Subjective Evaluation of Music Similarity System Based on Onsets

Shingchern D. You and Ho-Hsiang Shih

Department of Computer Science and Information Engineering
National Taipei University of Technology
Taipei, Taiwan
E-mail: you@csie.ntut.edu.tw

Abstract

We have proposed a music recommendation system based on the time difference between adjacent onsets previously. The system also employed the rough longest common subsequence (RLCS) algorithm for similarity measurement. In this paper, we report our results for subjective (listening) comparison of our system with the Musly system. The results show that, two out of five query songs, the audiences felt that the recommended clips from the proposed system are more similar to queries. Since our system uses only "onsets" for similarity measure, as opposed to both temporal and spectral features in other systems, to determine the similarity of audio clips the subjective results confirm that onset information is an important factor to determine the similarity of two audio clips. Thus, a music recommendation system might include the onset features for computing similarity scores.

Key words: Onset, RLCS, Music similarity

Introduction

Previously, we have proposed a system to evaluate the similarity of music clips [1]. Such a system can be used to recommend similar audio clips in the database if a preferred clip is given. The proposed system uses the (time) difference between adjacent onsets as bases to compute similarity. In the system, we employed four different onset detection methods, including temporal and spectral detection methods, to better detect onsets. Onsets detected from each method were individually scored with the rough longest common subsequence (RLCS) algorithm [2]. The final score is a weighted sum of individual scores from each detection method.

In the paper of [1], we did not conduct subjective (listening) tests to compare our system with other systems. To fully explore the capabilities and limitations of the proposed system, we further conduct four experiments. The first experiment is to compare the onset detection accuracy of the proposed approach versus other approach, such as Madmom [3] and Librosa [4]. The second experiment is to check the consistency of the detected onsets for different detection methods. The third experiment is to know whether the onset detection accuracy is a significant factor in similarity comparison. The final experiment is a subjective test to compare the proposed system with a commercial system called Musly [5]. Through the experiments, we can better assess the performance of the proposed system.

This paper is organized as follows. Section one is the introduction. It is followed by a brief description of the proposed system. Next, section three covers the detailed steps of the four experiments and results. And, finally, section four is the conclusion.

The Proposed System

The block diagram of the proposed system is depicted in Fig. 1. In the system, we employ four different onset detection methods, namely, the temporal method [6,7], the high-frequency component (HFC) method [8,9], the spectral difference method [10], and the up-count method [11]. The first method detects onsets through the majority votes based on the outputs of four bandpass filters. The second method uses the difference of total weighted spectral energy between two successive frames to detect the onsets. The third method also uses spectral coefficients for onset detection. However, this method computes onsets by using the differences of individual spectral energy, instead of the total energy, over successive frames. The fourth method is also a spectral-based method. This method, to reduce the influence of noise, counts only the number spectral coefficients which are increasing over two frames. By utilizing four different onset detection methods, we hope that the system is more robust.

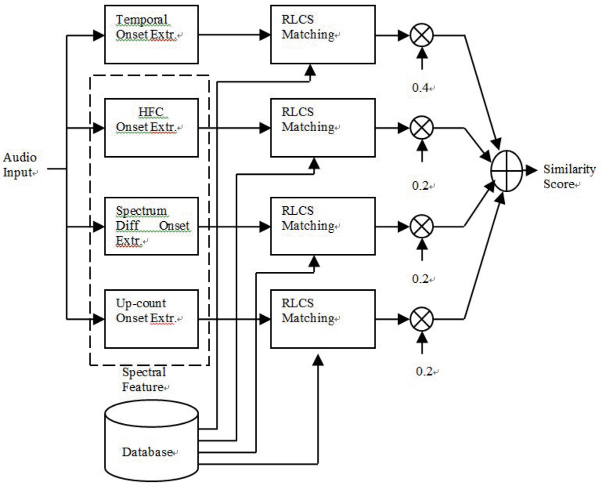

Fig. 1. The block diagram of the proposed system [1].

Since the query clip can start from any point in a soundtrack, we cannot use the absolute time indices. Thus, the time difference between two adjacent onsets is used as an element in a time series. With the four detection methods, we have four time series of onset information. The overall score is a weighted combination of the individual RLCS matching scores. We use RLCS, instead of longest common subsequence

(LCS) algorithm is that spurious or missing onsets occurs frequently, and the RLCS algorithm handles this problem better than the LCS algorithm. For more details about the proposed system, the reader is referred to [1].

Experiments and Results

As mentioned previously, we conducted four experiments to assess the performance of the proposed approach. The following describes the experiments and results.

A. Accuracy of Detected Onsets

The first experiment is to compare the onset detection accuracy of the methods used in the proposed system versus other widely recognized methods, namely, Madmom [3] and Librosa [4]. The authors of Madmom also release a dataset with annotations for onset detection training and testing [12]. Therefore, in the experiment, we randomly choose 10 audio clips from that dataset for comparison. Since we may encounter a miss or a false alarm when detecting onsets, we decide to use the F-measure [13] to evaluate the detection accuracy. In the measurement, if the position of the detected onset is within ± 25 ms to a "true" (from ground truth) onset position, this onset is considered as detected.

The scores of the F-measure of the detection approaches are listed in Table 1. We know from Table 1 that the Madmom approach is better than the rest ones, including Librosa. However, as the dataset is also provided by the authors of the Madmom, it is unclear whether the Madmom approach uses the dataset to fine-tune the Madmom meta-parameters. Consequently, it is not easy to say whether this comparison is biased or not. Nevertheless, the methods used in the proposed system, when compared with Librosa, do not have satisfactory F-measure. We also surprisingly note that the HFC method performs much worse than the other two spectral-based methods. The source to this situation has not been fully studied by us.

TABLE I
RESULTS FOR EXPERIMENT I

Song Title	Madmom	Librosa	Temp	HFC	SD	UC
Classics4-01	0.81	0.48	0.47	0.34	0.36	0.39
Classics4-11	0.73	0.78	0.43	0.50	0.42	0.46
Magic-04	0.83	0.42	0.24	0.43	0.42	0.29
Magic-09	0.70	0.67	0.47	0.58	0.51	0.51
Paradiso-05	0. 90	0.86	0.79	0.72	0.31	0.76
Paradiso-07	0.97	0.75	0.58	0.64	0.38	0.72
Chrisanne1-01	0.84	0.81	0.54	0.65	0.52	0.66
Chrisanne1-08	0.82	0.74	0.30	0.53	0.40	0.69
Chrisanne2-01	0.84	0.69	0.48	0.45	0.37	0.43
Chrisanne3-02	0.91	0.89	0.84	0.56	0.24	0.65
Average	0.84	0.71	0.51	0.54	0.39	0.55

B. Consistency of the Onset Detection

Although the onset detection methods used in the proposed system have lower accuracy, this situation may not be a significant shortcoming if the detected onsets have high consistency. Consider the following case: Two audio clips are conceptually "very similar" fi they are from the same musical score but performed by different artists, as in the case of cover version soundtracks. Recall that onsets are closely related to musical notes. Therefore, these two clips should have very similar onset patterns. If our methods also produce similar patterns for these two audio clips, the system will give high scores for these two clips regardless whether the detected onsets are real onsets or not.

To this end, we conduct the second experiment. In this experiment, we find an audio soundtrack which has two segments with almost identical melody. We then compare the detected onsets from the first segment and the second segment. Specifically, we use the obtained onset series in the first segment as "ground truth" to compute the F-measure of the onsets from the second segment. The results are given in Table II. It can be seen that the methods in the proposed system has satisfactory consistency results, especially the temporal method.

TABLE II
RESULTS FOR EXPERIMENT II

	Madmom	Librosa	Temp	HFC	SD	UC
F-measure	0.93	0.88	0.87	0.63	0.78	0.78

C. Comparison with Alternative Onset-based System

To know whether the lower onset accuracy will significantly affect the recommended candidates, we conduct the third experiment. In this experiment, we build a second system by using Librosa as the onset detection routine in conjunction with the RLCS scoring subsystem. The dataset has around 2,000 audio clips with various genres, such as classic, soft music, instrument solo, and popular music. Each audio clip has duration of 30 seconds with sample rate of 44,100 s/s. The clips are down-sampled to 11,025 s/s before sending to both systems for onset detection. Both systems are then given with five queries (exemplar audio clips) to find most similar clips in a dataset. Ten candidate clips are obtained by each of the systems from one query.

The experimental results are given in Table III, where the format of the number of matches is denoted as "*x*+*y*." In the notation, *x* represents the number of same song titles recommended by both systems (in the candidate list of 10 songs). If the song titles from both systems are different, but from the same album, *y* represents the number of such songs. For example, "2+1" means that out of the list of 10 recommended titles from each system, two of the song titles appear in both lists. In addition, two different song titles, one from each system, are from the same album. From the experimental results we know tthat both systems recommend many identical audio clips. Thus, it seems that the "actual" onset detection capability is not a critical condition, as long as the detection method produced "consistent" onset series.

From the viewpoint of probability, if both systems "randomly" recommend 10 song titles out of 2000, the chance that no identical song titles in both lists is more than 95%. Thus, to have results similar to the one we have in Table III, the chance is actually very small, estimated to be less than 0.003 %. Thus, although both lists are not identical, they are highly correlated.

ISBN 978-1-5386-4342-6

TABLE III
RESULTS FOR EXPERIMENT III

Query	#1	#2	#3	#4	#5
Match	6+0	1+1	3+0	0+1	5+0

D. Subjective Test

The final experiment is to compare the proposed system with a commercial system called Musly [3]. In this experiment, 15 graduate students are participated in the subjective test to pick the most similar clip among candidate clips. Before the test, the listeners were asked not to emphasize on any particular similarity criteria. Instead, they were asked to pick the clip based on their intuition or their own similarity preference. In this experiment, five query audio clips are used as the inputs to both systems. For each query, three most similar audio clips are received from each system. The subjective test consists of two steps: In the first step, the listeners are asked to choose one audio clip out of three candidate clips recommended from each system for each query. After this step, each query has one similar clip left from each system. In the second step, the listeners are asked to judge, between the clips recommended for the proposed system and the Musly, which one is more similar to the query clip. The results, given in Table IV, are that two out of five queries, candidates from the proposed system are favored by the listeners.

TABLE IV
RESULTS FOR SUBJECTIVE TEST

Song index	Musly	Proposed
#1	2	11
#2	13	2
#3	14	1
#4	12	3
#5	1	12

In order to have a better understanding the reasons behind listener’s judgement, they are asked to choose *all* appropriate reasons, including genre similarity, melodic similarity, rhythmic (or tempo speed) similarity, accompaniment instrument similarity, same singer (or similar vocal tone), etc. Therefore, it is possible that one listener chooses more than one reason for a particular clip. A summary of the reasons is given in Table V. It is observed that for song index #1 and #5 rhythmic similarity is indeed an important factor for the listeners to make decisions. Basically, among two candidates *A* and *B*, if *A* is similar to the query in rhythm and *B* is not, many listeners prefer *A*. If none of these has similar rhythm, other factors then dominate. Overall, this experiment shows that the rhythmic similarity is indeed a crucial factor for human beings to determine whether two audio clips are similar or not.

TABLE V
REASONS FOR SUBJECTIVE TEST

Index	Genre	Melody	Rhythm	Instrument	Singer	Other
#1	6	1	7	0	1	1
#2	5	3	2	1	5	1
#3	2	2	3	10	0	2
#4	6	6	1	0	0	4
#5	5	1	9	1	0	0

Conclusion

This paper presents the evaluation results of the proposed music similarity evaluation system. The proposed system does not have very high onset detection accuracy. However, in terms of consistency, the system performs satisfactory. In contrast to other recommendation system utilizing both spectral and temporal features, the proposed system only relies on temporal features, i.e., onsets. Despite this fact, the recommended clips from the proposed approach are favored by the audiences in two out of five queries. With this subjective experiment, we conclude that a system to recommend similar music from a query should seriously consider including the onset similarity as a portion of the scoring system.

Acknowledgement

This work was supported in part by the Ministry of Science and Technology (MOST) of Taiwan through the grant number MOST 105-2221-E-027-109, which is gratefully acknowledged.

References

[1] S. D. You and R-W Chao, “Music similarity evaluation based on onsets,” *Proceedings in Lecture Notes in Electrical Engineering*, vol. 422, pp. 153-163, Sep. 2017.

[2] H.-J. Lin, H.-H. Wu, C.-W. Wang, “Music matching based on rough longest common subsequence,” *J. Info. Sci. Eng.*, vol. 27, no. 1, pp. 95 – 110, 2011.

[3] Madmom: A new Python audio and music signal processing library, available at http://www.cp.jku.at/research/papers/Boeck_etal_ACMMM_2016.pdf

[4] Libroas, available at https://librosa.github.io/

[5] Audio music similarity for both open-source and commercial versions, available at http://www.musly.org/

[6] A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” *Proc. 1999 IEEE International Conference on. Acoustics, Speech, and Signal Processing*, vol. 6, 1999.

[7] J. Ricard, “An implementation of multi-band onset detection,” in Proc. 1st Annu. *Music Inf. Retrieval Evaluation eXchange (MIREX)*, pp. 1- 4, 2005.

[8] P. Masri and A. Bateman. "Improved modeling of attack transients in music analysis-resynthesis." *Proceedings of the International Computer Music Conference*, pp. 100 – 103, 1996.

[9] K. Jensen and T. H. Andersen, “Real time beat estimation using feature extraction,” In *Proc. Computer Music Modeling and Retrieval Symposium*, Lecture Notes in Computer Science. Springer Verlag, 2003.

[10] C. Duxbury, M. Sandler, and M. Davies. “A hybrid approach to musical note onset detection,” *Proc. Digital Audio Effects Conf. (DAFX,’02)*, pp. 33- 38, 2002.

[11] Hsin-Jung Huang, *A Study on Note Detection and Melody Matching Method for Query By Singing/Humming System*, Master Thesis, Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan, 2009 (in Chinese).

[12] https://github.com/CPJKU/onset_db

[13] https://en.wikipedia.org/wiki/F1_score