Identifying Cover Songs from Audio Using Harmonic Representation

Kyogu Lee
Center for Computer Research in Music and Acoustics
Department of Music, Stanford University
kglee@ccrma.stanford.edu

Abstract

This extended abstract describes in detail a submission to the task on Audio Cover Song in the Music Information Retrieval eXchange in 2006. The system uses as feature set a chord sequence identified by an HMM trained with audio-from-symbolic data, and computes a distance between two chord sequence pair using the Dynamic Time Warping algorithm to find the minimum alignment cost. The rational behind the system is that cover songs largely preserve harmonic content even if they vary in other musical attributes such as instrumentation, tempo, key, and/or melody.

Keywords: MIREX, Cover Song, Chord Sequence, Dynamic Time Warping

1. Introduction

A cover song is defined as a song performed by an artist different from the original artist [1]. Identifying cover songs given an original as a seed/query or finding the original given a cover version from the raw audio is a challenging task, and it has recently drawn attention in a Music Information Retrieval society. Cover songs are different from its original in terms of many musical attributes such as duration, tempo, dynamics, instrumentation, timbre, or even genre. Therefore, the raw audio in the time-domain or its frequency-domain representation like spectrogram is very different from each other. Such diversity found in cover songs requires a robust feature set that remains largely unchanged under various musical changes mentioned above.

Harmonic progression is a robust mid-level representation that is largely preserved under such musical variations. While other musical details such as melody, tempo, and/or timbre may vary from one to another, their harmonic progression over time undergoes minor changes compared with the others.

2. System Overview

Our system consists of two main blocks – (1) feature set is first extracted from the raw audio, and (2) distance measures are computed between a song pair based on the extracted feature set. Figure1 illustrates the overview of the system.

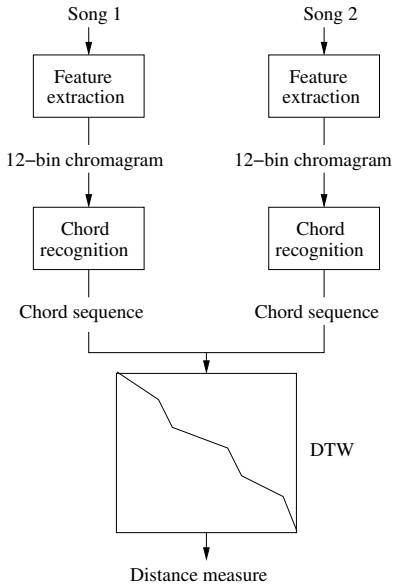

Figure 1. System overview

2.1. Chord recognition

Automatic chord recognition algorithm is decomposed of two parts. The first part computes a quantized 12-bin chromagram from the raw audio [1]. After a chromagram is obtained, automatic chord recognition algorithm based on the HMM is applied to get the chord sequence [2, 3], which has just one value per frame. Figure2 shows the overview of the automatic chord recognition system.

Transposition from one key to another is not rare in cover songs, and it may cause a serious problem in computing the distance because chord-to-chord distance becomes larger even though relative chord progression between the two sequences might be alike. To avoid this problem, key identification must precede. Instead of designing a sophisticated key identification algorithm, we simply estimated the key of a song to be the most frequent chord in the chord sequence, and transposed every song to a C major key before sending it to a distance computing algorithm (algo1). As an alternative, we also used the first chord to be the key (algo2).

2.2. Distance computing

[1] http://www.secondhandsongs.com/wiki/Guidelines/Cover

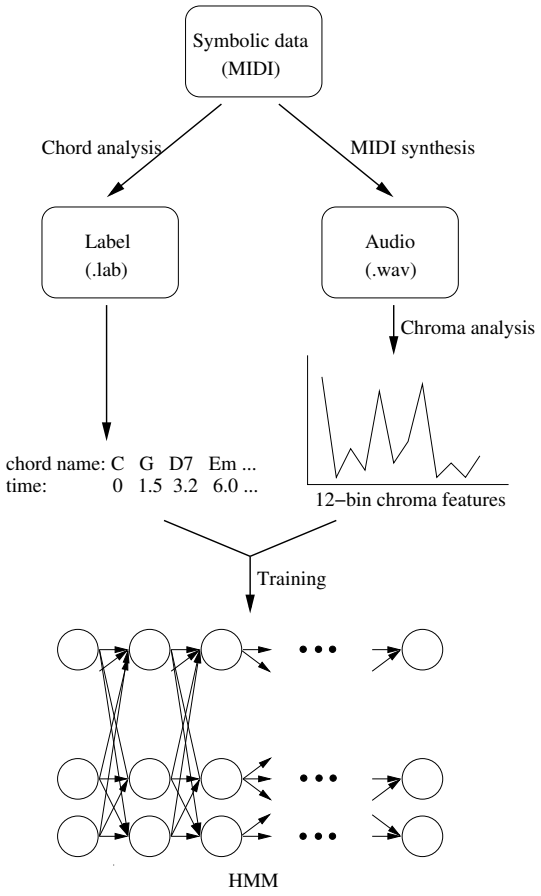

Figure 2. Overview of the chord recognition system.

After frame-level chord sequence is obtained for each song, we used the dynamic time warping algorithm (DTW) to find the minimum alignment cost between the two songs. DTW algorithm has been successfully used in the automatic speech recognition system to identify a word by finding the minimum-cost path between the input word and the word templates.

In order to use the DTW algorithm, we first need to obtain two-dimensional cost matrix from the two inputs. We defined a cost of being in aligned states by computing the chord-to-chord distance from the HMM parameters obtained above. In addition, we also defined a transition cost from the transition probability matrix in the HMM. Therefore, a combined cost at time step k for input a_k and b_k is given by

$$d(a_k, b_k) = d_S(a_k, b_k) + d_T(a_k, b_k | a_{k-1}, b_{k-1}), \qquad (1)$$

where d_S is a cost of being in aligned states, and d_T is a transition cost from the previous states to the current states. The total cost of alignment between the two sequences a and b is then given by

$$D(a, b) = \sum_{k=1}^{K} d(a_k, b_k). \qquad (2)$$

Figure3 displays the examples of the DTW for a cover-pair (on the left) and for a non-cover pair (on the right).

3. Results and analysis

3.1. Test material

Test data was composed of 30 queries with each query having 11 different cover versions including themselves. Therefore, total collection contains 30x11 = 330 songs. The collection includes a wide range of music from classical to hard rock.

3.2. Evaluation

Eight algorithms including the two by the author were submitted to the MIREX task on cover song identification. Four measures were used to evaluate the performance of the algorithms – (1) total number of cover songs identified; (2) mean number of cover songs identified; (3) mean of maxima; and (4) Mean reciprocal rank (MRR) of first correctly identified cover. Table1 shows the results using these measures.

As shown in Table1, two algorithms described in this paper are ranked at 2nd and 3rd places, respectively, using all four measures. Raw results reveal that some songs are difficult to identify for all systems while other songs are system-specific. In addition, the top four algorithms were specifically designed only for cover song identification whereas the bottom four were originally used in the similarity finding task as well. This proves that the two tasks are quite different from each other.

4. Conclusions

We proposed a system that identifies the cover songs from the raw audio using harmonic representation as a feature set and the dynamic time warping algorithm to score an alignment between the two songs abstracted through chord sequences. The rationale behind this idea was harmonic content would remain largely intact under various acoustical changes found in different versions of cover songs.

To this end, we first extracted a chord sequence from the chromagram at the frame rate using the HMM, and anchored the whole sequence to the most frequent chord or to the first chord to avoid the problem of transposition of keys. We then used the dynamic time warping algorithm to find the minimum alignment cost between a pair of chord sequences. In computing the total cost, we not only used a cost of being in aligned states but also a transition cost from chord to chord to reflect the theory of harmonic progression in Western tonal music.

Although we used our system only to recognize the cover songs, we believe it can be also used to find musical similarity since cover songs are extreme examples of similar music. Therefore, even if some songs which appear high in the list are not relevant, they might be evaluated similar to the query by human subjects, especially harmonic content is a key criterion in evaluating musical similarity.

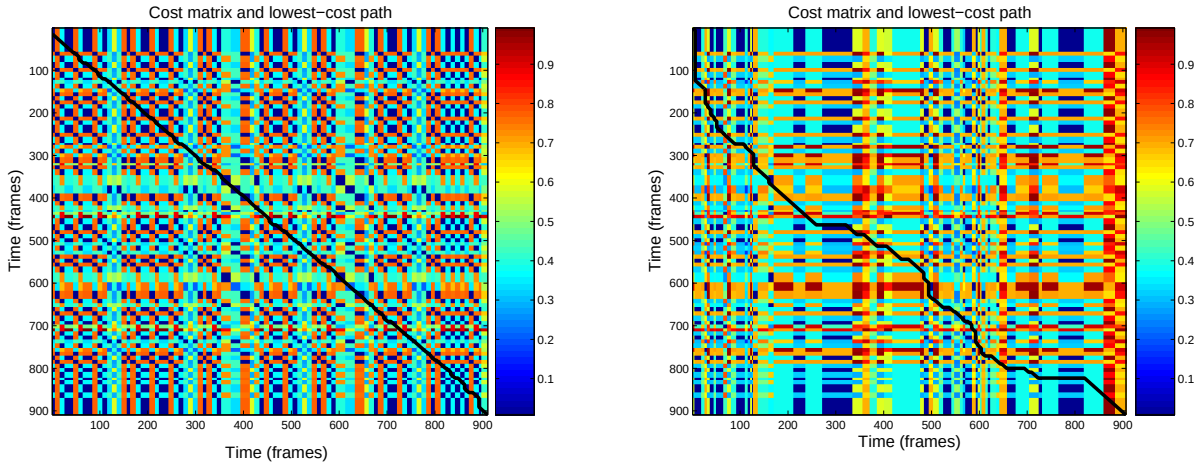

Figure 3. Examples of dynamic time warping.

Table 1. Summary results of eight algorithms.

Measure	Total number of covers identified	Mean number of covers identified	Mean of maxima	MRR of first correctly identified cover
1	D.P.W. Ellis (761)	D.P.W. Ellis (2.31)	D.P.W. Ellis (4.53)	D.P.W. Ellis (0.49)
2	**K. Lee [1] (365)**	**K. Lee [1] (1.11)**	**K. Lee [1] (2.50)**	**K. Lee [1] (0.22)**
3	**K. Lee [2] (314)**	**K. Lee [2] (0.95)**	**K. Lee [2] (2.27)**	**K. Lee [2] (0.22)**
4	Sailer & Dressler (211)	Sailer & Dressler (0.64)	Sailer & Dressler (2.13)	Sailer & Dressler (0.21)
5	Lidy & Rauber (149)	Lidy & Rauber (0.45)	Lidy & Rauber (1.57)	Lidy & Rauber (0.12)
6	K. West [1] (117)	K. West [1] (0.35)	T. Pohle (1.50)	K. West [1] (0.10)
7	T. Pohle (116)	T. Pohle (0.35)	K. West [1] (1.30)	K. West [1] (0.10)
8	K. West [2] (102)	K. West [2] (0.31)	K. West [2] (1.23)	T. Pohle (0.09)

In the future, we plan to include a melodic description in the feature set, which is another robust musical attribute that doesn't change much from cover to cover. In addition, we believe that applying the DTW to the most representative part of music will help not only increase the identification performance but also decrease the computation time to a great degree.

References

[1] C. A. Harte and M. B. Sandler, "Automatic chord identification using a quantised chromagram," in *Proceedings of the Audio Engineering Society*. Spain: Audio Engineering Society, 2005.

[2] K. Lee and M. Slaney, "Automatic chord recognition using an HMM with supervised learning," in *Proceedings of the International Symposium on Music Information Retrieval*, Victoria, Canada, 2006.

[3] ——, "Automatic chord recognition from audio using a supervised HMM trained with audio-from-symbolic data," in *Audio and Music Computing for Multimedia Workshop*, 2006.