



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

NeurIPS 2024 DB Track Spotlight

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

CONTENTS



01

Motivation

02

Dataset

03

Benchmarks

Motivation

➤ Background

- 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. These tasks have been progressively developed and applied in real life, like short videos and professional composition. However, 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.

Motivation

- The limitations of current open-source singing datasets
 - The low quality may lead to singing models producing off-pitch, unpleasant, or noisy results.
 - A limited variety in languages and singers restricts personalized singing models to learn diverse timbres and styles.
 - The absence of the controlled comparison and annotations for multiple singing techniques (like falsetto), constrains the technique modeling and control for singing models.
 - The lack of realistic music scores hinders human composers from using singing models in real-world musical composition.
 - Poor task suitability forces multiple emerging singing tasks to customize new datasets with high cost.

Motivation

- The limitations of current 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	✓	✓	✓	✓	✓

Motivation

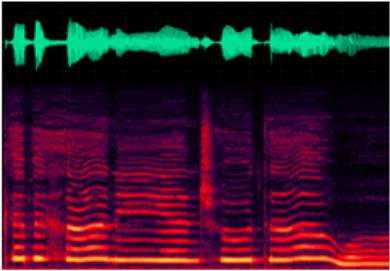
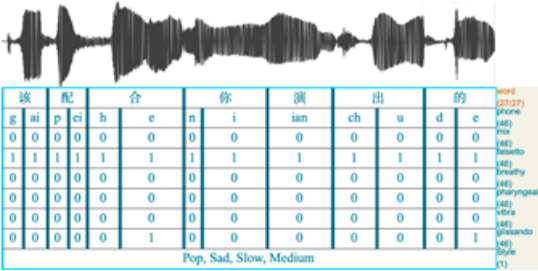
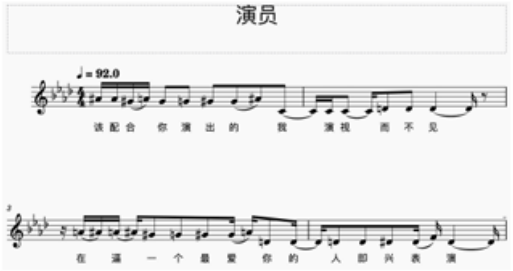
➤ The advantages of GTSinger

- **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** (Chinese, English, Japanese, Korean, Russian, Spanish, French, German, and Italian) 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** (mixed voice, falsetto, breathy, pharyngeal, vibrato, and glissando) 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** (singing method, emotion, range, and pace), and **16.16 hours of paired speech**, ensuring comprehensive annotations and broad task suitability.

Motivation

➤ The advantages of GTSinger

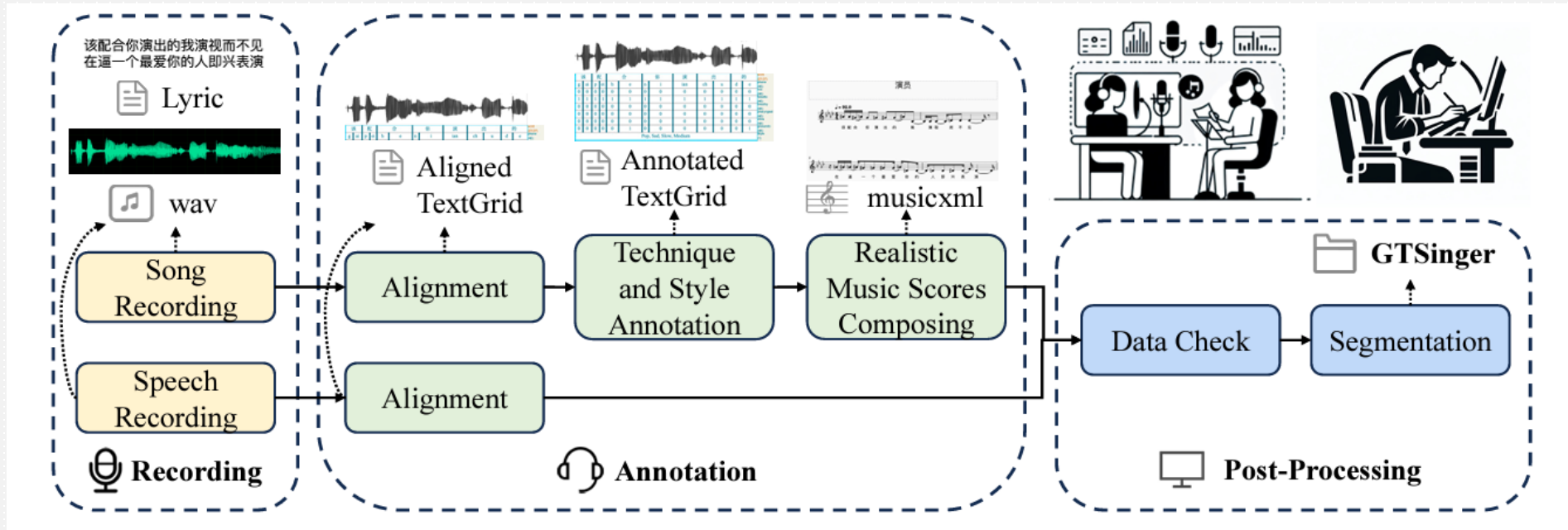
- Each song contains a technique group and a control group for the controlled comparison, along with a paired speech. Alignments, annotations, and realistic music scores are manually created for each group.

Language ID ▶ Singer ID ▶ Technique ID ▶ Song ID			
Controlled Comparison	 <p>wav</p>	 <p>TextGrid</p>	 <p>musicxml</p>
Technique Group	Singing Voice with Dense Specific Technique	Alignment, Style Label, Technique Annotation	Realistic Music Score
Control Group	Natural Singing Voice without Specific Technique	Alignment, Style Label, Technique Annotation	Realistic Music Score
Paired Speech	Paired Speech under the Same Lyrics	Alignment	

Dataset

➤ Pipeline

- Our GTSinger pipeline consists of three main stages: recording, annotation, and post-processing. Human double-checks exist in each process.



Dataset

➤ Songs and Singers

- To construct GTSinger, we first select nine widely spoken languages and six commonly used singing techniques. After rigorous auditions, we select 20 professional singers, covering all four vocal ranges (alto, soprano, tenor, bass). 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.

Dataset

➤ Songs and Singers

- 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

Dataset

➤ 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 densely applying the specific technique (technique group) and once for the natural singing voice without the specific technique (control group).
- 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.

Dataset

➤ Alignment

- We initially use the Montreal Forced Aligner (MFA) for a coarse alignment of the original lyrics and audio and store the results in TextGrid format.
- Next, annotators with a musical background use 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.
 - (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.

Dataset

➤ 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 for annotations based on their auditory perception, indicating the presence or absence of each technique for every phoneme.
- 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.

Dataset

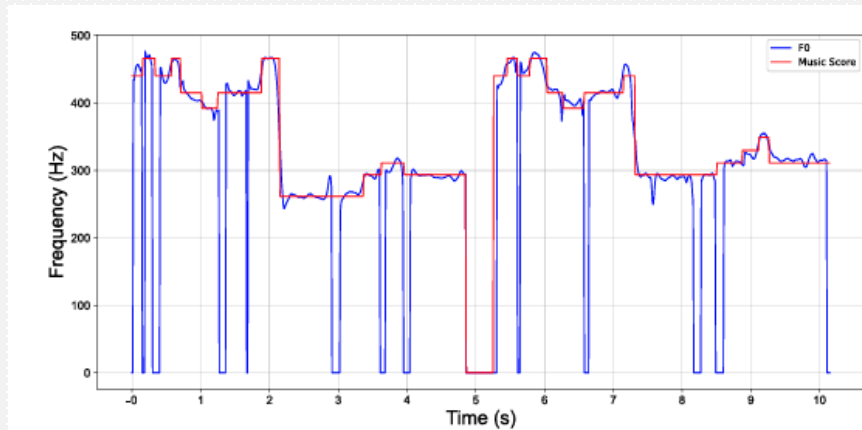
➤ Realistic Music Score Composing

- To compose realistic music scores, we initially employ RMVPE to extract F0 for each singing voice.
- Then, we use ROSVOT to derive the MIDI form of the music scores.
- 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.

Dataset

➤ Realistic Music Score Composing

- 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.



(b) Fine-gained music scores



Dataset

➤ 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.
- (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.

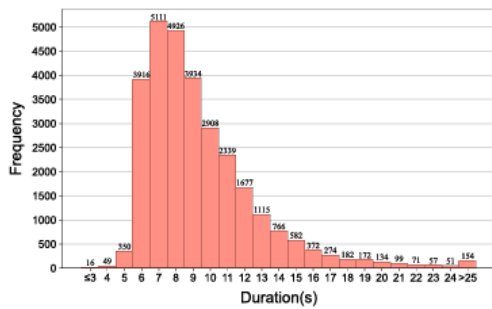
Dataset

➤ Segmentation

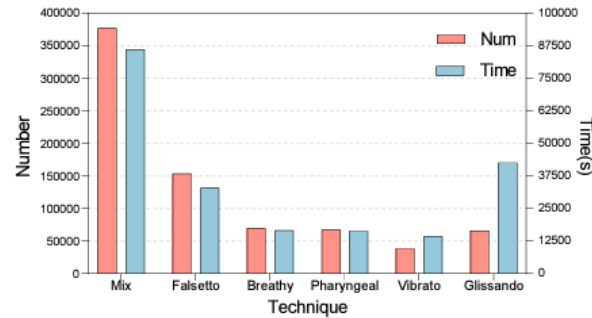
- After completing the data annotation and inspection, we segment the audio into smaller fragments to facilitate training for singing tasks.
- For the same song, the control group, technique group, and paired speech are synchronously segmented into sentence-level segments, with their alignments, annotations, 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.
- 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.

Dataset

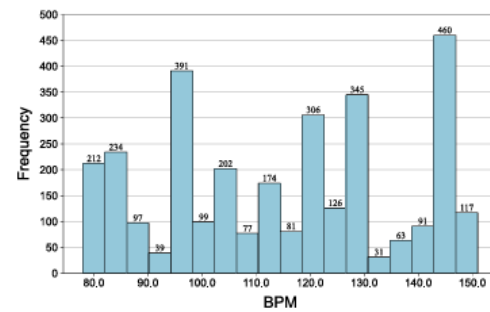
➤ Statistics



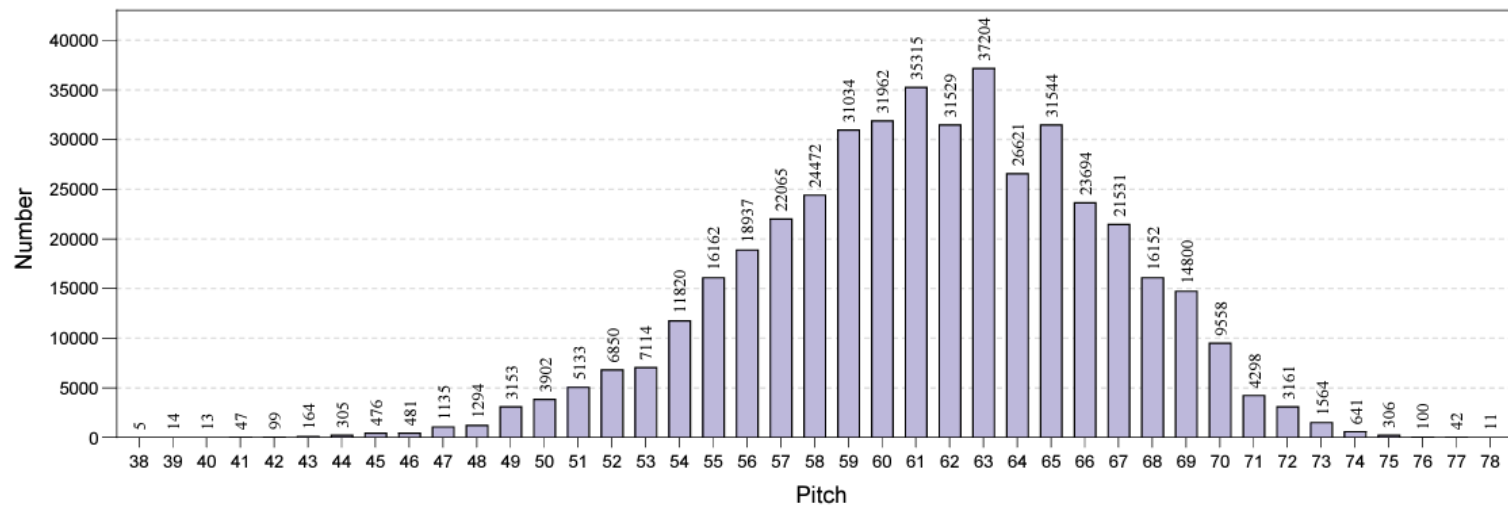
(a) The distribution of segment duration



(b) The distribution of techniques



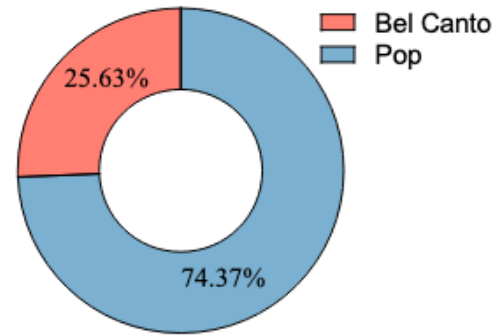
(c) The distribution of BPM



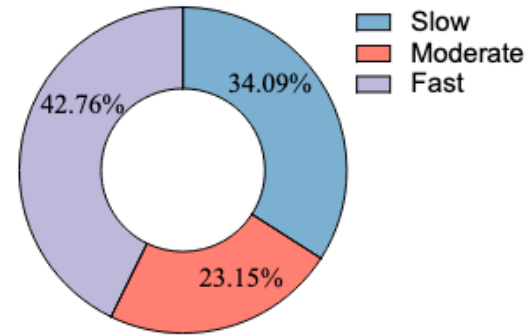
(d) The distribution of pitch

Dataset

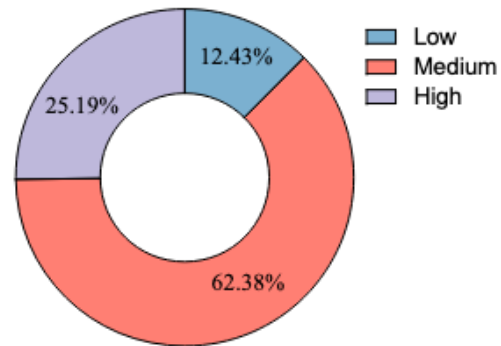
➤ Statistics



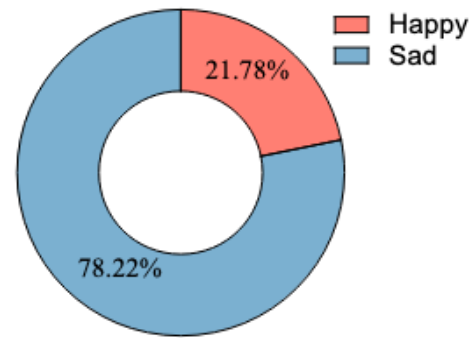
(a) The distribution of singing method



(b) The distribution of pace



(c) The distribution of range



(d) The distribution of emotion

Benchmarks

➤ Technique-controllable SVS

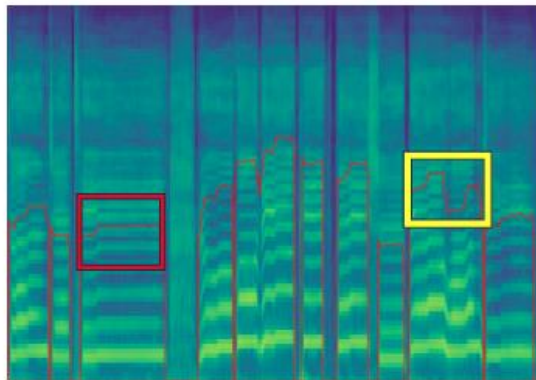
- Technique-controllable SVS in both parallel and non-parallel experiments. We use FFE, MCD, MOS-Q, and MOS-C.

Method	Parallel				Non-Parallel	
	FFE ↓	MCD ↓	MOS-Q ↑	MOS-C ↑	MOS-Q ↑	MOS-C ↑
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

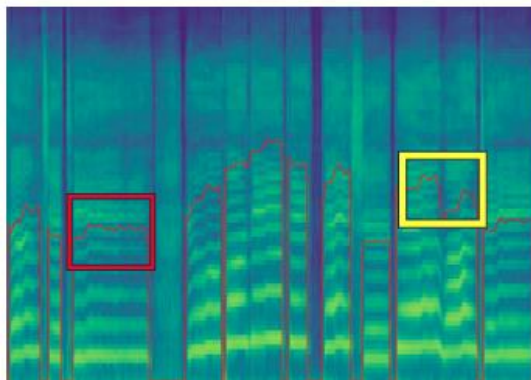
Benchmarks

➤ Technique-controllable SVS

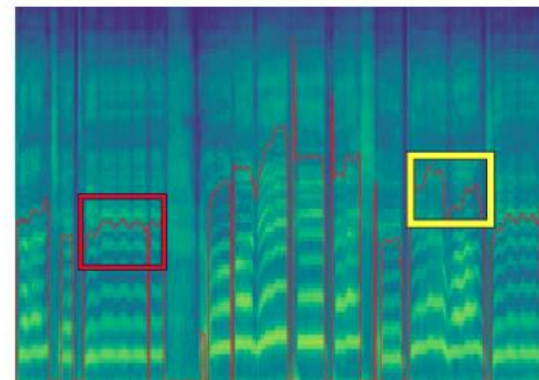
- The mel-spectrograms depict the results of technique-controllable SVS. The vibrato technique is indicated by red boxes, and yellow boxes show the mixed voice technique.



(a) DiffSinger



(b) RMSSinger



(c) StyleSinger

Benchmarks

➤ Technique Recognition

- 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	falsestto	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

Benchmarks

➤ Style Transfer

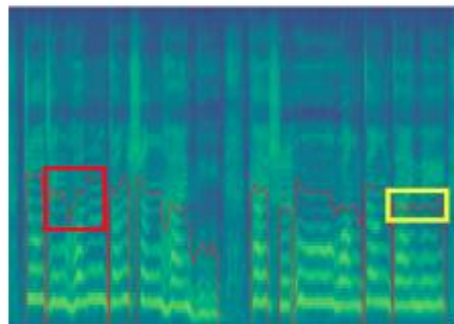
- Style Transfer performance in both parallel and cross-lingual experiments. We use FFE, MCD, Cos, MOS-Q, and MOS-C for comparisons.

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

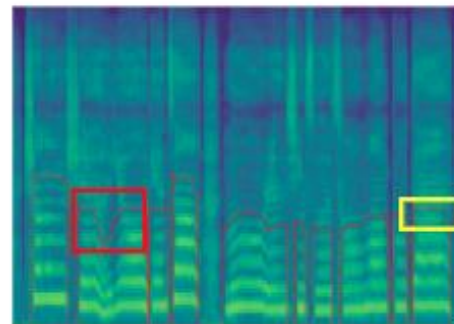
Benchmarks

➤ Style Transfer

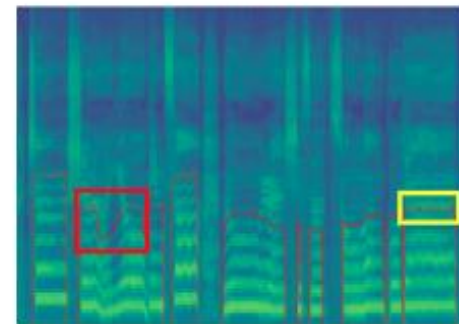
- The mel-spectrograms depict the results of style transfer. The vibrato style is indicated by yellow boxes, and the pronunciation and articulation skills are highlighted in red boxes.



(a) Reference Style



(b) RMSSinger



(c) StyleSinger

Benchmarks

➤ Speech-to-Singing Conversion

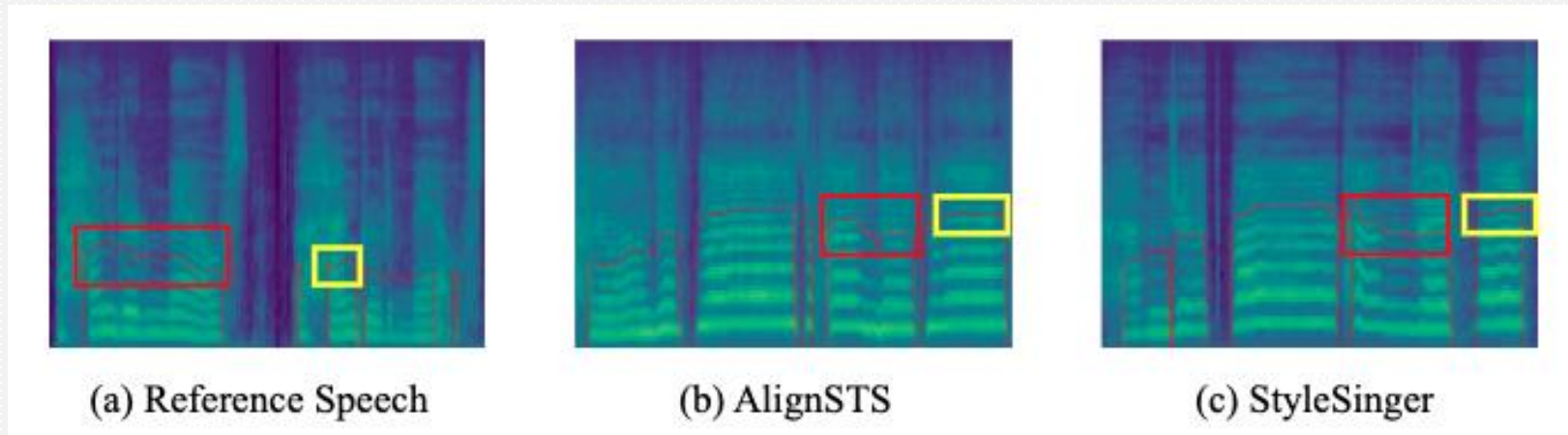
- Speech-to-singing performance in FFE, MCD, Cos, MOS-Q, and MOS-S metrics.

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

Benchmarks

➤ Speech-to-Singing Conversion

- 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.





THANKS