HSD: A hierarchical singing annotation dataset

Xiao Fu¹, Xin Yuan¹, Jinglu Hu¹

¹Graduate School of Information Production and Systems
Waseda University
Kitakyushu, Fukuoka, Japan
hirabara@toki.waseda.jp, sherryyuan@ruri.waseda.jp, jinglu@waseda.jp

Abstract—Commonly music has an obvious hierarchical structure, especially for the singing parts which usually act as the main melody in pop songs. However, most of the current singing annotation datasets only record symbolic information of music notes, ignoring the structure of music. In this paper, we propose a hierarchical singing annotation dataset that consists of 68 pop songs from Youtube. This dataset records the onset/offset time, pitch, duration, and lyric of each musical note in an enhanced LyRiCs (LRC) format to present the hierarchical structure of music. We annotate each song in a two-stage process: first, create initial labels with the corresponding musical notation and lyrics file; second, manually calibrate these labels referring to the raw audio. We mainly validate the labeling accuracy of the proposed dataset by comparing it with an automatic singing transcription (AST) dataset. The result indicates that the proposed dataset reaches the labeling accuracy of AST datasets.

Index Terms—Music information Retrieval, singing annotations, dataset preparation, hierarchical data

I. INTRODUCTION

Vocals in pop songs usually have a hierarchical structure as described in Figure 1. A piece of vocal music usually consists of several musical phrases and each phrase contains several musical notes. Meanwhile, a single musical note presents five basic properties: onset, pitch, duration, lyric, and offset. Recently, various singing annotation datasets have been proposed. According to the research motivations, the contents of these singing datasets are also different. For instance, AST [1] mainly requires the data of pitch, onset, and offset time for each musical note, symbolic music generation (SMG) [2] needs pitch and duration, and automatic lyric transcription (ALT) [3] calls for lyrics and the corresponding timestamps. However, there are few singing annotation datasets that fully record the hierarchical information.

In this work, we present a singing annotation dataset (namely HSD), which not only provides the information of onset/offset, pitch, duration, and lyric for each musical note but also records these data in a hierarchical structure. We initialize the vocal labels via a method (named N&L) using the corresponding musical notation and LRC file and then execute a manual calibration process to produce the final labels. We validate the labeling accuracy by comparing it with the MIRST-500 dataset [4] which is proposed for AST task.

II. RELATED WORK

In the last few decades, several vocal-music-related datasets have been proposed for various MIR tasks. The Jamendo lyrics database [5] was designed for lyrics alignment, transcription, and evaluation. The iKala dataset [6] and MIR-1K dataset [7] were offered for singing voice separation. The ISMIR 2014 dataset [8], TONAS dataset [9], and MIR-ST500 dataset [4] were provided for AST. The RWC music database [10], which labeled both the vocals and the instrumental accompaniment in real-world songs, was proposed for common use and for the purposes of musical information processing research. We compare the contents of HSD with several similar datasets as shown in Table I.

III. LABELING METHOD

The labeling process is executed in two-step: initialization and calibration. We use an N&L (i.e., notation and LRC) method to create the initial labels. For given raw audio, we first collect the corresponding musical notation and LRC file from the web (the music notation offers pitch and duration for each musical note and LRC file provides phrase-level lyrics and timestamps). Second, we segment the raw audio into music phrases according to the phrase onset time tags \( t_{1:n} \) from the LRC file, where \( n \) represents the phrase number. At last, we locate each musical note by computing its percentage in the phrase according to the duration \( d_{1:m} \) from the musical notation, where \( m \) represents the number of notes in the sentence. This initialization method requires the annotators to have some basic knowledge of music to read the musical notation and input each musical symbol into the initial program.

After the initialization, a manual calibration process is executed to correct the coarsely initialized labels to accurate labels. The initialized labels are processed to MIDI format first and then put into a parallel track with the raw audio track. The annotators play these two tracks simultaneously to compare the vocals with the annotations. Since the pitch and duration obtained from musical notation are almost correct, the annotators are mainly needed to fix the data of the onset/offset time. After two annotators (one for labeling and another...
for validation) consider the labels are already accurate, the annotations of a song are exported to an enhanced LRC format as the final output.

IV. DATA RECORD

The proposed dataset is available at https://github.com/hirabarahyt/HSD-Dataset. There are 68 annotation files in enhanced LRC formats in the dataset. The YouTube link for each song is offered. The enhanced LRC format is based on the ordinary LRC format with an onset time tag “[mm:ss.xx]” and an offset time tag “[mm:ss.xx]” added for each lyric, where “mm” is minutes, “ss” is seconds, and “xx” is hundredths of a second. We included the pitch and duration alongside each lyric in the format “[m p d]” after the onset time tag and before the offset time tag, where “m” is the lyric, “p” is the pitch, and “d” is the duration. The duration was transformed to floating point units, for example, a quarter note typically represents one beat, and a sixteenth note is one-quarter of a quarter note, so a sixteenth note duration is represented as 0.25. The phrase time tag is stated at the beginning of each phrase in the format “[mm:ss.xx].” Each line in the enhanced LRC file can thus be described as follows: [mm:ss.xx][mm:ss.xx] lp d[mm:ss.xx]…[mm:ss.xx] lp d[mm:ss.xx]. We also provide the fixed notations, LRC files, and the source codes used to create the annotations in the same repository. Users can reuse this framework to produce their own data if they have musical notations and LRC files.

V. VALIDATION

We mainly validate the proposed dataset by comparing the COnP with the MIRST-500 [4] to confirm the labeling accuracy. However, intuitively quantifying the quality of singing annotations is difficult. Therefore we employ a contrastive method in the evaluation process as follows. For a given raw audio r and its annotation a, we first shift the onset/offset time of all the musical notes t milliseconds earlier and later to a− and a+ Then we compare the compatibility of r with a, a−, and a+ to find the most suitable annotation. If a is the best for r, we will consider the annotation a accurate; otherwise a will be considered inaccurate. Since evaluating the whole dataset is time-consuming, we randomly select 5 songs from HSD and MIRST-500 respectively for the validation. The results based on t = 50ms and t = 100ms are shown in Table II. From the results we can observe that our dataset reaches the same labeling accuracy as MIRST-500.

VI. CONCLUSION

In this paper, we propose a singing annotation dataset that records the pitch, duration, lyric, onset, and offset of each musical note in a hierarchical format. Through a contrastive validation method, we found our dataset achieves the labeling accuracy of the MIRST-500 dataset [4] which is designed for the AST task. In the future, we will explore how to use this kind of hierarchical information for MIR tasks.

| Dataset | Lyric | Pitch | Duration | Onset/Offset | Phrase-Timestamp | Hierarchical structure |
|---------|-------|-------|----------|--------------|------------------|-----------------------|
| MIR-ST500 | ×     | ✓     | ✓        | ×            | ×                | ×                     |
| Jamendo lyrics | note-level lyrics | ×     | ×        | ✓            | ✓                | ×                     |
| RWC     | phrase-level lyrics | ✓     | ✓        | ✓            | ✓                | ✓                     |
| HSD (Ours) | note-level lyrics | ✓     | ✓        | ✓            | ✓                | ✓                     |

**Table II**

| Dataset  | COnP (50ms) | COnP (100ms) |
|----------|-------------|--------------|
| MIRST-500 | 96.49       | 98.42        |
| HSD (ours) | 97.07       | 98.53        |

REFERENCES

[1] R. Nishikimi, E. Nakamura, S. Fukayama, M. Goto, and K. Yoshii, “Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 161–165.

[2] X. Fu, H. Deng, X. Yuan, and J. Hu, “Generating high coherence monophonic music using monte-carlo tree search,” IEEE Transactions on Multimedia, pp. 1–1, 2022.

[3] M. McVicar, D. P. Ellis, and M. Goto, “Leveraging repetition for improved automatic lyric transcription in popular music,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 3117–3121.

[4] J.-Y. Wang and J.-S. R. Jang, “On the preparation and validation of a large-scale dataset of singing transcription,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 276–280.

[5] D. Stoller, S. Durand, and S. Ewert, “End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model,” in ICA2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 181–185.

[6] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, “Vocal activity informed singing voice separation with the ikala dataset,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 718–722.

[7] C.-L. Hsu and J.-S. R. Jung, “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, 2010.

[8] E. Molina, A. M. Barbancho-Perez, L. J. Tardon-Garcia, I. Barbancho-Perez et al., “Evaluation framework for automatic singing transcription,” 2014.

[9] E. Gómez and J. Bonada, “Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing,” Computer Music Journal, vol. 37, no. 2, pp. 73–90, 2013.

[10] 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.