Children's Song Dataset for Singing Voice Research

Soonbeom Choi **Wonil Kim** **Saebyul Park** **Sangeon Yong** **Juhan Nam**

Graduate School of Culture Technology, KAIST, South Korea

{cjb3549, ianwonilkim, saebyul_park, koragon2, juhan.nam}@kaist.ac.kr

ABSTRACT

We introduce the Children's Song Dataset (CSD) which contains vocal recordings of 100 children's songs in Korean or English. The audio recordings are precisely aligned with the MIDI transcriptions and lyrics annotations and so we expect that the dataset can be useful for various singing voice analysis and synthesis tasks.

1. INTRODUCTION

The primary method of singing voice synthesis has recently moved from concatenative synthesis or statistical parametric synthesis to deep learning-based approaches. While the deep learning approaches require significantly streamlined data processing pipeline than the previous methods, they require a large set of audio recordings aligned with melody and lyrics. However, there are few publicly available datasets possibly due to the copyright issue [1] [2]. The open-source datasets such as NUS-48E [3] and Kara1k [4] do not have time annotation for lyrics and has no melody data. Sinsy is a publicly available dataset used for singing voice synthesis but it is limited to the Japanese language [5].

To address the lack of open-source datasets in singing voice synthesis research, we collected our own singing voice recordings along with melody and lyrics annotations. The dataset contains vocal recordings in two different languages, Korean and English, and unlike the DALI dataset [6], they are manually annotated with MIDI notes and lyrics in time. While this dataset was originally designed for singing voice synthesis, it can be used for singing voice analysis as well, for example, singing voice transcription (MIDI or lyrics) from audio recordings and audio-to-score or audio-to-lyrics alignment using the time annotations.

2. DATASET CONTENTS

The dataset is composed of 50 Korean and 50 English songs sung by a Korean female professional pop singer. Each song is recorded in two separate keys, ranging from 3 to 5 semitones, resulting in a total of 200 audio recordings. We collected the children's songs to avoid the possible copyright issues in commercial pop music. 25 songs are recorded in both Korean and English. They are originated from western music (they share the same melody) but are rearranged with Korean lyrics.

Each audio recording is paired with a MIDI transcription file and a lyrics annotation file. Singing voice is an highly expressive sound and so it is hard to define precise onset timings and pitches compared to instrumental sounds. We guided the singer to maintain a consistent singing style during the recording session. Also, we tried to annotate the note onset and duration consistently for various expressions [1] .

2.1 Vocal Recording

Children's songs usually have various versions in different styles. We chose one of them that suites for the singer. While recording vocals, she sang along with the background music tracks. She deliberately rendered the singing in a "plain" style refraining from expressive singing skills. The recording took place in a dedicated soundproof room. We recorded three to four takes for each song and combined the best parts into a single audio track.

2.2 MIDI Transcription

We chose MIDI for melody annotation because they can be easily modified and visualized on music production software. The MIDI data consists of monophonic notes. Each note contains onset and offset times which were manually fine-tuned along with the corresponding syllable. We did not include any expression data or control change messages because we made this dataset primarily for end-to-end singing voice synthesis and those parameters can be ambiguous to define. There can be a variety of criteria to align audio recordings with MIDI notes. We assumed one syllable matches with one MIDI note and made the following criteria to represent various expressions in singing voice.

- A piano sound is used as a reference tone for the annotated MIDI to ensure the alignment with vocal

Attribution: S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Children's Song Dataset for Singing Voice Research", *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

[1] The dataset is available at : `https://github.com/emotiontts/emotiontts_open_db/tree/master/CSD`

Language (number of songs)		Korean (50), English (50)
Number of key per song		2
Pitch range		F3 - F5
File formats	Audio	44.1Hz, 16bit in WAV format
	MIDI	Monophonic MIDI without any expressions or control change messages
	Lyrics	Grapheme level text annotations in a plain text format

Table 1. Audio, MIDI and lyrics specification of the CSD dataset

- The rising pitch at the beginning of a note is included within a single note
- The end of syllable is treated as the offset of a note
- The breathing sound during short pauses is not treated as note onset or offset
- Vibrato is treated as a single sustaining note
- If a syllable is rendered with several different pitches, we annotated them as separate notes

2.3 Lyric Annotation

Lyrics are annotated in grapheme level with a plain text format. We did not include other information such as phonetic alignment because the onset and offset of MIDI notes can be used for syllable timings. English words in Korean songs are annotated with Korean syllables. The English lyrics include grapheme level and phoneme level annotations. When a syllable corresponds to multiple notes, it is replicated so that one syllable matches one note.

3. APPLICATIONS

3.1 Singing Voice Synthesis System

We proposed a Korean singing voice synthesis system based on auto-regressive boundary equilibrium Generative Adversarial Network (BEGAN) using the Korean set of this dataset [7]. The system is based on our assumption that musical notes, the onset and offset timings and lyrics text are the minimum requirements to model a singing voice. In the future, we plan to use both Korean and English sets to make the system generate singing voice in the two languages.

3.2 Automatic Singing Transcription and More

The dataset can be used to predict the MIDI melody or lyrics from the audio recordings or for automatic alignment between lyrics and audio or between MIDI to audio using the precise timing information.

4. ACKNOWLEDGEMENT

This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No. 10080667, Development of conversational speech synthesis technology to express emotion and personality of robots through sound source diversification).

5. REFERENCES

[1] H. Kenmochi and H. Ohshita, "VOCALOID - Commercial Singing Synthesizer based on Sample Concatenation," in *INTERSPEECH*, 2007, pp. 4011–4012.

[2] L. Zhang, C. Yu, H. Lu, C. Weng, Y. Wu, X. Xie, Z. Li, and D. Yu, "Learning Singing from Speech," *arXiv preprint arXiv:1912.10128*, 2019.

[3] Z. Duan, H. Fang, B. Li, K. 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*, 2013, pp. 1–9.

[4] Y. Bayle, L. Marsik, M.Rusek, M. Robine, P. Hanna, K. Slaninova, J. Martinovic, and J. Pokorny, "Kara1k: A Karaoke Dataset for Cover Song Identification and Singing Voice Analysis," in *Proceedings - 2017 IEEE International Symposium on Multimedia*, 2017, pp. 177–184.

[5] Y. Hono, S. Murata, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Recent Development of the DNN-based Singing Voice Synthesis System - Sinsy," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 211–216.

[6] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 431–437.

[7] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Korean Singing Voice Synthesis Based on Auto-Regressive Boundary Equilibrium Gan," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2020-May, 2020, pp. 7234–7238.