Automatic Lyrics Alignment for Cantonese Popular Music

Chi Hang WONG, Wai Man SZETO, and Kin Hong WONG
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{chwong1, wmszeto, khwong}@cse.cuhk.edu.hk

Appendix

A Experimental Results of Onset Detection

A.1 Results and Discussion

The onset detection algorithm was applied on all 70 segments in table 1 of section 8.1 with 75% window overlapping and three variables such as onset threshold ϵ^{onset}, onset window size w^{onset} and omitting window size w^{omit}. The following score function Ω^s was used to evaluate the onset detection algorithm:

$$\Omega^s = 1.0 - 0.65\frac{FN}{TP+FN} - 0.1\frac{FP}{TP+FN} \tag{28}$$

where TP (true positive) is the number of true onsets that the algorithm can detect, FN (false negative) is the number of true onsets that the algorithm failed to detect and FP (false positive) is the number of false onsets that the algorithm found. The higher the value of Ω^s the better is the result. For the evaluation, the onsets within 100ms of true onsets are considered as correct (TP).

The term $\frac{FN}{TP+FN}$ is the missing rate and the term $\frac{FP}{TP+FN}$ is the false alarm rate. The coefficients 0.65 and 0.1 of the score function are the weights which control the effect of the missing rate and the false alarm rate on the overall performance of our proposed system. They were derived from the the experiments of robustness of the DTW algorithm in section 8.2.3 as below. Figure 15 in section 8.2.3 shows the "In-Range Accuracy" and the Duration Accuracy after adding the spurious onsets. The accuracy dropped from 93% to 88%, 5% dropped after 50% spurious onsets, thus the DTW algorithm could compensate the false alarm errors which were introduced from the onset detection algorithm of the system. The weight of the false alarm rate in the score function was chosen as 0.1 because the "In-Range Accuracy" was dropped 5% when 50% spurious onsets were added. Figure 16 in section 8.2.3 shows the "In-Range Accuracy" and the Duration Accuracy after pruning the onsets. The accuracy dropped from 91% to 58%, 33% dropped after pruning 50% onsets, thus 0.65 was chosen as the weight of the missing rate in the score function.

Figure 17 in this appendix shows the onset detection result against onset threshold ϵ^{onset}, onset window size w^{onset} and omitting window size w^{omit}. The omitting window sizes w^{omit} with 100ms and 150ms (figures 17(b) and (c)) were better than that with 50ms and 200ms (figures 17(a) and (d)). In general, the time between two consecutive characters is between 100ms and 150ms, thus it is effective to use the omitting window size either 100ms and 150ms for pruning onsets that are too close.

From figures 17(b) and (c), the best onset window size (curve with circle) was 50ms. This result pretty much matched the same energy integration as human perception in [1], thus the experiment showed that the effective energy integration window was 50ms.

Figure 18 in this appendix shows the detailed result of the score with onset window size 50ms and omitting window size 150ms. The best threshold was within 0 and 0.1 and all the standard deviations of threshold 0.0-1.0 were about 0.07 which was small. Thus when the threshold is within 0 and 0.1, the result was similar.

Figure 19 in this appendix shows the detailed result of the false alarm rate with onset window size 50ms and omitting window size 150ms. The result showed the false alarm rate was around 50%-60% for small threshold values i.e. in 100 detected onsets, 50-60 detected onsets were wrong because the vocal enhancement method could not remove all the non-vocal instruments, thus there were many non-vocal onsets. In order to resolve the problem, the

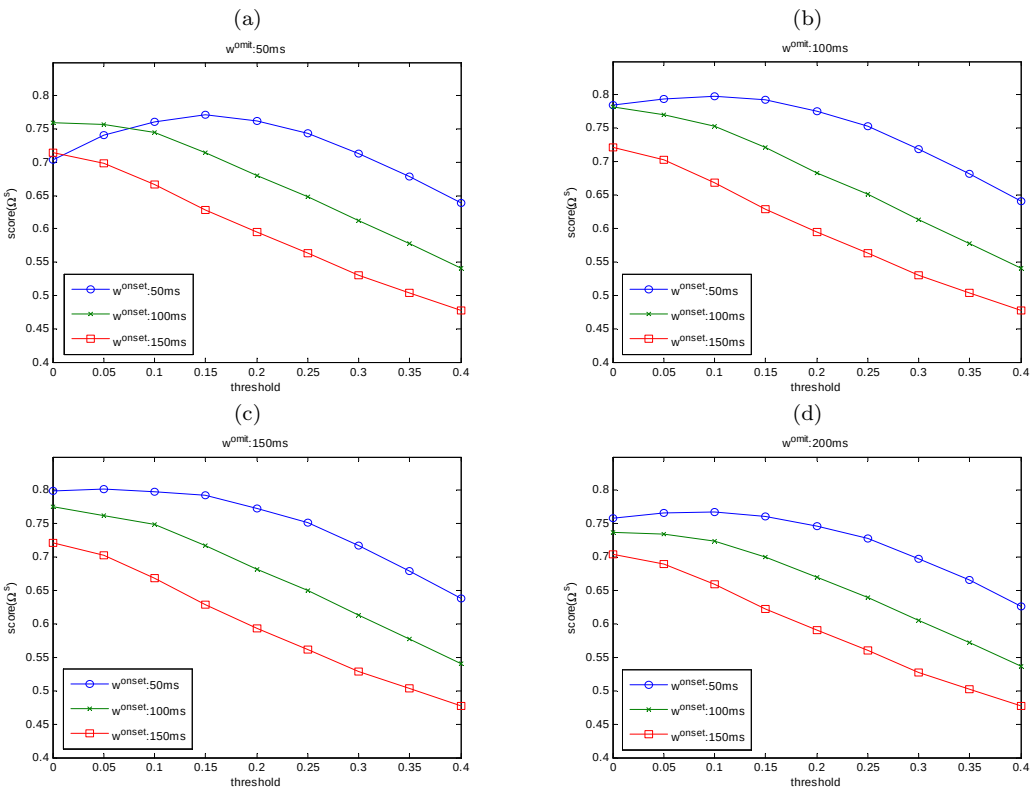

Figure 17: Onset detection result with different omitting window size w^{omit}: (a) 50ms, (b) 100ms, (c) 150ms and (d) 200ms.

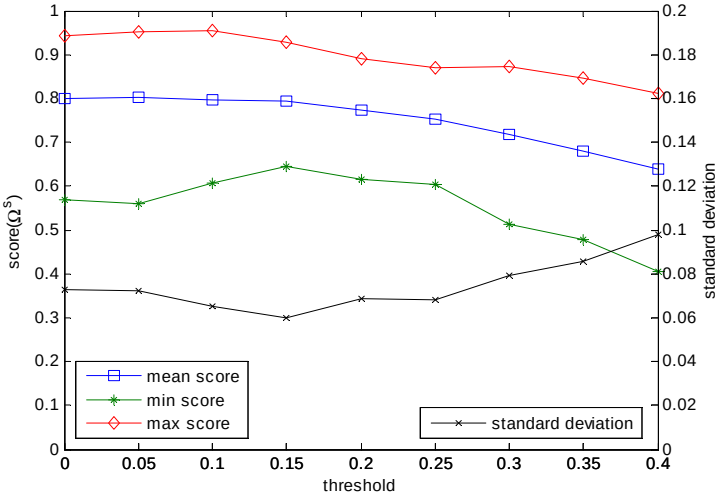

Figure 18: The mean (square), minimum (cross), maximum (diamond) scores and the standard deviation (σ) (cross) of onset detection for the omitting window size w^{omit}=150ms and the onset window size w^{onset}=50ms.

non-vocal pruning module was used.

In conclusion, our experimental results showed that the best omitting window size was 150ms. And also, onset window size 50ms and threshold 0.05 produced the best score for omitting window size 150ms.

References

[1] Eric D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of Acoustic Society of America*, (1):588–601, 1998.

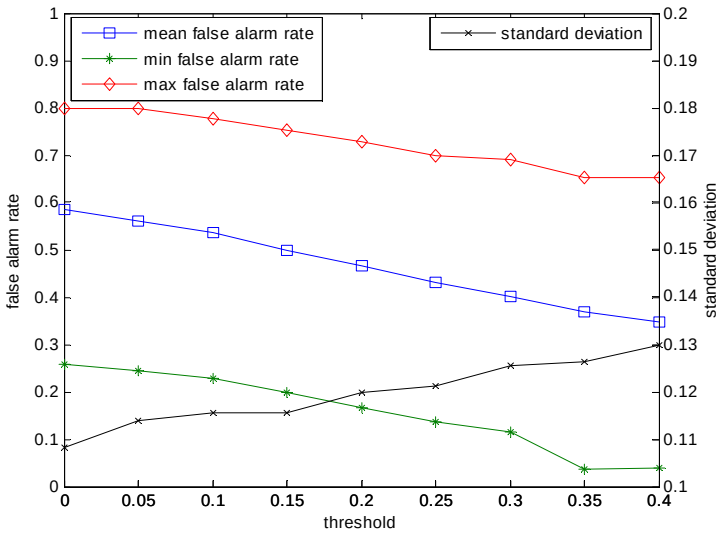

Figure 19: The mean (square), minimum (cross), maximum (diamond) false alarm rates and the standard deviation (σ) (cross) of onset detection for the omitting window size w^{omit}=150ms and the onset window size w^{onset}=50ms.