Audio Engineering Society
Convention Express Paper 212

Presented at the 156th Convention
2024 June 15-17, Madrid, Spain

*This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Automatic Surtitle Generation Scheme for Live Cantonese Opera

Weike He[1] and Kevin Chau[1]

[1] *The Hong Kong University of Science and Technology, Hong Kong, China*

Correspondence should be addressed to Kevin Chau (eekchau@ust.hk)

**ABSTRACT**

Presently reported is an automatic surtitle generation scheme for live Cantonese opera based on beat tracking. The individual Chinese characters in the surtitles are animated in a karaoke style in sync with the performance. The scheme requires a music-vocal score of the opera but no human intervention during the live performance. A major advantage of the method is that the next Chinese character can be predictively highlighted at the onset of its vocalization with almost zero latency after compensating for the digitizing and processing time of the streaming audio. Specifically, a regularly spaced pulse train registering the onset position of every monosyllabic Chinese character to be sung or spoken is initially prepared from the music-vocal score. On-going synchronization of the pulse train to the live performance is obtained by monitoring the beats in the audio. Assuming beat tracking is accurate, the alignment of the surtitle characters should be 100% correct. An exception is speech that does not have a beat structure for which only an estimated speech rate can be established. The objective of this research is to promote Cantonese opera to those who do not speak Cantonese with the hope that the present surtitle generation scheme can be extended to other regional Chinese operas like Peking opera, Chaozhou opera, and Kun Qu opera.

## 1 Introduction

Cantonese opera, also known as Yueju (粵劇) [1-3], is an indigenous art form of South China that involves music, singing, spoken lines, acting, acrobatics, and martial arts, a screenshot of which is shown in Fig. 1. Cantonese opera is on the UNESCO Representative List of the Intangible Cultural Heritage of Humanity [4]. The opera is delivered in Cantonese [5], which is a Chinese language that sounds very different from Mandarin and is intelligible to less than 10% of the Chinese population. However, the written language is common between Cantonese and Mandarin to a large extent. Therefore, by highlighting the Chinese characters being sung or spoken in real time, it will greatly help promote Cantonese opera to a much wider audience who do not speak Cantonese.




Figure 1. Screenshot of a Cantonese opera video [6] featuring animated karaoke-style subtitles highlighting the Chinese characters being sung. Those that have been sung are filled in blue.

Internal use



In live operas, surtitles if available are usually shown above the stage. There are also personal titling systems that have become available recently. In any case, one or more lines of text, which must be prepared beforehand, are displayed encompassing the words being sung or spoken. The surtitle content is manually advanced by an operator who follows the performance, by a keystroke on the computer for example. An automatic surtitle display system with no human intervention is desirable. Furthermore, an enhanced audience experience can be brought about if the individual words in the surtitles can be animated in a karaoke style in sync with the performance like those shown in Fig. 1.

Little has been reported on the automatic generation and display of surtitles for live operas. A sentence-by-sentence lyrics alignment scheme for Cantonese popular music based on the tonal characteristic of the language in relation to the melody of the song was described in [7]. The method, however, is not applicable to individual Chinese character alignment in real time.

## 2 Proposed Surtitle Scheme

For the best enjoyment of the audience in live Cantonese opera, each monosyllabic Chinese character in the surtitle should ideally be highlighted at exactly the time when it is being sung or spoken. Hence, a key task of the proposed surtitle generation scheme is the automatic alignment of the Chinese characters to the vocalization events on the stage. Such events can be singing or speech, the former usually predominates in Cantonese opera. Furthermore, the scheme should take advantage of the regular beat structures in Cantonese opera. Even the speech tends to be rhythmic or accompanied by percussion instruments in many cases.

Given the above considerations, a full scheme is proposed for the automatic surtitle generation and alignment for live Cantonese opera based on beat tracking. A major advantage of the method is that the next Chinese character can be predictively highlighted at the onset of its vocalization with almost zero latency after compensating for the nominal digitizing and processing time of the streaming audio data.

In the first step, essential information like the sequence of the sung and spoken parts, their estimated tempos (in beats per minute) or speech rates (in characters per minute), the melodies of some of the intro and bridge music, along with all the sung and

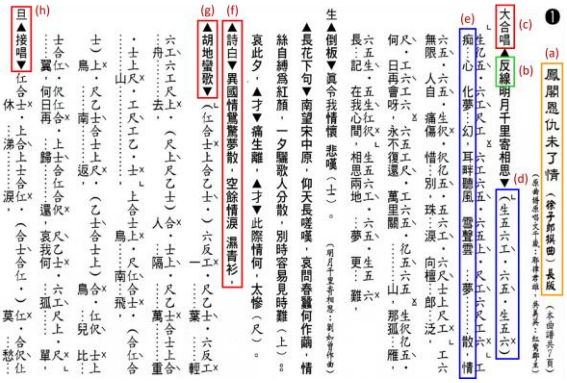



Figure 2. First page of the music-vocal score of the Cantonese opera “Romance of the Phoenix Chamber” (鳳閣恩仇未了情) in gongche notation [9]. (a) Title and composer, (b) mode/key information, (c) singing in unison, (d) musical note and beat information on the intro music, (e) lyrics, (f) male spoken line, and (g) a song that starts with male solo followed by (h) female solo.

spoken words are extracted from the music-vocal score. In Hong Kong and Macao, Cantonese opera scores are mainly available in the traditional gongche notation [3, 8]. Also known as gongchepu (工尺譜), an example of which is shown in Fig. 2 [9].

Next, for all vocalized events having a beat structure (this covers all the songs and several speech types in Cantonese opera), a regularly spaced pulse train that is an integral multiple of the beats and sufficient to register the onset position of every monosyllabic Chinese character to be sung or spoken is prepared from the score. Assuming beat tracking to the live performance is accurate, the surtitle character alignment should be 100% correct. Beat tracking is therefore a key part of the proposed scheme. In Cantonese operas, singing is normally preceded by a few seconds of intro or bridge. The music is often monophonic, which greatly facilitates its melody recognition by pitch tracking, and enables pulse train synchronization to the music before singing starts. However, the on-going tempo can vary according to the performers. There is therefore a need to realign the pulse train to the actual performance in real time. For this purpose, the beats are constantly monitored, and an adjustment mechanism is implemented on the pulse rate.

The remaining speech type has no beat structure and resembles ordinary conversation, the onset of which can be detected by voice activity detection (VAD) techniques. A pulse train is still generated in this case.

Internal use



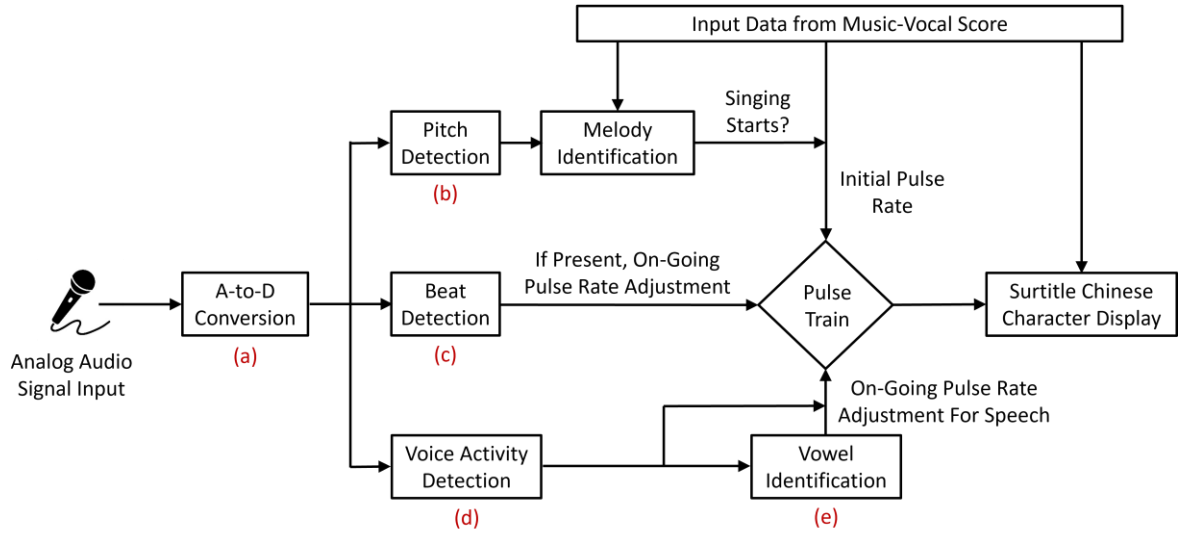



Figure 3. Block diagram of the proposed surtitle generation scheme. (a) Analog-to-digital conversion, *PyAudio* [10], (b) pitch detection, *Aubio* [11], (c) beat detection, *Librosa* [12], (d) voice activity detection, *Silero VAD* [13], and (e) vowel identification, *Wav2Vec2Phoneme* [14].

The spoken characters are then registered on the pulse train according to the cadence of the speech, sentence breaks, etc.

## 3 Implementation

The block diagram of the proposed surtitle generation scheme is shown in Fig. 3. The overall system is implemented in *Python.*

Audio signal recorded by the microphone is digitized by the analog-to-digital converter on the soundcard and streamed to *PyAudio* [10] at a sampling rate of 16 kHz in chunks of 0.4-s duration. The signal chunks are then sent to (a) *Aubio* [11] for pitch detection followed by melody identification, (b) *Librosa* [12] for beat detection, and (c) *Silero VAD* [13] for voice activity detection. The VAD output is further sent to *Wav2Vec2Phoneme* [14] for vowel identification.

```
little_book = [
    ['singing', 120,  # Initial pulse rate
    ['G', 'E', 'D',], 12, # Intro melody
    '痴', 1, '心', 2, '化', '夢', 1, '幻', 2, '耳', '畔', '聽',
    '風', 1, '雪', '聲', '愁', 2, '和', '夢', 3, '散', 5,
    '情', '無', '限', 2, '人', '自', 2, '痛', '傷', 1, '惜', 1, '別', 1,
    '珠', 1, '淚', 2, '向', '檀', '郎', 2, '泛', 7,
    '何', 2, '日', '再', '會', '呀', 2, '永', '不', '復', '還', 1,
    '萬', '里', '關', 5, '山', 2,
    '那', '孤', 1, '雁', 1, '長', 1, '記', 2, '在', '我', 1, '心', 1, '間', 4,
    '相', '思', '兩', '地', 3, '夢', 1, '更', 3, '難', 2, ''],

    ['speech', 130, # Initial pulse rate
    '異國', '情', '鴛', '鴦', '夢', '散', 5, ''],

    ['speech', 110,
    '空餘', '情', 1, '淚', 7, '濕', '青', '衫', 3, ''],
```

Figure 4. A portion of the input data showing coded information on a singing part and a speech part extracted from the music-vocal score of the Cantonese opera in Fig. 2.

The input data contains coded information on each singing and speech part. This includes the estimated pulse rate, the Chinese characters registered on the pulse train, and the musical notes of the intro or other musical cues. Figure 4 shows a portion of the input data extracted from the music-vocal score of the Cantonese opera in Fig. 2.

Central to the system is the pulse train generator, which can be started based on the melody identification or VAD outputs. The initial pulse rate is set by the input data. On-going pulse rate is then adjusted by beat tracking, VAD, or vowel identification.

The input data sends one or two lines of text to be displayed in the surtitle each time. The individual Chinese characters in it are then highlighted one by one according to their respective positions on the pulse train.

## 4 Results

The proposed surtitle generation scheme is tested on an online video of the Cantonese opera “Romance of the Phoenix Chamber” (鳳閣恩仇未了情) [15]. To

Internal use



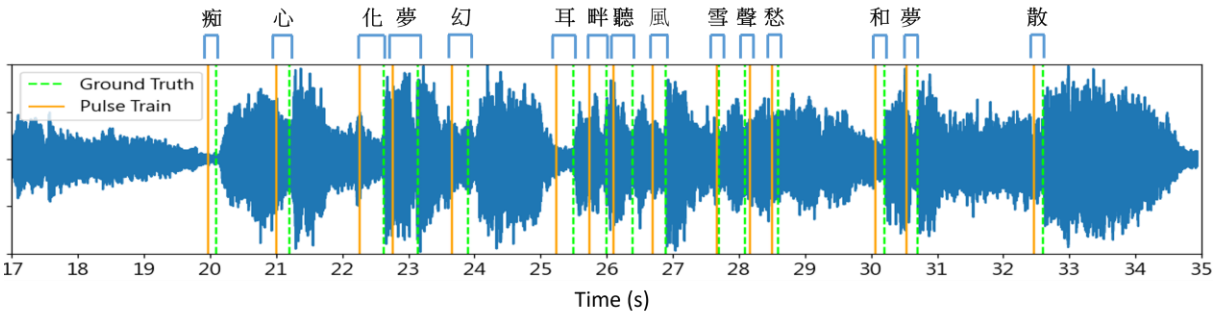



Figure 5. Audio waveform at the beginning of the unison singing part in the Cantonese opera "Romance of the Phoenix Chamber" (鳳閣恩仇未了情) marking the times at which the Chinese characters in the lyrics were highlighted by the proposed surtitle generation scheme (pulse train) vs. their respective onset times as perceived by the human ear (ground truth).

emulate a live opera situation, the audio output of the video player is rerouted to a virtual microphone through a virtual sound card before inputting to the surtitle system shown in Fig. 3.

The performance of the surtitle scheme was evaluated on the first four minutes of the opera video, which contains a chorus singing in unison, spoken lines, and male/female solo singing. Figure 5 shows the audio waveform at the beginning of the unison singing part. The times at which the Chinese characters in the lyrics were highlighted by the surtitle generation scheme are marked along with their respective onset times as perceived by the human ear (ground truth). Table 1 summarizes the timing precision of the proposed surtitle scheme. The percentage of surtitle Chinese characters that are highlighted within 0.5 s of the ground truth is 90% for singing and 71% for speech.

Two key factors are currently limiting the attainable precision of the surtitle generation scheme. The first factor is the limited responsiveness and accuracy of beat tracking, which mainly affects the singing parts. The second key factor is the lack of musical cues and rhythmic structures in some of the speech parts, which is exacerbated by the latency of a fraction of a second in VAD and up to a few seconds in vowel identification. As a result, it is sometimes difficult to predict when each sentence of a speech starts.

| | # Char | % Accurate to Within 0.5 s | Mean (s) | Stdev (s) |
|---|---|---|---|---|
| Singing | 142 | 90.1 | 0.00 | 0.33 |
| Speech | 14 | 71.4 | 0.23 | 0.39 |

Table 1. Timing precision of the proposed surtitle generation scheme evaluated on the first four minutes of the Cantonese opera "Romance of the Phoenix Chamber" (鳳閣恩仇未了情). Shown in the columns are: (a) number of surtitle Chinese characters, (b) percentage that are highlighted within 0.5 s of their respective onset times as perceived by the human ear, (c) mean, and (d) standard deviation of the lag time.

## 5 Discussion and Conclusions

An automatic surtitle generation scheme has been proposed for live Cantonese opera based on beat tracking. Reported in this work is the progress in its first implementation. While the various algorithms still need to be optimized, preliminary results are promising for a workable system to be realized in the future. In particular, the timing precision for singing is much better than that for speech. Both should continue to improve as better and faster modules for pitch tracking, beat tracking, VAD, and vowel identification become available.

Looking forward, since Chinese characters are monosyllabic in different dialects and the musical forms are similar, it is hoped that the present surtitle generation scheme can be extended to other regional Chinese operas like Peking opera, Chaozhou opera, and Kun Qu opera.

## Acknowledgment

The authors would like to thank the Undergraduate Research Opportunities Program (UROP) of The Hong Kong University of Science and Technology for supporting this work.

## References

[1] *Cantonese Opera*, Wikipedia. Accessed: May 12, 2024. [Online]. Available:

Internal use



https://en.wikipedia.org/wiki/Cantonese_opera

[2] *Cantonese Opera: History, Performance, Theatre and Costume Design*, South China Morning Post. Accessed: May 12, 2024. [Online]. Available: https://multimedia.scmp.com/infographics/culture/article/3036661/cantonese-opera/index.html

[3] C.-F. Wong, *Listening to Chinese music*, chapter 6, The Commercial Press (Hong Kong), 2009.

[4] *Yueju Opera*, Intangible Cultural Heritage, UNESCO. Accessed: May 12, 2024. [Online]. Available: https://ich.unesco.org/en/RL/yueju-opera-00203

[5] *Cantonese*, Wikipedia. Accessed: May 12, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Cantonese#:~:text=Cantonese%20is%20the%20predominant%20Chinese,official%20language%20in%20the%20world

[6] 艷曲醉周郎 黃偉坤 曹亞敏 (Sept. 17, 2017). Accessed: May 12, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=pYrTQ4PbcgM&list=PLTe4e4xDpQmlXdh6HklYpbVmosggyUwFp&index=14

[7] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic Lyrics Alignment for Cantonese Popular Music," *Multimedia Systems*, vol. 12, pp. 307–323, 2007. https://doi.org/10.1007/s00530-006-0055-8

[8] *Gongche Notation*, Wikipedia. Accessed: May 12, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Gongche_notation

[9] *Music-Vocal Score of "Romance of the Phoenix Chamber"* (鳳閣恩仇未了情), Sin Lok Music Art. Accessed: May 12, 2024. [Online]. Available: http://68893333.com/sllyric/Lyric/71416.pdf

[10] *PyAudio* (0.2.14). Accessed: May 12, 2024. [Online]. Available: https://people.csail.mit.edu/hubert/pyaudio

[11] *Aubio* (0.4.9), Zenodo. Accessed: May 12, 2024. [Online]. Available: https://aubio.org https://zenodo.org/records/2578765

[12] *Librosa* (0.10.1), Zenodo. Accessed: May 12, 2024. [Online]. Available: https://librosa.org https://zenodo.org/records/8252662

[13] *Silero VAD* (V4), GitHub. Accessed: May 12, 2024. [Online]. Available: https://github.com/snakers4/silero-vad

[14] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-Shot Cross-Lingual Phoneme Recognition," arXiv.org, 2021. https://doi.org/10.48550/arxiv.2109.11680

[15] *Romance of the Phoenix Chamber* (鳳閣恩仇未了情) (Aug. 8, 2011). Accessed: May 12, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=2ZoC9owD6OY