LARGE-SCALE ANALYSIS OF LYRICS AND MELODIES IN CANTONESE POP SONGS

Qiaoyu Yang, Panzhen Wu, and Zhiyao Duan
Audio Information Research Lab, University of Rochester
{qyang15, pwu10}@u.rochester.edu, zhiyao.duan@rochester.edu

ABSTRACT

In pop songs of tonal languages, researchers have found that the tones of lyrics characters and the melodies contours have similar patterns of motion. However, no large-scale quantitative analysis has been done to generalize the phenomenon. The current study explores the extent of relationship between lyrics and melodies quantitatively in a large dataset of pop songs written in Cantonese, a language with one of the richest tonal systems. To align the lyrics with corresponding melody, the singing voices are extracted from the polyphonic music tracks and automatic speech recognition (ASR) systems are applied to the singing voices to detect the lyrics content as well as character-wise timestamps. The transcribed lyrics are matched with true lyrics using Levenshtein distance and then further corrected to ensure the lyrics are precisely sung in each melody segment. Finally, the notes of the melody are extracted and compared with frequencies of generated speech to obtain quantitative relationship.

1. INTRODUCTION

Besides many similarities, there are also significant differences between music and language. The semantic structure of languages is much richer than music. There is no discriminative musical representation for even simple items such as an apple. On the other hand, if the features are restricted to signal properties of sound, music elements could be drawn in a larger spectrum than languages. While pitch information could define the content of the music such as melody contour and harmony, in spoken languages, the frequency of the sound is not usually directly contributing to the meaning of the words or sentences. Some languages have dialects with completely different pronunciations but using the same vocabulary and grammar [1]. However, in tonal language, the pitch information is essential. For example, in Cantonese, the same phoneme combination, fan, can be interpreted as different characters and have different meaning when pronounced at different tones pitches: 分 (fan1), 粉 (fan2), 訓(fan3), 焚(fan4). Since the pitch contour of sentences could distinguish the semantics of tonal languages, it is reasonable to hypothesize that a similar pitch pattern is also reflected in singing voices. In this paper, we are going to explore the relationship between the pitch motions of lyrics tones and melodies in songs with tonal languages, specifically Cantonese.

2. RELATED WORKS

The correspondence between lyrics and melodies has been studied before in Chinese. The earliest results in literature came in the 1980s. In 1987, Chan divided the Cantonese tone systems into three pitch categories and found significant correspondence in relative motion from six contemporary Cantonese songs [2]. In 1999, Jonathan Stock conducted a qualitative inspection on 7 excerpts of Beijing Opera. He concluded that Beijing Opera singers had the ability to plan melodic structures in a similar way as they executed the speech tones in the language traditions [3]. Ho looked closer into the songwriting process of Cantonese pop songs and found surprising results that native speakers could set appropriately matching text to melodies even without any musical training [4].

Despite significant melody-tone correspondence have been found in songs written in Chinese, all existing literature only provided qualitative analysis or small-scale quantitative analysis of very few samples. No large-scale quantitative analysis have been done in the domain. An obvious hurdle in front of researchers is the lack of large annotated datasets. In order to study the relationship between melodies and speech-tones, not only do we need the transcription of the pitch information of the songs, but the alignment between the lyrics characters and the melody notes is required. Although the past decade has witnessed significant progress in lyrics alignment and singing transcription [5–14], there are rarely any experiments in languages other than English. However, as seen in the musicology community, songs written in tonal languages in East Asia possess the tightest relationship between lyrics and music. In this paper, we attempt to use the current technology of speech processing to improve alignment results and quantify the relationship between lyrics and melodies in Cantonese pop songs in a semi-unsupervised method.

3. DATA

Around 750 Cantonese pop songs are used in the study. The original songs are extracted from the music streaming

Attribution: Q. Yang, P. Wu, and Z. Duan, “Large-scale analysis of lyrics and melodies in Cantonese pop songs”, in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

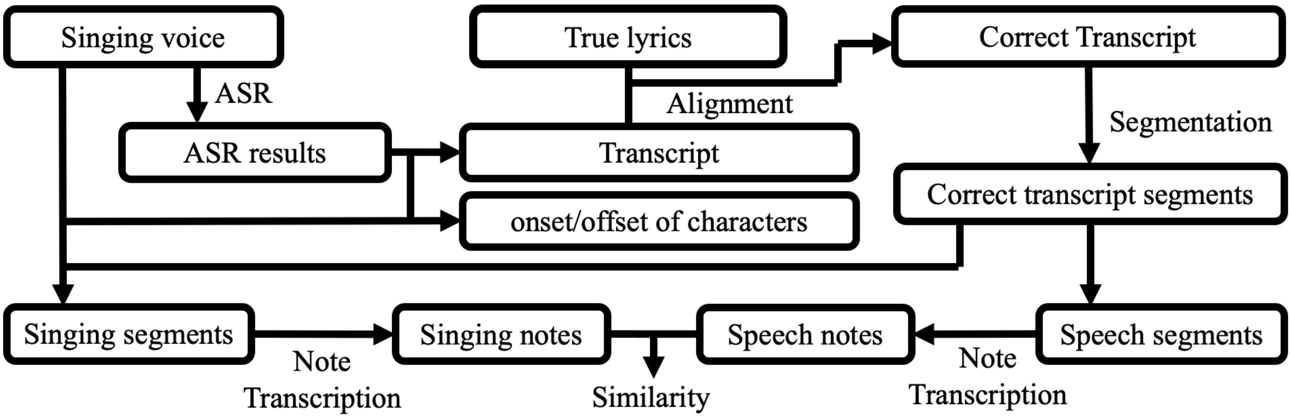

Figure 1: A complete pipeline of the the method

software, Spotify. The vocal part of each song is extracted using the music production software, iZotope's RX8. To further reduce the errors in later processes due to the extended durations of overtones that might cause overlapping voice content, reverberation reduction is applied on the separated vocal parts using RX8. The lyrics data are obtained from online karaoke websites such as kkbox.com. After listening and analyzing the sentence structures of the lyrics, we manually annotated the periods to help the segmentation process.

4. METHOD

To obtain valid conclusions of the relationship between lyrics and melody, the extracted notes should be estimating the real contours and the contours should be aligned with respect to the lyrical content. We propose the following procedures using characters as anchors to minimize the error due to imperfect results of note transcription and alignment. The complete pipeline of the procedures is summarized in Figure 1.

4.1 Alignment

We use the ASR system from the Kuai company to generate transcripts for each song. Inspired by [13], the transcripts are compared with the true lyrics to find unmatched regions using Levenshtein distance. If the unmatched regions are of the same length, wrong characters in the transcript will be replaced by the corresponding characters in the true lyrics. The matched regions are further segmented using the annotated periods as boundaries. The onset of the first character and offset of the last character in each sentence are used as time boundaries of each segment in the singing voices. The corresponding speech segment with the same sequence of characters is generated from Google's TTS system.

4.2 Note extraction

PYin and Praat are used to extracted pitch estimations of the singing voices and speeches, respectively. Since there are pitch changes within some Cantonese tones, multiple notes are extracted for each character. Within the time interval of a character, we apply K-means on the pitch estimations to obtain the expected note frequencies. The standard Cantonese tone system labels each tone with two pitch levels so k is set to be 2. The centers are initialized with the boundary estimations and the iteration stops when the centers converge. Additionally, in order to avoid the effect of the variability of note ranges to the contour relationships, the notes frequencies are normalized to lie within $[0, 1]$.

4.3 Relationship measures

After lyrics alignment and note extraction, we obtain the same number of notes in each contour segment for both lyrics and melodies. The notes in the same position of two contours represent the same lyrical content. We use two types of similarity measures to analyze the note contours. First, for pairs of close notes, the frequency of similar motions across the two contours is calculated. Notes within characters and across adjacent characters are considered separately. Motions across multiple characters are also compared to capture long-term relationships.

Another quantitative measure for the relationship is cosine similarity of the note contours as vectors. To analyze the local behavior of the contours, we also used a modified version of cosine. A short window slides across the contours and we take the average of the cosine similarity calculated within all windows.

5. RESULTS

The current results are shown in Table 1. There are considerable similar motions both within and across characters. An interesting observation is the windowed cosine similarity is significantly higher, suggesting a tighter relationship between the contours at the local level.

f_i	f_c	cos	$cos+$
0.701	0.681	0.801	0.850

Table 1: different similarity measures between the melody and speech contours

6. ACKNOWLEDGEMENT

We thank Yahui Shan, Jie Li and Xiaorui Wang at MMU of Kuaishou Technology for transcribing the singing voices using their ASR engine. This work has been partially supported by National Science Foundation grant 1846184.

7. REFERENCES

[1] J. Good and M. Cysouw, "Languoid, doculect, and glossonym: Formalizing the notion 'language'," *Language Documentation and Conservation*, vol. 7, December 2013.

[2] M. Chan, "Tone and melody in cantonese," in *Proc. of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 1987, pp. 26–37.

[3] J. Stock, "A reassessment of the relationship between text, speech tone, melody, and aria structure in beijing opera," *Journal of Musicological Research*, vol. 18, no. 3, pp. 183–206, January 1999.

[4] W. Ho, "The tone-melody interface of popular songs written in tone languages," in *Proc. of the 9th International Conference on Music Perception and Cognition*, 2006, pp. 1414–1422.

[5] C. Wang, R. Lyu, and Y. Chiang, "An automatic singing transcription system with multi-lingual singing lyric recognizer and robust melody tracker," in *Eighth European Conference on Speech Communica-tion and Technology*, 2003.

[6] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," in *Eighth IEEE International Symposium on Multimedia*, 2006, pp. 257–264.

[7] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–11, 2010.

[8] A. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Proc. of International Society of Music Information Retrieval*, 2016, pp. 358–364.

[9] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks," in *Proc. of International Society of Music Information Retrieval Conference*, 2016, pp. 737–743.

[10] Z. Fu and L. Su, "Hierarchical classification networks for singing voice segmentation and transcription," in *Proc. of International Society of Music Information Retrieval Conference*, 2019, pp. 900–907.

[11] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 161–165.

[12] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music us-ing an audio-to-character recognition model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 181–185.

[13] C. Gupta, H. L. R. Tong, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *Proc. of International Society of Music Information Retrieval Conference*, 2018, pp. 600–607.

[14] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 396–400.