Supplementary Materials

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

C Datasheet

In this section, we copy the questions from Datasheets for Datasets [8] and provide details about the GTSinger, which is published with a permanent DOI `10.57967/hf/2498`.

C.1 Motivation

- **Q: For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

  A: The scarcity of high-quality and multi-task singing datasets significantly hinders the development of diverse controllable and personalized singing tasks, like technique-controllable singing voice synthesis (SVS), technique recognition, style transfer, and speech-to-singing (STS) conversion. Existing open-source singing datasets suffer from low quality, limited diversity of languages and singers for global styles, absence of multi-technique information for technique modeling and control, lack of realistic music scores for real-world composition, and poor task suitability (like global labels for singing method and emotion recognition, paired speech for STS conversion). Therefore, 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 singing tasks. Thus different singing approaches that use this corpus can be easy to reproduce, and make fair comparisons between each other.

- **Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  A: Zhejiang University.

- **Q: Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

  A: Zhejiang University.

- **Q: Any other comments?**

  A: No.

C.2 Composition

- **Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  A: Singing voices, manual phoneme-to-audio alignments, phoneme-level technique annotations, global style labels (singing method, emotion, pace, and range), realistic music scores, paired speech under the same lyrics.

- **Q: How many instances are there in total (of each type, if appropriate)?**

  A: 80.59 hours of high-quality singing voices in nine languages with alignments, style and technique annotations, along with realistic music scores. 16.16 hours of paired speech with alignments.

[8] https://arxiv.org/abs/1803.09010

- **Q: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  A: GTSinger is specifically collected and created by us as a high-quality and multi-task singing corpus. It is practically impossible to cover all possible instances of all languages and singing techniques worldwide. Our dataset encompasses nine widely used languages and six of the most frequently used singing techniques. Additionally, we have selected 20 professional singers covering all four vocal ranges. Thus, GTSinger is a carefully designed, independent, global, multi-technique singing dataset aimed at supporting all singing tasks, without claiming to cover every possible singing instance. We provide a complete set of data processing codes to facilitate future expansions of the dataset in terms of languages, techniques, and duration.

- **Q: What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

  A: Each instance consists of raw data of singing voices or speech.

- **Q: Is there a label or target associated with each instance?** If so, please provide a description.

  A: Singing voices with manual phoneme-to-audio alignments, phoneme-level technique annotations, global style labels (singing method, emotion, range, pace), and realistic music scores. Speech with manual phoneme-to-audio alignments.

- **Q: Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  A: No.

- **Q: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  A: Yes. 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). 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.

- **Q: Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

  A: Yes, we have predefined several tasks and introduced our rules for dividing the training and test sets. We randomly selected five songs (including paired speech for STS conversion) from each singer, totaling 100 songs for the test set, with the remainder used as the training set. For cross-lingual technique recognition, models are trained on one group of languages (e.g., Asian) and tested on the other (e.g., European).

- **Q: Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  A: The singing voices are recorded in a professional studio, ensuring high quality, fidelity, and clarity. The annotation process is carried out by professionals with musical and linguistic backgrounds. Small errors are inevitable in annotations. Any errors present are due to slight differences in subjective perception, which may affect the annotations of technique, word, and phoneme duration annotations. However, these few and minor errors do not compromise the overall quality of the dataset. For instance, 0.001 second's difference in word boundaries won't influence the performance of SVS models.

- **Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  A: Self-contained.

- **Q: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

  A: No.

- **Q: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  A: No.

- **Q: Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

  A: Our 20 singers are categorized by their vocal ranges: 6 tenors, 3 basses, 7 sopranos, and 4 altos.

- **Q: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

  A: No.

- **Q: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

  A: Yes, we have biometric data in our dataset. Therefore, We first perform data desensitization and consider using techniques such as vocal watermarking to further protect personal privacy.

- **Q: Any other comments?**

  A: No.

C.3 Collection Process

- **Q: How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  A: Singing voices and speech are recorded in a professional recording studio.

- **Q: What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

  A: The audios were recorded in the recording studio.

- **Q: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  A: Our dataset encompasses nine widely used languages and six of the most frequently used singing techniques. Additionally, we have selected 20 professional singers covering all four vocal ranges. Thus, GTSinger is a carefully designed, independent, global, multi-technique singing dataset aimed at supporting all singing tasks, without claiming to cover every possible singing instance.

- **Q: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  A: Professional singers. 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.

- **Q: Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

  A: The singing voices with paired speech were recorded and collected over the time period July 2023 - May 2024.

- **Q: Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

  A: No.

- **Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

  A: From the individuals.

- **Q: Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

  A: Yes, all singers sign an agreement and agree to make their singing voice open-source for academic usage before collection.

- **Q: Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

  A: Yes, all singers sign an agreement and agree to make their singing voice open-source for academic usage before collection.

- **Q: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

  A: Yes, we add the mechanism to revoke their consent in the agreement.

- **Q: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

  A: No.

- **Q: Any other comments?**

  A: No.

C.4 Preprocessing/cleaning/labeling

- **Q: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

  A: We performed data cleaning, annotation, and segmentation.

- **Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

  A: The raw audio data is archived in our GitHub repository [9]. However, due to the excessive duration of WAV files, the raw data is not suitable for model training.

- **Q: Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

  A: The code for preprocessing, cleaning, and labeling is available in the same GitHub repository [9].

- **Q: Any other comments?**

  A: No.

C.5 Uses

- **Q: Has the dataset been used for any tasks already?** If so, please provide a description.

  A: Yes, it has been presented in the paper, including technique-controllable singing voice synthesis, technique recognition, style transfer, and speech-to-singing conversion.

- **Q: Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

  A: This has been presented in the paper.

- **Q: What (other) tasks could the dataset be used for?**

  A: all current singing tasks, like singing voice synthesis related, emotion and technique related, and music information retrieval related works.

- **Q: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

  A: No.

- **Q: Are there tasks for which the dataset should not be used?** If so, please provide a description.

  A: No.

- **Q: Any other comments?**

  A: No.

C.6 Distribution

- **Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

  A: Yes, the dataset is freely and publicly available and accessible.

[9] `https://github.com/GTSinger/GTSinger`

- **Q: How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  A: The dataset is free for download by everyone. Links are available in the GitHub repository [9]. The DOI of the dataset is `10.57967/hf/2498`.

- **Q: When will the dataset be distributed?**

  A: The dataset is distributed as of June 2024 in its first version.

- **Q: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

  A: The dataset is licensed under CC BY-NC-SA 4.0 license.

- **Q: Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

  A: No.

- **Q: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  A: No.

- **Q: Any other comments?**

  A: No.

C.7 Maintenance

- **Q: Who will be supporting/hosting/maintaining the dataset?**

  A: Zhejiang University.

- **Q: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  A: yuzhang34@zju.edu.cn.

- **Q: Is there an erratum?** If so, please provide a link or other access point.

  A: Currently, there is no erratum. If errors are encountered, the dataset will be updated with a fresh version. They will all be provided in the same GitHub repository [9].

- **Q: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

  A: Same as above.

- **Q: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

  A: No.

- **Q: Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

  A: No, all the updates will be synced on the website.

- **Q: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

A: The code for preprocessing, cleaning, and labeling is available in the GitHub repository [9]. You can follow the pipeline in our paper to extend/augment/build on/contribute to GTSinger.

- **Q: Any other comments?**

  A: No.

C.8 Reproducibility of the benchmarks

The training details and all benchmark models are provided in the GTSinger GitHub repository [9] for the reproducibility of reported results.

C.9 Data Format

Our dataset is organized hierarchically. It presents nine top-level folders, each corresponding to a distinct language. Within each language folder, there are five sub-folders, each representing a specific singing technique. These technique folders contain numerous song entries, with each song further divided into several controlled comparison groups: a control group (natural singing without the specific technique), a technique group (densely employing the specific technique).

Our singing voices and speech are recorded at a 48kHz sampling rate with 24-bit resolution in WAV format. Alignments and annotations are provided in TextGrid files, including word boundaries, phoneme boundaries, phoneme-level annotations for six techniques, and global style labels (singing method, emotion, pace, and range). We also provide realistic music scores in musicxml format. Notably, we provide an additional JSON file for each singing voice, facilitating data parsing and processing for singing models.