Style transfer aims to generate high-quality singing voices with the timbre and styles (like singing methods, rhythm, techniques, and pronunciation) of the reference audio. This technology can be applied in the dubbing of entertainment short videos, offering personalized experiences. We reuse the rule in Section 4.1 to split training and test sets. For baseline models, we reuse StyleSinger, as it is the first singing style transfer model, and enrich RMSSinger with additional singer and emotion embedding for conducting style transfer like previous works [28]. For subjective evaluation, we use MOS-Q for synthesis quality and MOS-S for singer similarity in terms of timbre and styles. For objective evaluation, we employ FFE to measure pitch accuracy, MCD to assess synthesis quality, and Cos for evaluation of singer similarity. We conduct both parallel and cross-lingual style transfer experiments. For parallel experiments, we use another singing voice by the same singer as the reference audio. For cross-lingual experiments, we also split languages into Asian and European groups like Section 4.2. Then we randomly select reference audio from one language group (like Asian), transferring singing styles to target lyrics in the other language group (like European).

Table 5: Style Transfer performance in both parallel and cross-lingual experiments. We use FFE, MCD, Cos, MOS-Q, and MOS-C for comparisons.

Method	Parallel					Cross-Lingual	
	FFE ↓	MCD ↓	Cos ↑	MOS-Q ↑	MOS-S ↑	MOS-Q ↑	MOS-S ↑
GT	-	-	-	4.53 ± 0.03	-	-	-
GT (vocoder)	0.05	1.34	0.96	4.18 ± 0.04	4.26 ± 0.03	-	-
RMSSinger	0.31	3.47	0.88	3.70 ± 0.04	3.79 ± 0.06	3.66 ± 0.04	3.76 ± 0.08
StyleSinger	0.26	3.29	0.93	3.95 ± 0.06	4.01 ± 0.05	3.89 ± 0.07	3.92 ± 0.09

As shown in Table 5, we can observe that StyleSinger achieves impressive results in both synthesized quality (MOS-Q) and singer similarity (MOS-S), which suggests that GTSinger's extensive style collection facilitates modeling and transfer, enabling the model to achieve high performance. Additionally, StyleSinger performs well in cross-lingual tasks, showcasing its ability to sing any music scores and lyrics, regardless of the singer's identity. However, there is still ample room for improvement in the cross-lingual style transfer performance. Future research can explore specialized models for handling singing style transfer tasks involving significant style differences. For more detailed and visualized results about style transfer, please refer to Appendix B.5.

4.4 Speech-to-Singing Conversion

Speech-to-singing (STS) conversion aims to transform speech into the corresponding singing voice preserving the timbre and phoneme information. STS can be applied to automatic music production or personalized entertainment. We randomly select five songs (including paired speech) from each singer, totaling 100 pairs as the test set, and others as the training set. Besides GT and GT (vocoder), we use AlignSTS [13] as a baseline model, and design another model based on StyleSinger, which inputs paired speech as the reference audio to conduct STS conversion. StyleSinger uses realistic musical scores (RMS), which are more practical, while AlignSTS requires GT singing F0 as input. We reuse evaluation metrics in Section 4.3 for evaluating synthesized quality and singer similarity.

Table 6: Speech-to-singing performance in FFE, MCD, Cos, MOS-Q, and MOS-S metrics.

Method	FFE ↓	MCD ↓	Cos ↑	MOS-Q ↑	MOS-S ↑
GT	-	-	-	4.53 ± 0.03	-
GT (vocoder)	0.05	1.34	0.95	4.17 ± 0.05	4.20 ± 0.04
AlignSTS	0.35	3.52	0.85	3.68 ± 0.12	3.73 ± 0.09
StyleSinger	0.28	3.38	0.92	3.83 ± 0.09	3.88 ± 0.08

As shown in Table 6, we observe that StyleSinger outperforms AlignSTS in both synthesis quality (MOS-Q) and singer similarity (MOS-S). This also demonstrates the quality of our annotations of realistic music scores. Using realistic music scores and speech to synthesize expected singing voices allows for more controllable and personalized singing experiences in real-world applications. We can observe the potential for improvement in speech-to-singing performance, indicating that there is ample room for enhancement in pitch modeling based on realistic music scores. Future work can explore using realistic music scores for better pitch modeling, as well as designing specialized intermediate models to better convert styles between speech and singing styles. For more detailed and visualized results about the speech-to-singing conversion, please refer to Appendix B.6.

5 Conclusion and Discussion

In this paper, we propose a novel dataset GTSinger, a large global, multi-technique, free-to-use, high-quality singing corpus with realistic music scores, designed for all singing tasks, comprehensively addressing the limitations of existing singing datasets. Furthermore, we provide the construction process and statistical analysis for GTSinger. In addition, we have conducted extensive experiments and established four benchmarks, thereby contributing further to future singing research.

Limitations and Future Directions: (1) Our dataset currently lacks comprehensive coverage of widely spoken languages, like Arabic, and does not include several commonly used singing techniques, such as vocal fry. Future efforts will be directed towards expanding the diversity of the singing data. (2) Although our annotation process is performed by professionals with musical expertise, accurately segmenting phoneme durations within words and identifying subtle singing techniques remains challenging for human ears. Future models that better utilize word-level annotations may mitigate some of the errors introduced by manual labeling. (3) While our dataset addresses various singing tasks, extending its utility to the broader music field may require integration with vocal-to-accompaniment models like SingSong [5] to assist in generating music that includes vocals.

Negative Societal Impact: The presence of sensitive biometric data in our dataset inherently carries potential risks. Therefore, We first perform data desensitization and consider using techniques such as vocal watermarking to further protect personal privacy.

Acknowledgements

References

[1] Agarwal, S., Ganapathy, S., and Takahashi, N. Leveraging symmetrical convolutional transformer networks for speech to singing voice style transfer, 2022.

[2] Boersma, P. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[3] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[4] Choi, S., Kim, W., Park, S., Yong, S., and Nam, J. Children's song dataset for singing voice research. In *International Society for Music Information Retrieval Conference (ISMIR)*, volume 4, 2020.

[5] Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., Agostinelli, A., Verzetti, M., Simon, I., Pietquin, O., Zeghidour, N., et al. Singsong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*, 2023.

[6] Duan, Z., Fang, H., Li, B., Sim, K. C., and Wang, Y. 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*, pp. 1–9. IEEE, 2013.

[7] He, J., Liu, J., Ye, Z., Huang, R., Cui, C., Liu, H., and Zhao, Z. Rmssinger: Realistic-music-score based singing voice synthesis. *arXiv preprint arXiv:2305.10686*, 2023.

[8] Huang, R., Chen, F., Ren, Y., Liu, J., Cui, C., and Zhao, Z. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3945–3954, 2021.

[9] Jayashankar, T., Wu, J., Sari, L., Kant, D., Manohar, V., and He, Q. Self-supervised representations for singing voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[10] Kim, K. L., Lee, J., Kum, S., Park, C. L., and Nam, J. Semantic tagging of singing voices in popular music recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1656–1668, 2020.

[11] Kim, S., Kim, Y., Jun, J., and Kim, I. Muse-svs: Multi-singer emotional singing voice synthesizer that controls emotional intensity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[12] Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[13] Li, R., Huang, R., Zhang, L., Liu, J., and Zhao, Z. Alignsts: Speech-to-singing conversion via cross-modal alignment. *arXiv preprint arXiv:2305.04476*, 2023.

[14] Li, R., Zhang, Y., Wang, Y., Hong, Z., Huang, R., and Zhao, Z. Robust singing voice transcription serves synthesis, 2024.

[15] Liu, J., Li, C., Ren, Y., Chen, F., and Zhao, Z. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11020–11028, 2022.

[16] Liu, J., Li, C., Ren, Y., Zhu, Z., and Zhao, Z. Learning the beauty in songs: Neural singing voice beautifier. *arXiv preprint arXiv:2202.13277*, 2022.

[17] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017.

[18] Ogawa, I. and Morise, M. Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs. *Acoustical Science and Technology*, 42 (3):140–145, 2021.

[19] Ren, Y., Tan, X., Qin, T., Luan, J., Zhao, Z., and Liu, T.-Y. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1979–1989, 2020.

[20] Sharma, B., Gao, X., Vijayan, K., Tian, X., and Li, H. Nhss: A speech and singing parallel database. *Speech Communication*, 133:9–22, 2021.

[21] Takahashi, N., Singh, M. K., and Mitsufuji, Y. Robust one-shot singing voice conversion. *arXiv preprint arXiv:2210.11096*, 2022.

[22] Tamaru, H., Takamichi, S., Tanji, N., and Saruwatari, H. Jvs-music: Japanese multispeaker singing-voice corpus. *arXiv preprint arXiv:2001.07044*, 2020.

[23] Wang, Y., Wang, X., Zhu, P., Wu, J., Li, H., Xue, H., Zhang, Y., Xie, L., and Bi, M. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022.

[24] Wei, H., Cao, X., Dan, T., and Chen, Y. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*, 2023.

[25] Wilkins, J., Seetharaman, P., Wahl, A., and Pardo, B. Vocalset: A singing voice dataset. In *ISMIR*, pp. 468–474, 2018.

[26] Zhang, L., Li, R., Wang, S., Deng, L., Liu, J., Ren, Y., He, J., Huang, R., Zhu, J., Chen, X., et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926, 2022.

[27] Zhang, Y., Cong, J., Xue, H., Xie, L., Zhu, P., and Bi, M. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7237–7241. IEEE, 2022.

[28] Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., Duan, X., Huai, B., and Zhao, Z. Stylesinger: Style transfer for out-of-domain singing voice synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19597–19605, 2024.

[29] Zhang, Z., Zheng, Y., Li, X., and Lu, L. Wesinger: Data-augmented singing voice synthesis with auxiliary losses. *arXiv preprint arXiv:2203.10750*, 2022.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Abstract and Section 1.

    (b) Did you describe the limitations of your work? [Yes] See Section 5.

    (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [N/A]

    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See GitHub page http://gtsinger.github.io.

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix B.

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report confidence intervals of subjective metric results in Section 4

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [N/A]

    (b) Did you mention the license of the assets? [Yes] See section 3.

    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See GitHub page http://gtsinger.github.io.

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3 and Appendix A.

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 5.

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix B.1.

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix A and B.1.

Appendices

GTSinger: A Global Multi-Technique Singing Corpus with Realistic Music Scores for All Singing Tasks

A   Details of Dataset

A.1   Details of Recording

We have recruited 20 professional singers, each proficient in at least one language, and signed formal contracts with them. They are hired at a rate of $300 per hour of audio recording to perform specified language skill songs. In total, we spend $30,000 on the recording process. Before recording, all singers agree to make their vocal performances open-source for academic research. All recordings are conducted in a professional studio, with singers listening to the song's accompaniment through headphones. This ensures clean vocal tracks without accompaniment while preserving rhythm and timing. Each audio is recorded at a 48kHz sampling rate with 24-bit depth in WAV format, ensuring high-quality data for further statistical analysis and research. The 20 singers cover all four vocal ranges: tenor, alto, bass, and soprano. They also perform in nine widely spoken languages: Chinese, English, Japanese, Korean, Russian, Spanish, French, German, and Italian. We require the singers to perform controlled comparison recordings using six singing techniques: mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando. Each technique covers controlled comparisons of multiple songs. For each song, they maintain consistent rhythm, lyrics, and key, recording twice: once densely applying the specific technique (technique group) and once using their natural singing voice without the specific technique (control group). We uniquely manage the falsetto and mixed voice techniques due to their strong correlation, necessitating special contrast for subsequent research. These techniques form a distinct group, recording a natural singing version (control group) and two versions for technique groups, for both falsetto and mixed voice. The falsetto version can raise the key of the same song to better showcase the falsetto technique while maintaining other consistencies like rhythm. Furthermore, each song includes an additional spoken lyric sentence recorded by the same singer, providing paired speech for speech-to-singing tasks.

A.2   Details of Annotation

We hire numerous experts with backgrounds in music and language for our annotation and review process, compensating each at a rate of $15 per hour. In total, we spend $18,000 on the annotation process. Before beginning their tasks, each expert is informed about the use of the annotated data, and they agree to make their annotation results open-source for academic research. Initially, some experts organize the submitted data from the singers into the required format to facilitate subsequent annotation. Then, we use the Montreal Forced Aligner (MFA) [17] for a coarse alignment of the original lyrics and audio, storing the results in TextGrid format. Chinese phonemes are extracted using pypinyin [3], English phonemes follow the ARPA standard [4], Italian phonemes follow the Epitran standard [5], while others follow the MFA standard [6]. We choose these standards because Chinese uses pinyin for pronunciation, ARPA includes English stress patterns, Epitran performs better in Italian phonemes, and the MFA dictionaries better capture the phonetic characteristics of other languages. Next, annotators use Praat [2] to correct the rough annotation results. Following the alignment process, we instruct our annotators to perform phoneme-level annotations of six singing techniques on the TextGrid, including mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando. Then, annotators also need to label the singing method (pop and bel canto), emotion (happy and sad), pace (slow, moderate, and fast), and range (low, medium, and high) as global style labels for each group. To compose realistic music scores, we initially employ RMVPE [24] to extract F0 and ROSVOT[14] to derive the MIDI form of the scores. Subsequently, we engage music experts to listen to the recorded songs, refer to original accompaniments, and annotate realistic music scores in the musicxml format.

---

[3] https://github.com/mozillazg/python-pinyin
[4] https://en.wikipedia.org/wiki/ARPABET
[5] https://github.com/dmort27/epitran
[6] https://mfa-models.readthedocs.io/en/latest/dictionary/

Each step is double-checked by other music experts. Finally, for each language data, we employ an additional music expert proficient in that language to randomly inspect 25% of the annotations.

A.3 Statistics of Global Styles

We annotate each group (each song comprising control and technique groups) with global style labels provided by music experts. These labels include singing method (pop and bel canto), emotion (happy and sad), pace (slow, moderate, and fast), and range (low, medium, and high). As shown in Figure 5, we have summarized the distribution of these four labels. It can be observed that GTSinger predominantly features songs with the pop singing method, reaching 74.37%. This is because we selected representative songs from various languages, which are mainly contemporary and widely popular. Characteristics of pop songs, such as a higher occurrence of fast pace, a predominant pitch range in the medium category, and a higher frequency of sad emotions, align with our statistical results. We can observe that fast pace accounts for 42.67%, medium range reaches 62.38%, and sad emotion reaches 78.22%. Additionally, the inclusion of diverse types of global styles demonstrates the stylistic variety present in GTSinger.
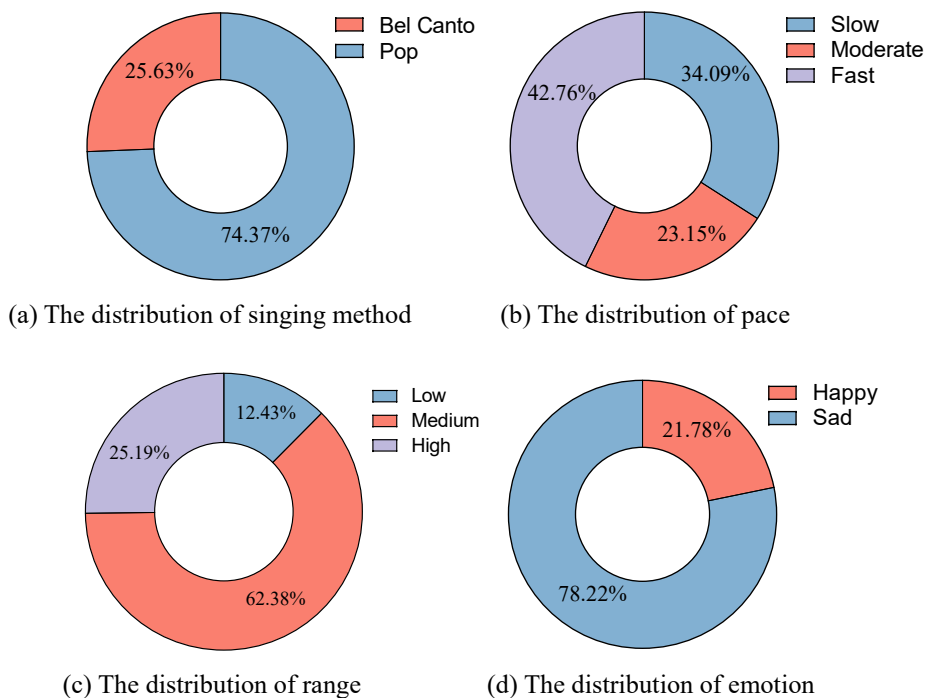

(a) The distribution of singing method   (b) The distribution of pace

(c) The distribution of range   (d) The distribution of emotion

Figure 5: The statistical distribution of global style labels.

A.4 Statistics of Chinese

There are two Chinese singers, namely ZH-Tenor-1 and ZH-Alto-1. As shown in Figure 6 (a), the overall BPM distribution ranges from 80 to 150, with a dense distribution between 110 and 150 due to the frequent use of their pop singing method, which is typically faster. As illustrated in Figure 6 (b), the two singers performed a variety of six techniques, with mixed voice being the most prevalent. This predominance is because mixed voice is the most commonly used technique. Vibrato is the least frequent because it can only be used on sustained notes, resulting in a very low distribution density. Figure 6 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 52 (E3, 164.81 Hz) to 69 (A4, 440 Hz). ZH-Tenor-1 primarily ranges from 52 (E3, 164.81 Hz) to 66 (F#4, 370 Hz), peaking at 59 (B3, 246.94 Hz), and ZH-Alto-1 primarily ranges from 55 (G3, 196 Hz) to 69 (A4, 440 Hz), peaking at 63 (D#4, 311.13 Hz). This distribution aligns with their vocal ranges. As mentioned above, Chinese phonemes are entirely annotated according to pypinyin , comprising a total of 56 phonemes. This annotation fully adapts to the pronunciation rules of Chinese and represents the most suitable set of Chinese phonemes, as used in previous

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

(d) The distribution of Chinese phonemes
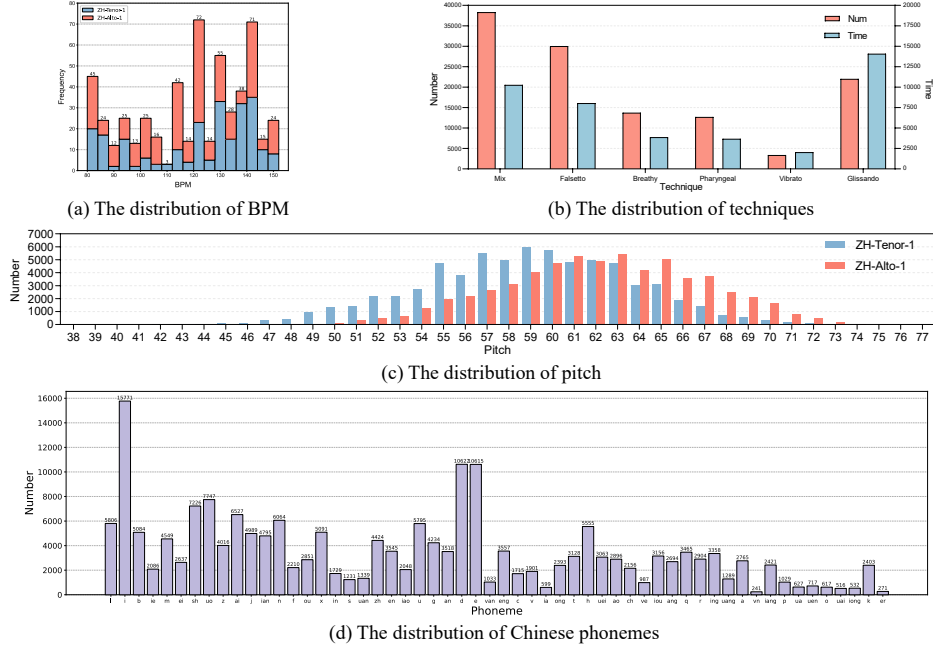
Figure 6: The statistical distribution of the BPM, techniques, pitch, and phonemes in Chinese.

Chinese singing datasets [26]. Figure 6 (d) illustrates that the most common phoneme is "i", with 15,771 occurrences, while the least common is "vn", with 241 occurrences. The broad distribution of phonemes is highly suitable for singing models.

A.5 Statistics of English

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

(d) The distribution of English phonemes

Figure 7: The statistical distribution of the BPM, techniques, pitch, and phonemes in English.

There are three English singers, namely EN-Tenor-1, EN-Alto-1, and EN-Alto-2. As shown in Figure 7 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 7 (b), the three

singers performed a variety of six techniques. Like Chinese, mixed voice is the most prevalent and vibrato is the least used technique. Figure 7 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 55 (G3, 196 Hz) to 69 (A4, 440 Hz). EN-Tenor-1 primarily ranges from 55 (G3, 196 Hz) to 68 (G#4, 415.30 Hz), peaking at 59 (B3, 246.94 Hz). EN-Alto-1 primarily ranges from 59 (B3, 246.94 Hz) to 68 (G#4, 415.30 Hz), peaking at 63 (D#4, 311.13 Hz). EN-Alto-2 primarily ranges from 57 (A3, 220 Hz) to 69 (A4, 440 Hz), peaking at 63 (D#4, 311.13 Hz). This distribution aligns with their vocal ranges. As mentioned above, English phonemes are entirely annotated according to the ARPA standard. The ARPA standard includes English stress patterns, therefore, it's appropriate for English phoneme modeling. Figure 7 (d) illustrates that the most common phoneme is "N", with 14,393 occurrences, while the least common phoneme is "AE0", with 1 occurrence. The missing phonemes are "AW0", "OY2", "AA0", and "OY0", which are barely seen in English songs, and the total usage frequency of all words containing them in daily life is less than 0.1%. The same applies to phonemes with an occurrence of less than 10.

A.6 Statistics of Japanese

(a) The distribution of BPM

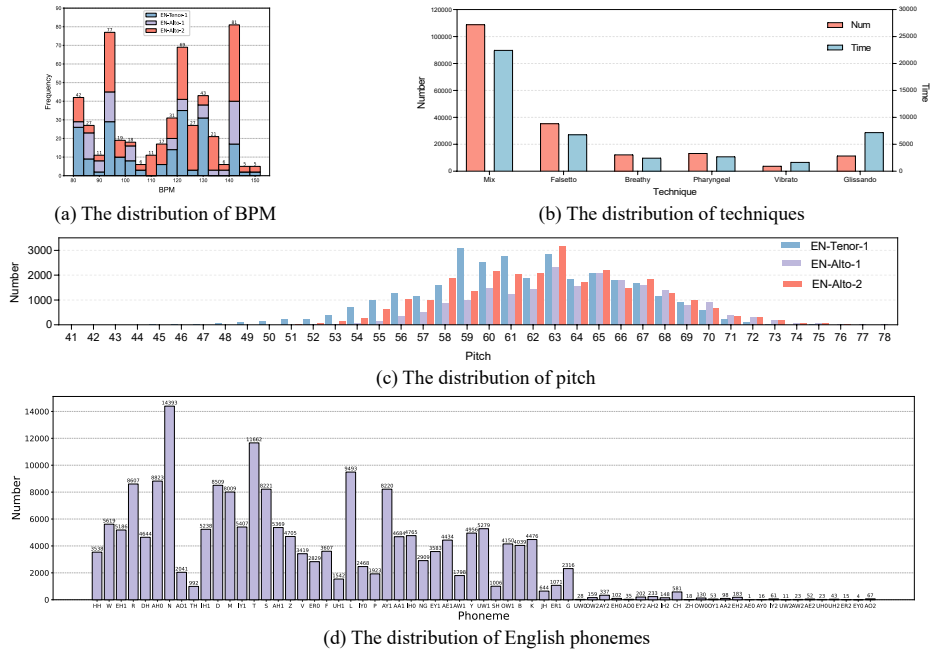(b) The distribution of techniques

(c) The distribution of pitch
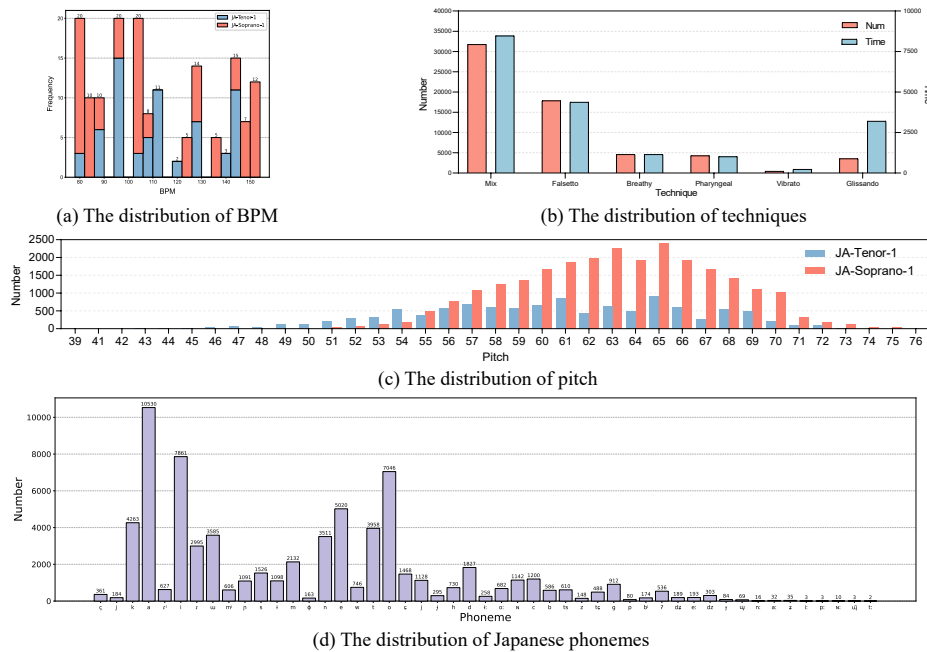
(d) The distribution of Japanese phonemes

Figure 8: The statistical distribution of the BPM, techniques, pitch, and phonemes in Japanese.

There are two Japanese singers, namely JA-Tenor-1 and JA-Soprano-1. As shown in Figure 8 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 8 (b), the two singers perform a variety of six techniques, with mixed voice being the most prevalent and vibrato being the least used technique. Figure 8 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 55 (G3, 196 Hz) to 70 (A#4, 466.16 Hz). JA-Tenor-1 primarily ranges from 54 (F#3, 185 Hz) to 69 (A4, 440 Hz), peaking at 65 (F4, 349.23 Hz). JA-Soprano-1 primarily ranges from 57 (A3, 220 Hz) to 70 (A#4, 466.16 Hz), peaking at 65 (F4, 349.23 Hz). Since the popular songs in Japanese typically have a relatively high range, we choose two singers whose vocal range is high. And this distribution aligns with their vocal ranges. As mentioned above, Japanese phonemes are entirely annotated according to the MFA standard. Figure 8 (d) illustrates that the most common phoneme is "a", with 10,530 occurrences, while the least common phoneme is "tː", with 2 occurrences. Phonemes with an occurrence of less than 10 are rarely found in Japanese songs. Additionally, the total usage frequency of all words containing these phonemes is less than 0.1%.

A.7 Statistics of Korean

There are three Korean singers, namely KO-Tenor-1, KO-Soprano-1, and KO-Soprano-2. As shown in Figure 9 (a), the overall BPM distribution ranges from 80 to 150, with a dense distribution between

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch
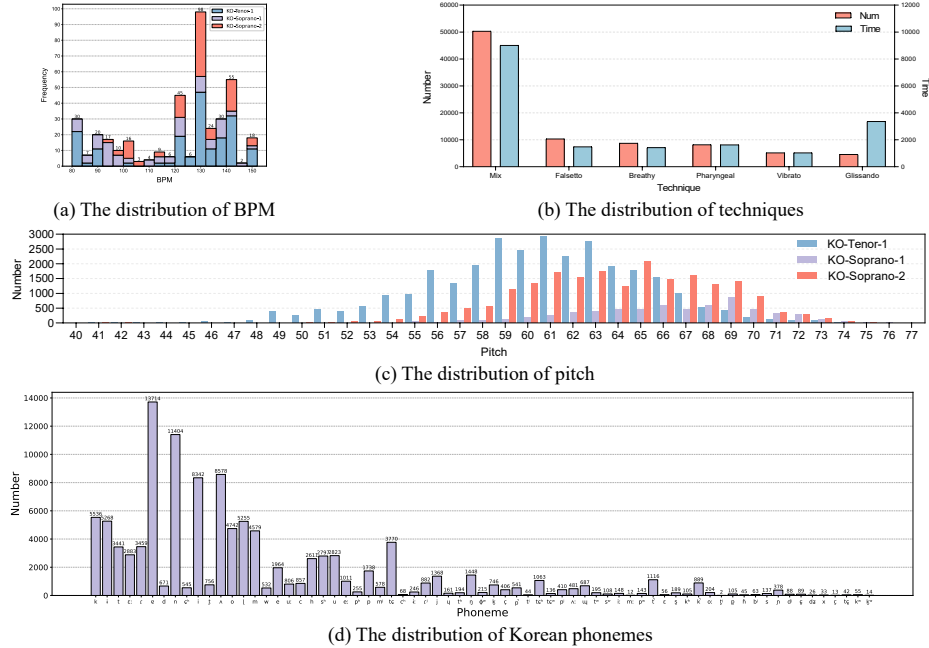
(d) The distribution of Korean phonemes

Figure 9: The statistical distribution of the BPM, techniques, pitch, and phonemes in Korean.

110 and 150 due to the frequent use of their popular singing styles, which are typically faster. As illustrated in Figure 9 (b), the three singers performed a variety of six techniques, with mixed voice being the most prevalent, glissando being the least used technique in number, and pharyngeal being the least used technique in time. Figure 9 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 55 (G3, 196 Hz) to 69 (A4, 440 Hz). KO-Tenor-1 primarily ranges from 53 (F3, 174.61 Hz) to 68 (G#4, 415.30 Hz), peaking at 61 (C#4, 277.18 Hz). KO-Soprano-1 primarily ranges from 61 (C#4, 277.18 Hz) to 72 (C5, 523.25 Hz), peaking at 69 (A4, 440 Hz). KO-Soprano-2 primarily ranges from 57 (A3, 220 Hz) to 70 (A#4, 466.16 Hz), peaking at 65 (F4, 349.23 Hz). Since the popular songs in Korean typically have a relatively high range, we choose three singers whose vocal range is high. And this distribution aligns with their vocal ranges. As mentioned above, Korean phonemes are entirely annotated according to the MFA standard. Figure 9 (d) illustrates that the most common phoneme is "ɐ", with 14,393 occurrences. Like Japanese, phonemes with an occurrence of less than 10 are rarely found in Korean songs, with the total usage frequency of all words containing these phonemes in daily life less than 0.1%.

A.8 Statistics of Russian

There is one Russian singer, namely RU-Alto-1. As shown in Figure 10 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 10 (b), the singer performs a variety of six techniques. Like the majority of technique distribution, mixed voice is the most prevalent and vibrato is the least used technique. Figure 10 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 57 (A3, 220 Hz) to 69 (A4, 440 Hz), peaking at 63 (D#4, 311.13 Hz). This distribution aligns with the singer's vocal range. As mentioned above, Russian phonemes are entirely annotated according to the MFA standard and shown in Figure 10 (d). Like other languages, the total usage frequency of all words containing phonemes with an occurrence of less than 10 in daily life is less than 0.1%.

A.9 Statistics of Spanish

There are two Spanish singers, namely ES-Bass-1 and ES-Soprano-1. As shown in Figure 11 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 11 (b), the two singers perform a variety of six techniques, with mixed voice being the most prevalent and vibrato being the least used technique. Figure 11 (c) shows that the overall note pitch ranges of realistic music

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

(d) The distribution of Russian phonemes

Figure 10: The statistical distribution of the BPM, techniques, pitch, and phonemes in Russian.

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

(d) The distribution of Spanish phonemes

Figure 11: The statistical distribution of the BPM, techniques, pitch, and phonemes in Spanish.

scores span from MIDI note number 49 (C#3, 138.59 Hz) to 69 (A4, 440 Hz). ES-Bass-1 primarily ranges from 49 (C#3, 138.59 Hz) to 65 (F4, 349.23 Hz), peaking at 54 (F#3, 185 Hz). ES-Soprano-1 primarily ranges from 56 (G#3, 207.65 Hz) to 69 (A4, 440 Hz), peaking at 61 (C#4, 277.18 Hz). This distribution aligns with their vocal ranges. As mentioned above, Spanish phonemes are entirely annotated according to the MFA standard. Figure 11 (d) illustrates that the most common phoneme is "e", with 18,798 occurrences, while the least common phonemes are "ɟ", with 12 occurrences. The missing phoneme is "ʃ", which is barely seen in Spanish songs, and the sum usage frequency of all words containing it in daily life is less than 0.1%.

A.10   Statistics of French

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

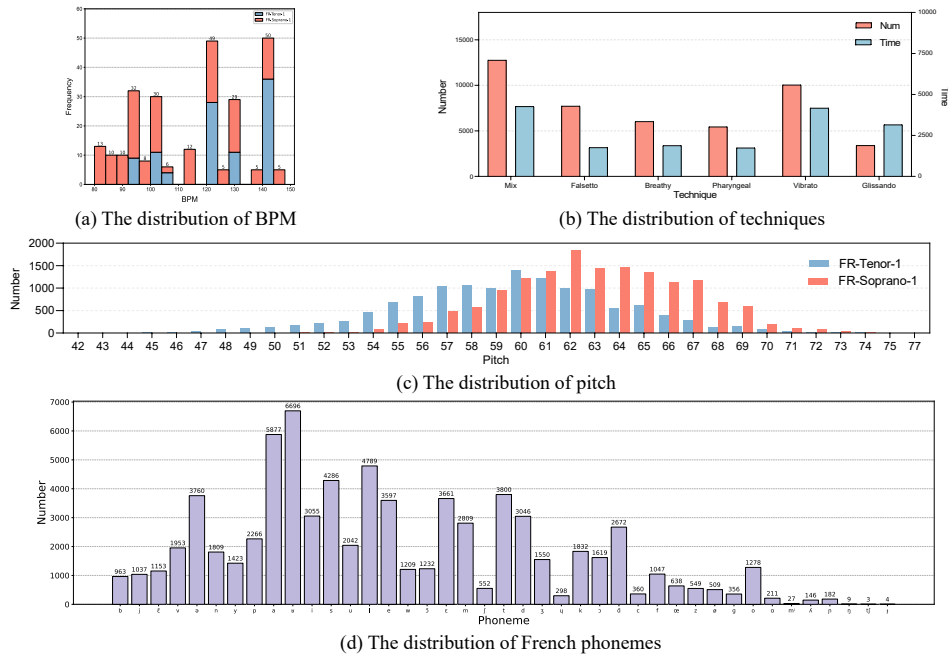(d) The distribution of French phonemes

Figure 12:   The statistical distribution of the BPM, techniques, pitch, and phonemes in French.

There are two French singers, namely FR-Tenor-1 and FR-Soprano-1. As shown in Figure 12 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 12 (b), the two singers perform a variety of six techniques, with mixed voice being the most prevalent, pharyngeal being the least used technique in number, and glissando being the least used technique in time. Figure 12 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 54 (F#3, 185 Hz) to 69 (A4, 440 Hz). FR-Tenor-1 primarily ranges from 54 (F#3, 185 Hz) to 65 (F4, 349.23 Hz), peaking at 60 (C4, 261.63 Hz). FR-Soprano-1 primarily ranges from 57 (A3, 220 Hz) to 69 (A4, 440 Hz), peaking at 62 (D4, 293.66 Hz). This distribution aligns with their vocal ranges. As mentioned above, French phonemes are entirely annotated according to the MFA standard. Figure 12 (d) illustrates that the most common phoneme is " ʁ", with 6,696 occurrences, while the least common phonemes are "tʃ" with 3 occurrences. Like other languages, the total usage frequency of all words containing phonemes with an occurrence of less than 10 in daily life is less than 0.1%.

A.11   Statistics of German

There are two German singers, namely DE-Tenor-1 and DE-Soprane-1. As shown in Figure 13 (a), the overall BPM distribution ranges from 80 to 150, which has a dense distribution between 80 and 110 due to the bel canto singing method for the singing, which is typically slow. As illustrated in Figure 13 (b), the three singers perform a variety of six techniques, with mixed voice being the most prevalent and vibrato being the least used technique. Figure 13 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 52 (E3, 164.81 Hz) to 71 (B4, 493.88 Hz). DE-Tenor-1 primarily ranges from 52 (E3, 164.81 Hz) to 65 (F4, 349.23 Hz), peaking at 59 (B3, 246.94 Hz). DE-Soprano-1 primarily ranges from 58 (A#3, 233.08 Hz) to 71 (B4, 493.88 Hz), peaking at 63 (D#4, 311.13 Hz). This distribution aligns with their vocal ranges. As mentioned above, German phonemes are entirely annotated according to the MFA standard. Figure 13 (d) illustrates that the most common phoneme is "n", with 12,284 occurrences, while the least common is "tʃ", with 6 occurrences. Like other languages, the total usage frequency of all words containing phonemes with an occurrence of less than 10 in daily life is less than 0.1%.
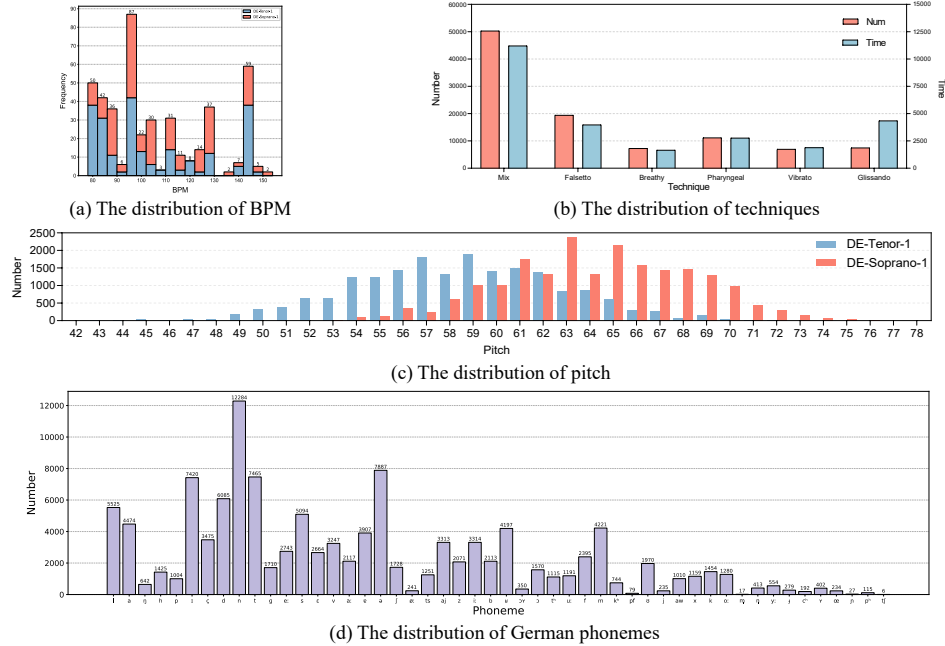
19

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch

(d) The distribution of German phonemes

Figure 13: The statistical distribution of the BPM, techniques, pitch, and phonemes in German.

(a) The distribution of BPM

(b) The distribution of techniques

(c) The distribution of pitch
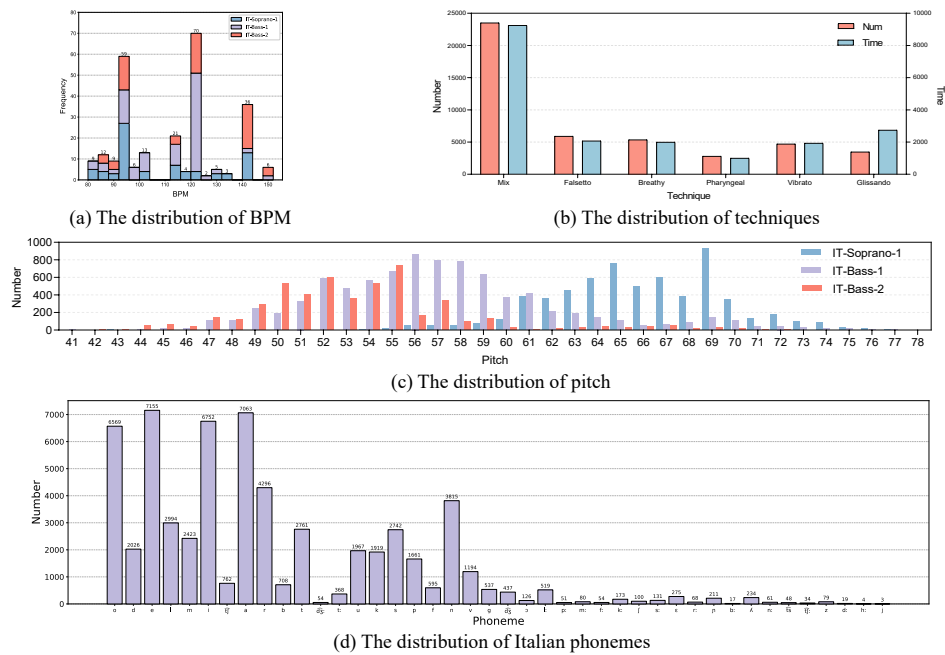

(d) The distribution of Italian phonemes

Figure 14: The statistical distribution of the BPM, techniques, pitch, and phonemes in Italian.

A.12 Statistics of Italian

There are three Italian singers, namely IT-Bass-1, IT-Bass-2, and IT-Soprano-1. As shown in Figure 14 (a), the overall BPM distribution ranges from 80 to 150. As illustrated in Figure 14 (b), the three singers perform a variety of six techniques, with mixed voice being the most prevalent and pharyngeal being the least used technique. Figure 14 (c) shows that the overall note pitch ranges of realistic music scores span from MIDI note number 49 (C#3, 138.59 Hz) to 70 (A#4, 466.16 Hz). IT-Bass-1 primarily ranges from 61 (C#4, 277.18 Hz) to 71 (B4, 493.88 Hz), peaking at 69 (A4, 440

20

Hz). IT-Bass-2 primarily ranges from 49 (C#3, 138.59 Hz) to 63 (D#4, 311.13 Hz), peaking at 56 (G#3, 207.65 Hz). IT-Soprano-1 primarily ranges from 49 (C#3, 138.59 Hz) to 59 (B3, 246.94 Hz), peaking at 55 (G3, 196 Hz). This distribution aligns with their vocal ranges. As mentioned above, Italian phonemes are entirely annotated according to the Epitran standard, which performs best for Italian phoneme alignment. Figure 14 (d) illustrates that the most common phoneme is "e", with 7,155 occurrences, while the least common phonemes are "j" and "h:", with 3 and 4 occurrences, respectively. Like other languages, the total usage frequency of all words containing phonemes with an occurrence of less than 10 in daily life is less than 0.1%.

B Details of Experiments

B.1 Subjective Evaluation

For each task, we randomly select 50 sentences from our test set for subjective evaluation, each of which is listened to by at least 20 professional listeners. To evaluate the model performance, we conduct the MOS (Mean opinion score) evaluation. In the context of MOS-Q evaluations, these listeners are instructed to concentrate on synthesis quality (including clarity, naturalness, and rich stylistic details), irrespective of singer similarity (in terms of timbre and styles). Conversely, during MOS-S evaluations, the listeners are directed to assess singer similarity (singer similarity in terms of timbre and styles) to the audio reference, disregarding any differences in content or synthesis quality (including quality, clarity, naturalness, and rich stylistic details). For MOS-C, the listeners are informed to evaluate technique controllability (accuracy and expressiveness of technique control), disregarding any differences in content, timbre, or synthesis quality (including quality, clarity, naturalness, and rich stylistic details). In both MOS-Q, MOS-S, and MOS-C evaluations, listeners are requested to grade various singing voice samples on a Likert scale ranging from 1 to 5.

The screenshots of instructions for testers are shown in Figure 15. It is important to note that all participants are fairly compensated for their time and effort. We compensated participants at a rate of $15 per hour, resulting in a total expenditure of approximately $2000 on participant compensation. Participants are informed that the results will be used for scientific research.

B.2 Objective Evaluation

For our evaluation, we select various objective metrics tailored to different tasks. First, we use F0 Frame Error (FFE), which combines voicing decision error and F0 error metrics to capture F0 information comprehensively. Next, we employ Mean Cepstral Distortion (MCD) for measuring audio quality as the formula:

$$\text{MCD} = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (c_t(d) - \hat{c}_t(d))^2}, \tag{1}$$

where $c_t(d)$ and $\hat{c}_t(d)$ represent the d-th MFCC of the target and predicted frames at time t, respectively, and D is the number of MFCC dimensions. Additionally, Cosine Similarity (Cos) is utilized to quantify the resemblance between the synthesized and reference singing voices. We calculate the average cosine similarity between the embeddings extracted from the synthesized voices and the ground truth. We use the WavLM [3] fine-tuned for speaker verification [7] to extract singer embedding.

B.3 Technique-Controllable Singing Voice Synthesis

For each model, we incorporated a simple technique embedding. This module takes as input six binary sequences indicating the presence or absence of each technique on each phoneme. These sequences are then encoded and concatenated. In the non-parallel experiments, we randomly generate technique sequences and select the appropriate ones to input into the model, assigning zero, one, or multiple techniques to each target phoneme.

We used a sequence incorporating mixed voice and vibrato techniques for technique control. As shown in Figure 16, DiffSinger failed to control vibrato, whereas StyleSinger enhanced the expressiveness of mixed voice through detailed styles.

---

[7]https://huggingface.co/microsoft/wavlm-base-plus-sv

(a) Screenshot of MOS-Q and MOS-C testing of technique-controllable SVS

(b) Screenshot of MOS-Q and MOS-S testing of style transfer

(c) Screenshot of MOS-Q and MOS-S testing of STS conversion

Figure 15: The statistical distribution of screenshots of MOS testings for diverse tasks.

B.4 Technique Recognition

We design a technique recognition model based on ROSVOT [14]. This model is originally designed for coarse notes of music score detection. Simply, we change the loss to the cross entropy loss between the GT and predicted technique labels. The inputs of the technique recognition model include mel-spectrogram, pitch, and phoneme boundaries, with the output being the predicted probabilities of six distinct techniques.

As shown in Table 7, our model achieves high Precision, Recall, F1, and Accuracy across all six techniques, both in overall and cross-lingual experiments. This performance also underscores the quality of controlled comparison and phoneme-level technique annotations in GTSinger.

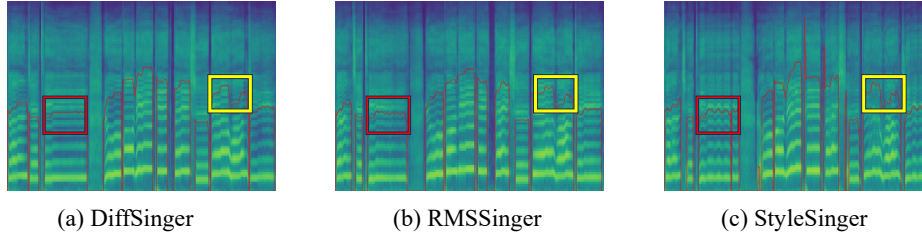|		
(a) DiffSinger	(b) RMSSinger	(c) StyleSinger

Figure 16: The mel-spectrograms depict the results of technique-controllable SVS. The vibrato technique is indicated by red boxes, and yellow boxes show the mixed voice technique.

Table 7: Precision, Recall, F1, and Accuracy of each technique in both overall and cross-lingual technique recognition. We provide both Asian-to-European and European-to-Asian results.

Experiment	Metric	Technique Recognition Accuracy					
		mixed voice	falsetto	breathy	pharyngeal	vibrato	glissando
Overall	Precision	0.95	0.98	0.99	0.96	0.75	0.75
	Recall	0.71	0.95	0.99	0.82	0.72	0.71
	F1	0.78	0.96	0.99	0.85	0.70	0.70
	Accuracy	0.78	0.84	0.78	0.80	0.89	0.85
Asian-to-European	Precision	0.92	0.97	0.99	0.97	0.62	0.63
	Recall	0.55	0.93	0.98	0.79	0.68	0.63
	F1	0.65	0.94	0.98	0.84	0.58	0.51
	Accuracy	0.74	0.78	0.71	0.78	0.87	0.76
European-to-Asian	Precision	0.89	0.95	0.93	0.87	0.72	0.61
	Recall	0.64	0.90	0.95	0.77	0.61	0.68
	F1	0.71	0.92	0.93	0.79	0.57	0.58
	Accuracy	0.76	0.8	0.73	0.76	0.81	0.81

B.5   Style Transfer

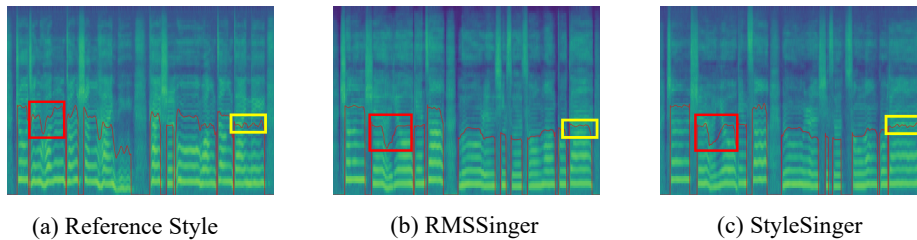|		
(a) Reference Style	(b) RMSSinger	(c) StyleSinger

Figure 17: The mel-spectrograms depict the results of style transfer. The vibrato style is indicated by yellow boxes, and the pronunciation and articulation skills are highlighted in red boxes.

We proceed to visualize mel-spectrograms and pitch contours of style transfer experiments in Figure 17. StyleSinger excels at capturing the intricate nuances of the reference style. The pitch curve generated by StyleSinger exhibits a greater range of variations and finer details, effectively capturing the vibrato technique, as well as the nuances of pronunciation and articulation skills. In contrast, the curves generated by RMSSinger appear relatively flat. Additionally, StyleSinger excels in modeling mel-spectrograms with higher quality and stylistic details.

B.6   Speech-to-Singing Conversion

We modify StyleSinger to use the paired speech reference input for generating the singing voice, enabling it to perform STS conversion and fully model the style transfer between speech and singing voice. While AlignSTS requires the input of the singing voice's f0 for conversion, StyleSinger only needs realistic music scores, making it more suitable for practical applications.

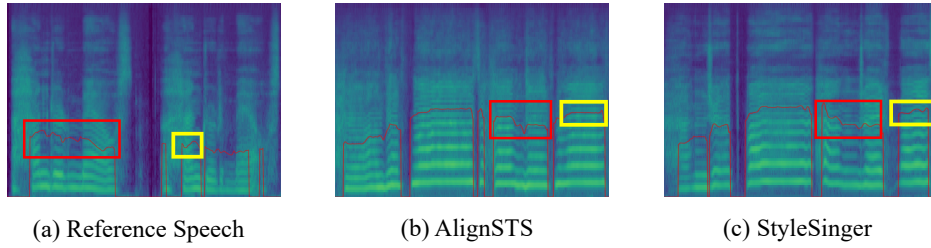| (a) Reference Speech | (b) AlignSTS | (c) StyleSinger |

Figure 18: The mel-spectrograms depict the results of STS conversion. The pronunciation and articulation skills are highlighted in yellow boxes, and the pitch transitions are shown in red boxes.

We visualize mel-spectrograms and pitch contours of STS experiments in Figure 18. We observe that StyleSinger successfully transfers the pronunciation and articulation skills, as well as the pitch transition styles. Compared to AlignSTS, which exhibits flat pitch with a lack of style variation and mel-spectrograms lacking in detail, StyleSinger demonstrates more pitch variations that closely match the reference speech style, along with higher quality mel-spectrograms.