SINGSTYLE111: A MULTILINGUAL SINGING DATASET WITH STYLE TRANSFER

Shuqi Dai[1] **Yuxuan Wu**[1] **Siqi Chen**[2] **Roy Huang**[1] **Roger B. Dannenberg**[1]

[1] Computer Science Department, Carnegie Mellon University , USA
[2] University of Southern California, USA

shuqid@cs.cmu.edu, rbd@cs.cmu.edu

ABSTRACT

There has been a persistent lack of publicly accessible data in singing voice research, particularly concerning the diversity of languages and performance styles. In this paper, we introduce SingStyle111, a large studio-quality singing dataset with multiple languages and different singing styles, and present singing style transfer examples. The dataset features 111 songs performed by eight professional singers, spanning 12.8 hours and covering English, Chinese, and Italian. SingStyle111 incorporates different singing styles, such as bel canto opera, Chinese folk singing, pop, jazz, and children. Specifically, 80 songs include at least two distinct singing styles performed by the same singer. All recordings were conducted in professional studios, yielding clean, dry vocal tracks in mono format with a 44.1 kHz sample rate. We have segmented the singing voices into phrases, providing lyrics, performance MIDI, and scores with phoneme-level alignment. We also extracted acoustic features such as Mel-Spectrogram, F0 contour, and loudness curves. This dataset applies to various MIR tasks such as Singing Voice Synthesis, Singing Voice Conversion, Singing Transcription, Score Following, and Lyrics Detection. It is also designed for Singing Style Transfer, including both performance and voice timbre style. We make the dataset freely available for research purposes. Examples and download information can be found at `https://shuqid.net/singstyle111`.

1. INTRODUCTION

In recent years, deep learning technologies have significantly advanced the field of Artificial Intelligence Generative Content (AIGC) [1], leading to breakthroughs in Computer Vision for image synthesis and manipulation [2–5], Natural Language Processing (NLP) for text generation and summarization [6–8], and audio signal processing for Text-to-Speech (TTS) generation [9–11]. In particular, advanced generative models such as Variational Autoencoders (VAEs) [12–14], Generative Adversarial Networks (GANs) [15, 16], Transformer-based models [17, 18], and Diffusion Models [19, 20] resulted in a series of exceptional TTS models that achieve not only realistic results [9–11, 21] but also explore stylistic and emotional speech synthesis [22, 23] in a more controllable way. However, the development of singing tasks such as Singing Voice Synthesis (SVS) [24–28] and Singing Voice Conversion (SVC) [29] have yet to progress as fast as TTS. One primary reason is the lack of data on several key aspects:

- Lack of high-quality data. Tasks such as SVS and SVC require monophonic, clean, and dry sound singing data with studio quality. Unfortunately, due to the limitations of Source Separation and Denoising technologies [30–33], as well as copyright issues, most available cover songs online cannot meet these quality requirements. Datasets recorded with studio quality are predominantly composed of amateur performances, which often exhibit off-key and cracking issues that could mislead the generative models and diminish their quality.
- Lack of diversity. Most available singing datasets cover only one language, resulting in a severely imbalanced language distribution. For example, there is a fair amount of Chinese singing data, while clean English data is very scarce. In addition, most datasets only focus on one pop singing style, and the distributions of different singing styles and vocal ranges are too narrow.
- Lack of annotations. Many datasets lack proper phrase-level segmentation, lyrics, and scores, and are not aligned at the phoneme level, making it impossible to conduct score-based SVS and more detailed performance control.
- Lack of large-scale data. The current data volume of high-quality singing is still insufficient for deep generative models.

Furthermore, current SVS results are primarily confined to modeling the timbre of singing voices. While there are several good vocoders [11, 21, 34] and acoustic models [10, 35] for SVS based on Ground-Truth control signals (e.g., inputting F0 control signals to the model), the truly creative and artistic aspects of singing, such as expressive performance control, singing styles, vocal techniques, and creative improvisation, have yet to be explored. Again, data limitations play a significant role in this, as most datasets consist of amateur performances or have not

Dataset	Language	Style	#Hour	#Singer	Quality	Musicality	Score	Align-ment	Style Transfer
Opencpop [41]	Chinese	Pop	5.25	1	Studio	Ama.	Perform. MIDI	✓	✗
M4Singer [42]	Chinese	Pop	29.77	20	Studio	50% Ama. 50% Prof.	Perform. MIDI	✓	✗
Children Song [43]	Korean English	Children	4.86	1	Studio	Prof. but plain	Perform. MIDI	word	✗
Tohoku Kiritan [44]	Japanese	Pop	0.95	1	Studio	Prof.	Score	✓	✗
PopCS [28]	Chinese	Pop	5.89	6	Not Clean	Ama.	✗	✗	✗
Open-Singer [35]	Chinese	Pop	50	66	Studio	Ama.	✗	✗	✗
VocalSet [45] Annotated [46]	Five Vowels	Opera	10.1	20	Studio	Prof.	Score	✓	technique transfer
NHSS [47]	English	Pop	3.5	10	Studio	Ama.	✗	✓	✗
NUS-48E [48]	English	Pop Children	1.41	12	Studio	Ama.	✗	✓	✗
RWC [49]	Japanese English	Pop	4	27	Not Solo	Prof.	Both	✗	✗
TONAS [50]	Spanish	Flamenco	0.34	> 40	Not Clean	Prof.	✗	✗	✗
Vocadito [51]	Seven Languages	Pop Children	0.23	29	Not Clean	Ama.	✗	✗	✗
MIR-1K [52]	Chinese	Pop	2.22	19	Not Solo	Ama.	✗	✗	✗
StyleSing111 (Ours)	English Chinese Italian	Opera Pop Folk Jazz etc.	12.8	8	Studio	Prof.	Both	✓	✓

Table 1. A comparison of existing singing datasets. Score means if there is score or performance MIDI file provided. "Perform. MIDI" stands for "Performance MIDI". "Both" means both performance MIDI files synchronized with the singing audio and sheet music scores are provided. Alignment means whether or not there is duration annotation at the phoneme level for lyrics. "Ama." stands for "Amateur," and "Prof." stands for "Professional."

yet begun to address the issue of artistic expression.

For example, Style Transfer [36, 37] is a popular technique in deep learning that combines the content of one image or sound with the style of another. For audio processing, some researchers [38, 39] have recently transferred the timbre from one audio source to another while preserving the speech content (similar to SVC). However, the transfer of expressive performance styles embedded below the timbre level remains elusive, mainly because (1) disentangling performance style is much more challenging than timbre features [40] and (2) the scarcity of relevant datasets providing examples of performance styles.

To help address these issues, we introduce a new singing corpus, SingStyle111. We summarize the main contributions as follows:

(1) SingStyle111 is a large and high-quality singing dataset. It contains 111 songs performed by eight professional singers, spanning 12.8 hours of clean monophonic vocal recordings in studio quality.
(2) It is a diverse dataset with creative singing. It covers English, Chinese, and Italian songs and incorporates various singing styles, such as bel canto opera, Chinese folk, pop, jazz, and children. Some performances are creative improvisations based on the original score.
(3) It demonstrates style transfer in both performance and timbre levels. 80 songs contain at least two distinct singing styles performed by the same singer.
(4) It includes proper annotations and extracted features. We manually segmented voices into phrases, labeled Performance MIDI files and music score notes and aligned them with the phonemes of lyrics, extracted acoustic features such as Mel-Spectrogram, F0 contour, and loudness curves.
(5) It applies to different MIR tasks such as SVS, SVC, Singing Transcription, Score Following, Expressive Performance, Lyrics Detection, Singing Style Transfer.
(6) It is publicly available for research purposes for free.

The rest of this paper is organized as follows: after a brief review of related works, we describe how we collect and process the dataset (Section 3) and show the annotations and analysis (Section 4). Finally, we discuss potential applications in Section 5 followed by conclusions.

2. RELATED WORK

Existing singing voice datasets still have many limitations in fulfilling the requirements for singing research tasks such as Singing Voice Synthesis (SVS) [24, 26–28] and Singing Voice Conversion (SVC) [29]. Table 1 provides an overview of the available public datasets. Datasets such as MIR-1K [52], TONAS [50], and Vocadito [51] are restricted by the absence of separated solo vocal tracks or suffer from subpar recording environments with noise, reverberation, and other interferences. These issues hinder their usability in SVS-related tasks. While NHSS [47] and OpenSinger [35] contain clean and dry human vocals, they lack essential musical scores or phoneme-level duration alignment. Consequently, these datasets are unsuitable for training end-to-end synthesis models that convert scores to vocals. Moreover, datasets such as Opencpop [41] and M4singer [42] offer good annotations and recording quality but primarily focus on Mandarin songs and a limited range of pop styles. Additionally, the singing proficiency of performers is inconsistent, with many being amateurs, which affects the overall quality of the dataset.

Another issue that has long been overlooked and misunderstood in singing voice datasets is the difference between Performance MIDI and the actual sheet music score. In Table 1, only Tohoku Kiritan [44], Vocalset [45, 46] and RWC [49] have music scores, while other datasets claimed to have scores that are indeed performance MIDI files. Performance MIDI features expressive performance timings rather than score timings with regular note durations in beats. The melodic pitches in performance MIDI can also differ from those in score melody. Utilizing performance MIDI for singing voice synthesis and claiming it as score-based is, in reality, a deceptive approach that takes advantage of real singing data.

As for the Style Transfer task, Vocalset [45, 46] provides relevant examples, but its scope is limited to singing technique transfer within the bel canto singing style. Furthermore, the dataset predominantly consists of scale exercises using only five vowels and includes only three short songs, which restricts its applicability. Given the limitations of existing datasets, there is a need for a large-scale, high-quality, professional, multilingual, and diverse singing dataset that caters to various styles and includes style transfer examples. In this paper, we introduce a novel dataset designed to address these requirements and facilitate research in SVS-related tasks and style transfer.

3. DATASET DESCRIPTION

3.1 Overview

SingStyle111 is a multilingual singing dataset with style transfer demonstrations. Figure 1 illustrates the data collection pipeline. Following the completion of the recording process, we post-process all recordings and retain all high-quality segments. Thus, our dataset offers two versions: the first version consists of edited full-length songs, and the second version comprises all usable, high-quality vocal segments, incorporating redos from the recording process.

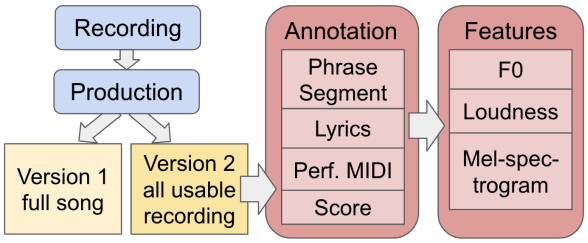

Figure 1. Data collection pipeline.

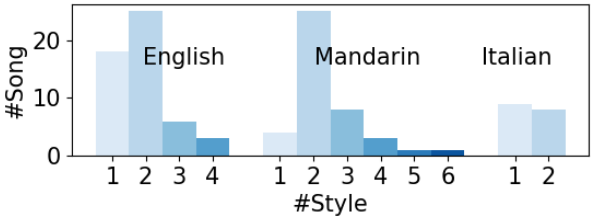

Figure 2. Distribution of songs according to languages and the number of style demonstrations. For example, English songs have 18 songs with only one style version, 25 songs with two different styles, six songs with three styles, and three songs with four styles.

We preserve these redos for two primary reasons. First, during recording, singers often need to restart due to minor errors, resulting in many redos that far exceed the quantity required for a single song. The high-quality vocals in these redo segments are perfect for segmented training in deep learning and effectively augmenting the dataset. Second, even when the same singer performs the same song using the same style, each rendition exhibits subtle differences. Capturing these variations provides valuable training data for learning multi-modes in singing performance and disentangling a singer's style with music content. This paper focuses on describing the second version of the dataset.

Upon obtaining the clean and dry vocal segments in audio, we manually annotate them into phrases (music sentences), provide lyrics and score alignment with audio at the phoneme level. We then extract acoustic attributes such as F0 contour, loudness curve, and Mel-spectrogram. Finally, we partition and package the data, incorporating relevant attributes. Section 4 describes this process in detail.

In the following subsections, we delve into the dataset's repertoire and styles, singer profiles, recording environments, and post-production methods, accompanied by pertinent statistics.

3.2 Repertoire and Style

SingStyle111 comprises 111 songs, of which 80 have at least two different versions performed in distinct styles by the same singer, resulting in a total of 224 song versions. The dataset encompasses three languages: English (372 minutes), Chinese Mandarin (307 minutes), and Italian (88 minutes). Figure 2 illustrates the number of song versions for each language. During song selection, we sought to diversely represent various styles, singing techniques, tempos, and eras.

Figure 3 presents the styles of the original songs and all

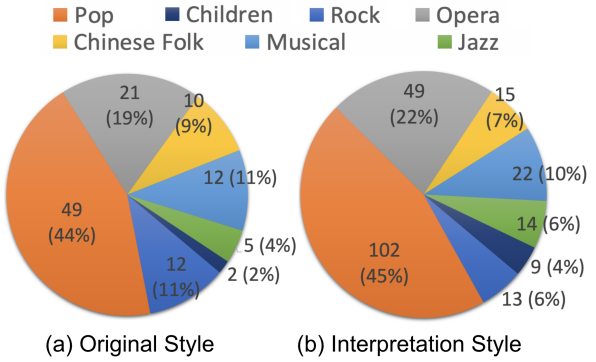

Figure 3. Distribution of song styles. Chart(a) describes the original style of the 111 songs, while chart(b) indicates the 224 different style interpretations in the dataset.

style demonstrations. We consolidated several sub-genres into seven broader styles to streamline the pie chart. For instance, Country, Western folk, Chinese pop, and other pop styles were combined into a single pop genre. Likewise, the Rock category contains Soft Rock, Hard Rock, Alternative Rock, etc.

Throughout the data collection process, we instructed singers to exhibit significant differences in style transfer. Sometimes they made appropriate adaptations or improvisations to the original song for better style transfer while preserving the original lyrics, melody, and structure. For example, it is easier for singers to transfer vocal timbres when the key changes. Also, tempo changes and rhythmic variations can dramatically help alter styles, such as transferring a fast and happy song into a slow and melancholic one. Converting singing techniques or adding ornamentations are also prevalent in our style transfer examples. For instance, the dataset includes many demonstrations interchanged among pop, bel canto, and Chinese traditional folk singing; or singing the same song in the distinct pop styles of Adele Adkins and Teresa Teng. In addition, some styles include deliberate emotional changes, for example, contrasting a "plain and lyrical style" with an "exaggerated and highly emotional style."

3.3 Singers

We paid eight professional singers (Table 2) to sing the songs. They have diverse vocal ranges, singing styles, and vocal techniques. They are aged 20 to 63, and all have received formal musical training for more than six years. Six of them are graduates or current students in the voice major at music conservatories. “Male1” is a native American English speaker, and all the others are Chinese. “Female1” has lived in the US for more than five years and received formal English singing training at a music academy. We also removed the English song phrases that have strong foreign accents. All singers have signed agreements to release the dataset for research purposes.

3.4 Recording

We recorded the songs in a professional recording studio with little reverberation or noise. We use a Shure Model SM81-LC microphone, an Apollo X8 Thunderbolt 3 audio interface, Heritage Audio 73jr as the pre-amplifier, and Pro Tools Studio as DAW software. All singings are pure vocal only and recorded at 44,100 Hz sampling rate with 24 bits per sample in wav format.

In most recording sessions, singers wear headphones to listen to the accompaniment. However, in some style transfer demonstrations, accompaniments and headphones may not always be used. Despite this, singers must ensure they maintain the correct key and consistently stay within it throughout the performance.

Singer	Language	Style	#Hour	Range
Female1	en, cn	P. C. O. R. F. M. J.	3.73	F#3-A5
Female2	it, en, cn	O. F. M.	1.24	E4-C6
Female3	cn, en	P. C. O. R. F. M. J.	1.58	F#3-F5
Female4	cn, en	P.	1.63	D3-C5
Male1	en	P. R. M.	0.59	D2-G4
Male2	cn, en	P. M. J.	1.35	A2-C5
Male3	it, cn	O. M.	1.16	C4-G5
Male4	cn	P. O. F.	1.51	D#3-A4

Table 2. Singer Information. Here the vocal range is the used range in the dataset. en: English, cn: Chinese, it: Italian, P: Pop, C: Children, R: Rock, O: Opera, F: Chinese Traditional Folk, M: Musical, J: Jazz.

3.5 Production

We employed several essential post-production techniques to refine and clean the recorded data. First, we edited the raw recordings to retain only high-quality clips, filtering out noisy sections, mistakes, and mispronunciations. A small portion of singer Male3’s singing clips were further edited with pitch-tuning. To achieve a consistent volume balance, we applied different gain levels in each clip. Moreover, we incorporated a compressor for all recording clips to prevent extreme dynamic fluctuations. Lastly, we maximized the output volume using a limiter, setting the output ceiling at -0.6 dB. After production, we obtained clean and dry vocal tracks with similar output volumes.

4. ANNOTATION AND ANALYSIS

This section presents the annotation process, including both manual annotation and automatic analysis. We first segment the audio clips into music phrases, for which we then manually identify corresponding lyrics and music scores. By combining automatic algorithms and manual efforts, we align lyrics phonemes and score notes to their corresponding audio. Next, we utilize algorithms to extract acoustic attributes such as F0 contour, loudness curve, and Mel-spectrogram. Finally, we highlight the key attributes and explain dataset partitioning and packaging.

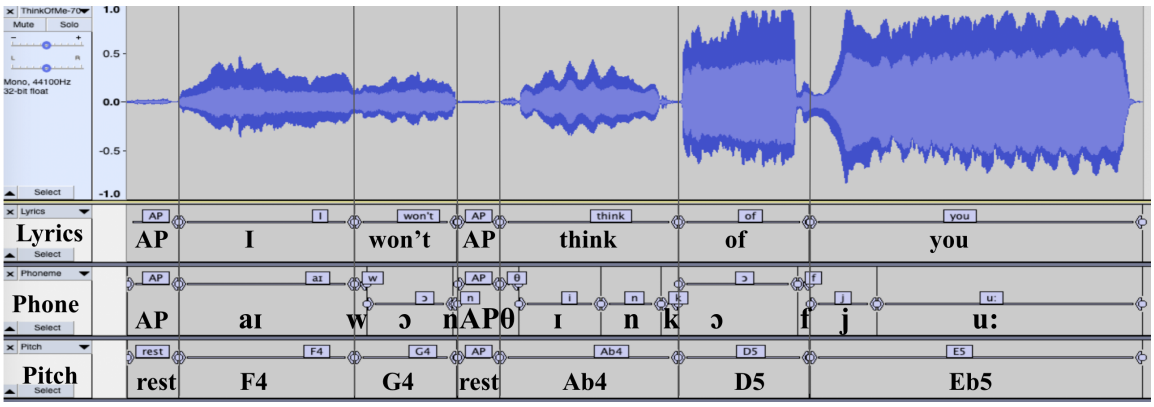

Figure 4. An example of phoneme-level annotation using Audacity. The lyrics word, IPA phoneme, and pitch label tracks are aligned with the corresponding audio. "AP" here stands for "aspirate".

4.1 Phrase-level Segmentation

We further divide the audio segments into smaller musical phrases for two reasons. First, this additional segmentation accelerates model training. Second, from a music perspective, phrases serve as one of the basic music structure units, with emotional expressions and performance controls being highly related to phrase-level structure. Inappropriate segmentation might compromise musical expression due to insufficient phrase structure information. Given the low accuracy of automatic algorithms for phrase segmentation, we manually label them. The result shows that most segmented phrases have lengths between 3 and 12 seconds with one to three breaths. We obtain 6588 phrases in total. No silence is included at the beginning or end of the phrases, except for breath events.

4.2 Lyrics and Score Alignment at Phoneme Level

In this subsection, we describe the lyrics and score annotation process.

Lyrics annotation We first manually find lyrics for each song online and then segment and align the lyrics with each phrase. We manually correct the lyrics to match the actual singing in the data. Secondly, we employ algorithms to translate the lyrics into phonemes. For English [1] and Italian [2], we utilize tools to translate them into International Phonetic Alphabet (IPA) phonemes [53]. For Mandarin, we use Pinyin [3] for phoneme-level alignment and provide a mapping of Pinyin to IPA phonemes for later phoneme-set processing for model training. We did not directly convert Chinese to IPA due to annotation complexity. Thirdly, we obtain an approximate phoneme alignment with audio using the Montreal Forced Aligner [54] and output it into TextGrid files. Finally, we (1) use *Praat* software [55], or (2) convert the TextGrid into txt files and input them to *Audacity* [56] for further manual adjustment of phoneme text and boundaries, as well as breath and silence event annotation (Figure 4).

Performance MIDI and Score annotation We annotated the performance MIDI file and music score for each singing phrase in the dataset as follows:

(1) We manually input performance MIDI files that strictly align to singing audio using MIDI piano, including multiple rounds of correction.
(2) We automatically align MIDI notes with phonemes based on their corresponding time stamps in the audio.
(3) We search online for music score MIDI files; if no reliable sources are found, we quantize and derive the score from annotated performance MIDI file.
(4) For online score files, we develop an algorithm that automatically matches each singing phrase's performance MIDI data to the corresponding phrase in the score MIDI file. Manual matching is required for non-original-style style transfer versions.
(5) We use the Dynamic Time Warping algorithm to match the performance MIDI data with the score MIDI file within each phrase. We manually verify the mapping results for non-original-style style transfer versions.

All these above steps allow us to annotate the lyrics, performance MIDI, and music score at the phoneme level for our singing voice dataset, ensuring accurate and comprehensive representations of the musical content.

4.3 Acoustic Feature Extraction

F0, or fundamental frequency, is the lowest frequency of a periodic waveform. F0 contour is critical in singing synthesis as it determines the pitch variations of singing performance and largely influences singing quality. It can capture pitch modulations in various singing techniques, such as vibrato, ornaments, and glissando. Many current SVS systems still require the input of ground-truth F0 as a condition to guide the synthesis process. To ensure accurate F0 extraction, we employ a combination of two widely-used models, pYIN [57] and PENN [58]. First, we use pYIN algorithm to identify unvoiced parts, including breaths, silence, and consonants. Then, the PENN algorithm is applied to extract F0 for the voiced parts.

Loudness represents the energy of a sound. It is crucial in singing performance since it largely reflects the dynamic and emotional changes that contribute to the expressiveness of the singing voice. To extract loudness, we first calculate the root-mean-square (RMS) amplitude values from audio and then convert them to decibels. We further ap-

[1] https://github.com/mphilli/English-to-IPA
[2] https://espeak.sourceforge.net/
[3] https://github.com/mozillazg/python-pinyin

ply a moving average window of frame size 30 to obtain a smoother loudness curve.

Finally, we use the Short-Time Fourier Transform (STFT) with a window size of 1024, FFT size of 1024, and hop size of 256 to extract the mel-spectrogram, which shares the same settings with loudness extraction.

5. POTENTIAL APPLICATIONS

This dataset is intended to promote research into a number of different MIR tasks. We consider a variety of interesting relevant problems in this section.

5.1 Singing Style Transfer

Style transfer has to do with music interpretation. Here, "style" refers to performance details that are not constrained by symbolic representations such as traditional notation. If notation gives a song its "identity," styles are performance characteristics that are shared across performances of different songs. Styles are often associated with genre, e.g., a song can be interpreted in rock, pop, or jazz styles. Styles can be more or less specific than genre, e.g., the style of Louis Armstrong (more specific) or symphonic (less specific). Style transfer is a process of identifying the style of one or more performances and applying it to a new song to create a stylistic performance. SingStyle111 contains many performances where a single singer performs in multiple styles, offering the potential to abstract styles from other information (singer identity, melodies) which is held constant. In the multi-style recordings, singers were asked to exaggerate differences, which should help to learn features that characterize different styles.

5.2 Singing Voice Synthesis

A large motivation for SingStyle111 is the difficulty of finding high-quality musical examples of singing. In particular, the presence of accompaniment and reverberation complicate the process of learning to create the sound of singing voices. Furthermore, lower recording and singing quality are a barrier to learning high-fidelity sounds of professional singing. In addition, SingStyle111 also provides necessary phoneme-level annotations for score-based SVS.

5.3 Singing Voice Conversion

In SVC, we hope to substitute the sound of one voice with the sound of another while maintaining the same melody and style. To promote progress in this area, SingStyle111 has performances of the same song by multiple singers, including male and female voices. Since we have performances of the same song in the same style, SVC can be cast as a sequence-to-sequence problem analogous to many other machine learning tasks such as language translation.

5.4 Expressive Performance

Expressive performance is the general problem of creating a musical performance given a symbolic description such as a melody in common music notation. Notation omits many details, including loudness, vibrato, pitch variations, changes in vocal timbre, the details of pronouncing lyrics within pitch and rhythmic constraints, and breathiness. Often, connections and transitions from one note to the next are as important as how notes are performed. To learn expressive performance, it helps to have symbolic notation, which can be considered as input constraints, context, or conditioning. In addition, it helps if the notated events are aligned with corresponding time points in the audio. SingStyle111 includes symbolic representations (performance MIDI files and music scores) aligned with audio. The data is especially designed to support machine learning using sequence-to-sequence models from notation to control signals such as pitch contours, loudness, spectrograms, or directly to audio.

5.5 Automatic Singing Transcription

Singing transcription can be regarded as the inverse of expressive performance control: Rather than converting notation to sound, we wish to convert sound into music notation. With transcriptions for all of the singing examples, SingStyle111 provides a wealth of transcription examples for training and evaluating transcription models.

5.6 Score Alignment and Following

Score following [59] is the problem of aligning an audio performance to symbolic notation. Vocal score following is particularly difficult because, unlike most other instruments, voices do not have keys, valves, or frets, so singing cannot be easily reduced to a sequence of distinct discrete states corresponding to musical notes [60]. Real-time score following is the first step in the task of computer accompaniment, in which a computer synchronizes a pre-composed accompaniment to a live performance by a soloist. Score following has also been used for automatic page turning, delivering synchronized comments via mobile phones to symphony orchestra audiences, and as a data collection method for learning music segmentation and other tasks. SingStyle111 contains accurate alignments for learning and evaluation of automatic alignment and real-time score following.

5.7 Lyrics Detection

The common task of understanding lyrics is one that even humans struggle with. SingStyle111 includes the lyrics used by the singers, and lyrics are aligned to the audio down to the phoneme level, facilitating learning and evaluation of various lyrics transcription and alignment tasks.

6. CONCLUSION

In conclusion, we introduce SingStyle111, a large-scale, high-quality, multilingual singing voice dataset that caters to various styles and includes style transfer examples. We provided detailed annotations of lyrics and scores at the phoneme level, together with extracted acoustic features. We will make the dataset freely available for research purposes to facilitate relevant MIR tasks.

7. REFERENCES

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[4] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] OpenAI, "Gpt-4 technical report," 2023.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel-spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*.

[11] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations*.

[12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[13] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[14] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*.

[21] 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.

[22] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," *Proc. Interspeech 2018*, pp. 3067–3071, 2018.

[23] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6179–6183.

[24] 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.

[25] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*.

[26] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "Xiaoicesing: A high-quality and integrated singing voice synthesis system," *Proc. Interspeech 2020*, pp. 1306–1310, 2020.

[27] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[28] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[29] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "Ppg-based singing voice conversion with adversarial representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7073–7077.

[30] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[31] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2391–2395.

[32] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7402–7406.

[33] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *Proc. Interspeech 2020*, pp. 4506–4510, 2020.

[34] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

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

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[37] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[38] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[39] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6284–6288.

[40] S. Dai, Z. Zhang, and G. G. Xia, "Music style transfer: A position paper," *arXiv preprint arXiv:1803.06841*, 2018.

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

[42] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen *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.

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

[44] 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.

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

[46] B. Faghih and J. Timoney, "Annotated-vocalset: A singing voice dataset," *Applied Sciences*, vol. 12, no. 18, p. 9257, 2022.

[47] 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.

[48] 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.

[49] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases." in *Ismir*, vol. 2, 2002, pp. 287–288.

[50] J. Mora, F. Gomez Martin, E. Gómez, F. J. Escobar-Borrego, and J. M. Díaz-Báñez, "Characterization and melodic similarity of a cappella flamenco cantes." International Society for Music Information Retrieval Conference, ISMIR, 2010.

[51] R. M. Bittner, K. Pasalo, J. J. Bosch, G. Meseguer-Brocal, and D. Rubinstein, "vocadito: A dataset of solo vocals with f_0, note, and lyric annotations," 2021.

[52] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 310–319, 2009.

[53] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[54] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[55] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International 5:9/10, 341-345*, 2001.

[56] D. Mazzoni and R. Dannenberg, "Audacity [software]," *The Audacity Team, Pittsburg, PA, USA*, vol. 328, 2000.

[57] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[58] M. Morrison, C. Hsieh, N. Pruyne, and B. Pardo, "Cross-domain neural pitch and periodicity estimation," in *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, TODO 2023.

[59] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, August 2006.

[60] L. Grubb, "A probabilistic method for tracking a vocalist," PhD thesis, Carnegie Mellon University, 1998.