Evaluation of Chord and Chroma Features and Dynamic Time Warping Scores on Cover Song Identification Task

Ladislav Maršík[3](✉), Martin Rusek[1], Kateřina Slaninová[1,2], Jan Martinovič[1,2], and Jaroslav Pokorný[3]

[1] IT4Innovations, VŠB - Technical University of Ostrava,
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic
{martin.rusek,katerina.slaninova,jan.martinovic}@vsb.cz

[2] Department of Computer Science, FEECS, VŠB - Technical University of Ostrava,
17. listopadu 15/2172, 708 33 Ostrava-Poruba, Czech Republic

[3] Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University, Malostranské nám. 25, Prague, Czech Republic
{marsik,pokorny}@ksi.mff.cuni.cz

Abstract. Cover song identification has been a popular task within music information retrieval in the 20th century. The task is to identify a different version or performance of a previously recorded song. Unlike audio search for an exact matching song, this task has not yet been popularized among users, due to an ambiguous definition of a cover song and the complexity of the problem. With a great variety of methods proposed on the benchmarking challenges, it is increasingly difficult to compare advantages and disadvantages of the features and algorithms. We provide a comparison of three levels of feature extraction (chroma features, chroma vector distances, chord distances) and show how each level affects the results. We further distinguish five scores for dynamic time warping method, to find the best performance in conjunction with the features. Results were evaluated on covers80 and SecondHandSongs datasets and compared to the state-of-the-art.

Keywords: Music information retrieval · Chroma features · Chord distance · Chroma vector distance · Cover song identification · Dynamic time warping

1 Introduction

A cover song is an alternative version, performance, or recording of a previously published musical piece. The author of a cover song may choose to differ from the original piece in several musical aspects: instrumentation and arrangement, tonality and harmony, song structure, melody, tempo, lyrics, or language. Thus, Cover Song Identification (CSI) task is a difficult challenge in choosing the

K. Saeed et al. (Eds.): CISIM 2017, LNCS 10244, pp. 205–217, 2017.
DOI: 10.1007/978-3-319-59105-6_18

best techniques, features and algorithms, and has been a vivid field of research within music information retrieval in the last decade. The state-of-the-art methods are evaluated annually on the benchmarking challenge MIREX (Music Information Retrieval Evaluation Exchange)[1] with up to 8 algorithms posted every year since 2006.

Musicians often use *chord progressions* (a sequence of chord labels) when replicating a musical piece originating from another author. Chord progressions and other harmony features contain deeper underlying music information, which is typically retained in cover songs, while tempo, style or instrumentation changes [7]. The common technique to obtain these features is to use the Discrete-Time Fourier transform (DFT), resulting in so-called *chroma features* [1], from which chords are derived by further processing. Both chroma features and chord progressions are commonly used for CSI [3,13,22].

The recent CSI studies focus on finding a proper harmony fingerprint of the song, to be able to work with large music databases [4,13,19]. Such fingerprint should reduce the complex information from the audio, keeping only the relevant harmony movements. In the light of the recent proposals [6,21], *chord distances* can be used in addition to the chord labels, to extract meaningful harmony features. If used properly, chord distances can help with the problem of a *key transposition* [12], when the cover song is played in a different key from the original.

To understand the concept of a chord distance, we can imagine chords as points in space, with assigned distances (forming a chord distance model). One of the straightforward applications of this model could be to treat the music as a path in this space. In our work we test this idea: we derive a chord distance from every pair of successive chords to obtain a time series descriptor, which we then use for CSI. Furthermore, we apply the same idea also for chroma features, by defining *chroma vector distances* as another time series for comparison. Lastly, we are aware of the danger of losing relevant information by extracting these features. Our study therefore compares the fingerprints with the lower-level *"raw" chroma vectors*, from which the fingerprints were extracted.

To determine a similarity between the time series, it is convenient to use Dynamic Time Warping (DTW), which has long been used in signal processing [18,22]. In our work we are interested not only in the different levels of feature extraction, but also in the comparison of five different DTW scores. As such, our task is to provide leads for the best feature-score combination for CSI.

We continue in Sect. 2 with the survey of CSI works most related to us. Then in Sect. 3 we provide the insight into the music harmony features that we use. In Sect. 4 we overview the DTW algorithm and the different DTW scores. In Sect. 5 we show our results on covers80[2] and SecondHandSongs[3] datasets and compare with the state-of-the-art results. We wrap up in Sect. 6 with the conclusion and our future work.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME.
[2] https://labrosa.ee.columbia.edu/projects/coversongs/covers80.
[3] https://labrosa.ee.columbia.edu/millionsong/secondhand.

2 Related Work

CSI task has been added to MIREX benchmarking in 2006, which was the first time that the task was given an academic attention. Now it is one of over fifteen annually evaluated tasks, including Audio Melody Extraction, Audio Chord Estimation, and others. The LabROSA system achieved the best results in the first year of the CSI benchmarking, with beat-synchronous chroma features [7]. Chroma vectors were calculated for beat-length segments of the songs, where beats were extracted by preprocessing. The result of 34 correctly identified covers from covers80 dataset was a promising start, proving that harmony features such as chroma vectors are meaningful for CSI task.

The winning submission in the 2007 benchmarking was a system developed by Serra et al. [22]. Based on an extensive series of experiments, the authors have proposed improvements, notably: the use of improved chroma features (HPCP [12], with an improved resolution of 36 bins), tackling the key transposition problem using a shift based on the global song profile, binary similarity matrix, dynamic programming and a local alignment (Smith-Waterman algorithm). DTW with various constrains was used for the similarity measures.

The best results in MIREX at the time of writing this paper were achieved in 2009 by an improved version of [22] (Serra et al. [23]). The system correctly identified 2426 out of 3300 cover songs (73.5%). The most important improvements were: using a cross-recurrence plot instead of the binary similarity matrix, and reasoning that a maximal length of diagonal lines (L_{max}) experimentally proved to have the highest discriminative power. The final measure (Q_{max}) was a slight modification of L_{max}, accounting for the tempo changes (the curvature of the line) and gaps.

The idea that the harmonic content should be invariant over the cover songs is a centerpiece of most of the CSI works. While some authors used chroma vectors (or their variants) to capture the harmonic content, other experimented with using chord sequences. Lee [15] used Hidden Markov Models to extract chords from chroma vectors and then compared the chord sequences using DTW. Bello [2] and Robine et al. [20] have studied string alignment methods. The chord-based algorithms were outperformed by the above mentioned chroma-based systems on MIREX, however, a full potential of chords has not been explored yet, especially with the possibility of evaluating chord distances for the progression.

The most recent trend in CSI is to pioneer new ways of feature extraction (2D Fourier Transform Magnitude [4], usage of MFCC [24]), new similarity measures (Shannon information [10]), or database pruning and combining various features based on machine learning [19]). The focus is shifting to the large-scale methods rather than outperforming the results on the smaller benchmarking datasets.

3 Chord and Chroma Features

3.1 Chroma Features

Chroma features are commonly referring to a series of 12-dimensional vectors of floating-point numbers, capturing the presence of each tone in a short music moment [12]. The concept has been proposed and studied at the beginning of 21st century by multiple authors; we first became aware of it in the work of Bartsch and Wakefield [1] and Fujishima [11]. Chroma features are obtained directly from DFT output by grouping frequencies that belong together in one frequency bin, where the bins are set by the tempered tuning of the piano scale. The resulting vector has the form:

$$< c_A, c_{A\#}, c_B, c_C, c_{C\#}, c_D, c_{D\#}, c_E, c_F, c_{F\#}, c_G, c_{G\#} >$$

where $c_A \in \mathbb{R}$ represents the presence of the A tone, $c_{A\#} \in \mathbb{R}$ represents the presence of the $A\#$ tone, etc. The value distribution of $c_A, c_{A\#}, \ldots$ depends on the algorithm used, but it is a common practice to normalize to $[0, 1]$ interval, where the loudness ratio is preserved in between the tones. For more information on chroma features, we refer the reader to Bartsch and Wakefield [1].

Since chroma features have long been used for CSI, we choose them as our first (and reference) feature for comparison. To achieve a proper search for a cover song transposed to a different key, we perform a circular shift to all 12 possible transpositions of the song before the comparison, as was done in [8] or [13].

3.2 Chord Distances

The use of chord distances for CSI is our first innovation: we form a time series of chord distances as a reasonably small fingerprint for the work with large music datasets. Chord distances are a subject of the recent research in music information retrieval, but the concept of a chord distance still remains ambiguous, as there are many definitions to choose from [21]. As a result, further study on how all of these distances may be used for CSI, has not been executed, to the best of our knowledge.

Thanks to Rocher et al., we have a thorough review available of the common chord distances, including musicology, geometric or computational approaches [21]. Nine chord distances are compared, and tested on the chord estimation task, with Tonal Pitch Space (TPS) distance [6] performing the best for the root and mode estimation.

In this paper our goal is not to compare the various chord distances available, but rather to choose a representative distance, that can best suite the CSI task. From the recent studies, TPS [6] and Chord Complexity Distance [16] are suitable candidates. If the distance is evaluated simply for each pair of successive chords, the resulting series can represent the characteristic curvature of the song, as seen on Fig. 1.

Chord Complexity Distance. We proceed by employing Chord Complexity Distance (ChordCD) of the two chords, since it is a simple, yet discriminative alternative to the more complex TPS distance.

Based on the same principles as TPS, ChordCD also takes into account the key for the transition, and the exact construction of the two chords. Our simplification lies in weighting the presence of non-chord and non-key tones, achieving

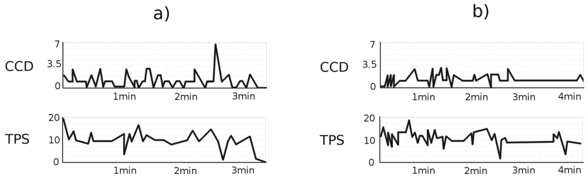

Fig. 1. Analysis of songs: (a) Hallelujah by Bastian Baker and (b) Wonderwall by Oasis, for both Tonal Pitch Space distance (TPS) and Chord Complexity Distance (CCD). Peaks on (a) represent a complex chord progression before the chorus in 0:35, 1:15, or 2:30. Peaks on (b) around 0:40, 1:30 and 2:30 correspond to the A5 chord followed by B7sus4 chord.

a complexity distance (as opposed to the use of Circle of Fifths in TPS). The system for deriving ChordCD can be considered as a grammar-like system. The chord consists of multiple tones, forming a sentential form in the music grammar. Chord distance is a number of steps of the derivation from one sentential form to another (see Fig. 2). We define *add* and *remove* derivations for adding/removing a tone from the same key. If the tone has to be moved outside the key, the derivation *alter* can perform alteration, with the inverse derivation *alter-back*. ChordCD is therefore a variation of an edit distance, while the constrains for adding/altering the tones are making it relevant in the tonal space.

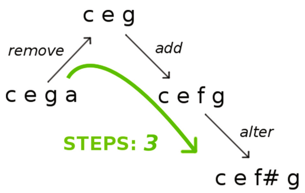

Fig. 2. ChordCD model for chord distances based on adding tones from the same key, and altering the tones outside the key. The key is evaluated for each chord tuple. Separate rules apply for changes between Tonic, Subdominant and Dominant [16].

The transitions between Tonic, Subdominant and Dominant triads are considered non-complex and yield 0 complexity of the transition. On the other hand, transitions in between chords with added dissonances yield higher complexity measures, as seen on Fig. 2. For more information and implementation details, we refer the reader to the complexity model explanation in [16].

3.3 Chroma Vector Distances

To the best of our knowledge, there have been no recent efforts in deriving distances in between the chroma vectors. We attribute this to the fact that the methods proposed [7,23] were successful for the tasks given with the 12-dimensional vectors. However, to obtain a fingerprint for the song, which would be more suitable for the DTW computation and large datasets, we proceed with the chroma vector distance definition as an intermediary feature for our comparison.

Chroma Complexity Distance. We propose a Chroma Complexity Distance (ChromaCD) definition based on a simple vector difference of the chroma vectors. We propose to improve this difference by weighting each bin according to the tone it represents. The details are in the following definition:

Definition 1. *For chroma vectors x and y, chroma complexity distance $\delta(x, y)$ is defined as:*

$$\delta(x, y) = \sum_{i=1}^{12} |w(x)_i x_i - w(y)_i y_i|$$

where $w(x)$ and $w(y)$ are 12-dimensional weight vectors similar to chroma vectors, having their 12 weight values dependent on the context for the transition from x to y, by the following rules:

$$\begin{aligned} w(x)_i &= 0 \Leftrightarrow c(x)_i = 1 \\ w(x)_i &= 1 \Leftrightarrow k(c(x), c(y))_i = 1 \wedge c(x)_i = 0 \\ w(x)_i &= 2 \Leftrightarrow k(c(x), c(y))_i = 0 \wedge c(x)_i = 0 \end{aligned}$$

where function $c(x)$ returns the chord estimation for the chroma x, and function $k(c, d)$ returns the key estimation for the two chords c and d.

Remark 1. Return values of the functions c and k (chords and keys) are 12-dimensional binary vectors, where value 1 is assigned to the tones present in the chord (key), and value 0 is assigned to the tones not present in the chord (key).

Functions for chord and key estimation can be chosen independently from the definition. We have employed a simple chord estimation, where $c(x)$, is the closest chord (by Euclidean distance) to the chroma vector x. The key estimation can be accomplished by a ranked list of chord usage for a given key (more used chords have higher ranks in the list). Function k then simply chooses a key for which the sum of ranks of the two chords is minimal. We employ the tonal harmony ranking described in [16] (chord complexity), but any valid ranking can be used, e.g. based on statistics or perception.

In a simple example, chroma vectors x and y may both be estimated to *A major* chord ($c(x) = c(y) = <1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0>$), with *A major* being the common key ($k(c(x), c(y)) = <1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1>$). The weight vectors represent the context of this transition: $w(x) = w(y) = <0, 2, 1, 2, 0, 1, 2, 0, 2, 1, 2, 1>$ (notice the chord tones having $w(x)_i = 0$, scale tones having $w(x)_i = 1$ and non-chord, non-scale tones having $w(x)_i = 2$). We treat chord tones in the chroma vectors as non-complex, therefore we do not weight the changes among the chord tones (e.g. simple music accompaniment); we assign weights to the more dissonant changes.

The proposed definition is very close to the ChordCD defined earlier for chords: adding a non-key tone is twice as complex as adding a non-chord tone from the key. Both distances are robust to the key transposition of the cover song, but it comes with a common disadvantage - dependency on chord and key estimation techniques.

4 Dynamic Time Warping

Dynamic Time Warping [18] is a method used for determining an optimal alignment between two time series. Based on this, an alignment score (distance) is calculated. Main advantages of DTW are the ability to compare sequences which differ in length and its generality. When introduced, DTW was used mainly for speech recognition, but thanks to the benefits mentioned, it quickly spread into other areas such as robotics, medicine, video games, music processing and many more. Importantly for our work, DTW is a straightforward technique to identify a cover song played in a different tempo.

4.1 Principles of DTW

Let us consider time series S and T. The basic idea behind DTW is to compare every data point from time series S with every data point in time series T. By this the distance matrix is built in which minimal warping path is found.

Definition 2. *Warping path in the distance matrix M is a sequence $W = (w_1, ..., w_l)$, where $w_l = M(i,j)$ and $l \in [1, |W|]$ that satisfies the following conditions.*

1. Condition of boundaries: $w_1 = (1,1)$ and $w_l = (|S|, |T|)$
2. Condition of a step size: $w_{l+1} - w_l \in \{(1,0),(0,1),(1,1)\}$ for $l \in [1, |W| - 1]$

In the end DTW distance is calculated using Eqs. (4.1) or (4.2)[17].

$$DTW(S,T) = \min \left\{ \sqrt{\sum_{l=1}^{L} w_l} \right. \tag{4.1}$$

$$DTW(S,T) = \min \left\{ \frac{1}{L} \sqrt{\sum_{l=1}^{L} w_l} \right. \tag{4.2}$$

Where L is the length of warping path W. Equation (4.2) for calculating the score is divided by the length of the warping path to account for the situations when compared sequences have different lengths. *Identity* occurs when $DTW = 0$.

4.2 Similarity in DTW

Standard Eqs. (4.1) and (4.2) have unfortunately one main problem when we use them for calculating DTW score. The score is not normalized in the interval $[0, 1]$, which means that it is difficult to compare DTW results across different datasets. For that reason we are testing three equations for calculating similarity by using the information from building the distance matrix and the warping path.

$$DTW_{sim}(S,T) = \frac{|S| + |T|}{|S| + l + |T| + u} \tag{4.3}$$

The first similarity score (4.3) [25] which we use is simply obtained by dividing the original length of the sequences by the length of the sequences containing the edit information. Compared to other methods which deal with the alignment of sequences, for example the Needleman-Wunsch algorithm (used for alignment of DNA sequences), DTW edits do not prolong the sequences in the resulting alignment. Edits in the DTW method occur in the situation when a continuous sequence of elements from one time series is collapsing into one point of another time series and vice versa (see Fig. 3). In Eq. (4.3) l and u are the values of horizontal and vertical movements in the warping path, where every l or u movement means an insertion of edit information into the sequence.

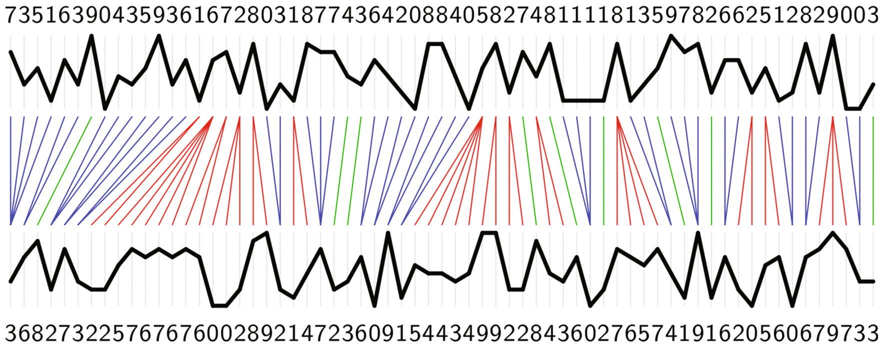
7351639043593616728031877436420884058274811118135978266251282900З
3682732257676760028921472360915443499228436027657419162056067973З

Fig. 3. Example of DTW alignment between two time series. Directions of collapsing points are depicted by blue and red colors. The green lines connect elements mapped in one to one relationship.

This score is in the interval $[0, 1]$, where $identity = 1$. One disadvantage of this score is that when $DTW_{sim} = 1$ it doesn't necessarily mean that the compared sequences are identical. It only means that the compared time series are aligned to each other by one point to one point relationship. In the classic DTW this score would never be 1 if the lengths of the compared sequences are different.

The second similarity score that we test in this paper is the following equation [14].

$$DTW_{sim}(S,T) = 1 - \frac{\sqrt{DTW(S,T)}}{\sqrt{DTW_{max}(S,T)}} \tag{4.4}$$

We calculate the similarity by dividing the score from $DTW(S,T)$ by maximal value DTW_{max} which can theoretically occur. Let us consider time series $S \in O^{|S|}$ and $T \in O^{|T|}$, where O is the domain of the analyzed sequences. Maximal DTW score can be obtained by calculating DTW for sequences S', T', where $S' \in \{s'_1, ..., s'_i\}, \forall n \in \{1, ..., i\}(s'_n = min(O))$ and $T' \in \{t'_1, ..., t'_j\}, \forall n \in \{1, ..., j\}(t'_n = max(O))$. The result is subtracted from 1 so that *identity* occurs when $DTW_{max} = 1$ and difference when $DTW_{max} = 0$.

Our third similarity score for the comparison (4.5) [14] is a slightly adjusted version of the Eq. (4.4).

$$DTW_{sim}(S,T) = \left(1 - \frac{\sqrt{DTW(S,T)}}{\sqrt{DTW_{max}(S,T)}}\right) \cdot \frac{min(|S|,|T|)}{max(|S|,|T|)} \tag{4.5}$$

The Eq. (4.4) is multiplied by a coefficient calculated from differences in lengths of the compared sequences. By this we penalize the comparison of time series with the different lengths.

Obviously, it would be extremely inefficient to calculate DTW two times, one for standard DTW and one for DTW_{max}. Fortunately, the calculation of DTW_{max} can be simplified as seen in the Eq. (4.6).

$$DTW_{max}(S,T) = \sqrt{(max(O) - min(O))^2 \cdot max(|S|,|T|)} \tag{4.6}$$

5 Experiments

We have performed a series of experiments on two commonly used music datasets: covers80 [9] and SecondHandSongs [5]. For feature extraction, we have used a combination of Vamp Plugins[4] (NNLS Chroma, Chordino and Key Detector). We have further processed the resulting features with our Harmony Analyser software[5] to obtain the chord and chroma vector distances. All experiments were run on the Anselm[6] supercomputer in IT4Innovations National Supercomputing Center, where we used one computational node (Two Intel Sandy Bridge E5-2665 processors, each having 8 cores, 2.4 GHz and 32 GB RAM).

The covers80 dataset consists of 160 songs organized as 80 musical works, each in two versions. The first versions of the songs are used as queries for the search in the whole set of the second versions of the songs.

[4] http://www.vamp-plugins.org.

[5] http://www.harmony-analyser.org.

[6] https://docs.it4i.cz/anselm/hardware-overview.

SecondHandSongs is a set of 18,196 tracks with 5,854 cover song clusters (average cluster size is 3.11). For our comparison, we have taken a chunk of 999 songs (295 clusters) from SecondHandSongs train set - the first 999 songs listed in the official dataset information file.

5.1 Comparison of Features and DTW Scores

Our first results are the comparison in between the feature-score combinations within the same dataset. A standard evaluation metric for MIREX is the use of Mean arithmetic of Average Precision (MAP). MAP score takes into account the ranking of each cover song, and assigns the weights according to the rank. The result is in the interval $[0, 1]$, higher values mean that the cover songs ranked high and close together. For datasets of thousands of songs and small cover song clusters, MAP values are likely to be ~ 0.1 [4].

We can see the comparison in the Tables 1 and 2. Raw chroma vectors have outperformed chord and chroma vector distances. We attribute this to the fact that 12-dimensional chroma vectors contain much more information than the one-dimensional ChromaCD. The information is reduced even more with the chord distances (ChordCD). The execution time to obtain similarity matrices was: $\simeq 56$ s for raw chroma vectors, $\simeq 51$ s for ChromaCD and 25 ms for ChordCD time series for covers80 dataset. We can see a marginally better performance for ChromaCD. However, the performance difference is reduced on the larger dataset: on SecondHandSongs dataset the execution time is $\simeq 550$ s for raw chroma vectors and $\simeq 100$ s for ChordCD time series.

Comparison of all results over all data formats shows, that the best results are provided by DTW scores (4.1), (4.3) and (4.4) in this order. Scores (4.2) and (4.5) provide very poor results and we deem them as not suited for CSI task.

5.2 Comparison to the State-of-the-art

Besides MAP results we have examined Mean Average Rank (MAR) of the covers and the number of correctly identified covers (for covers80 dataset), to be able to compare our results with the state-of-the-art methods. MAR values are between $[1, N]$ where N is the number of songs, and we aim to achieve the lowest value possible.

Table 1. MAP results for each DTW score and feature for covers80 dataset.

Score	Raw chroma vectors	ChromaCD	ChordCD
Score (4.1)	**0.482**	0.094	0.142
Score (4.2)	0.103	0.070	0.071
Score (4.3)	0.417	0.174	0.156
Score (4.4)	0.454	0.061	0.114
Score (4.5)	0.082	0.041	0.034

Table 2. MAP results for each DTW score and feature for SecondHandSongs dataset.

Score	Raw chroma vectors	ChromaCD	ChordCD
Score (4.1)	**0.107**	0.031	0.019
Score (4.2)	0.021	0.014	0.014
Score (4.3)	0.029	0.035	0.021
Score (4.4)	0.043	0.015	0.012
Score (4.5)	0.008	0.008	0.009

An overview of MAR results can be seen in the Tables 3 and 4. For the SecondHandSongs dataset, the MAR metrics were evaluated before us by Bertin-Mahieux and Ellis [3,4]. They achieved average rank of 2,939 for over 12,960 songs [4], and 308,369 for 1 million songs [3]. In our experiments, Raw chroma vectors have the average rank **321 out of 999** songs, closely followed by ChromaCD (average rank 341). That is a worse result than the state-of-the-art, but a promising result for ChromaCD. For covers80 dataset, the results were similar, raw chroma vectors outperforming the simple distances (best average rank **14 out of 80** songs). The number of correctly identified covers was **33**, which was in fact achieved by the score (4.1) with a slightly worse average rank. This is similar to the first LabROSA system results [7]. On the other hand, ChromaCD achieved only 6 identified covers, despite the promising average rank. This shows that ChromaCD keep the relevant tonal information, but are not self-contained (too much extraction, and the differences between the songs are getting lost).

We have also gathered the first experimental results with TPS chord distance [6] for CSI task. TPS has outperformed the more simple ChordCD, with the best result **0.198** MAP and **27.575** MAR for covers80 dataset. Note that these results are better in MAP than ChromaCD. These first experiments are promising for a further study of TPS.

Table 3. Comparison of mean average rank score on covers80 dataset.

Score	Raw chroma vectors	ChromaCD	ChordCD
Score (4.1)	15.688	30.450	32.825
Score (4.3)	15.025	21.538	26.688
Score (4.4)	**13.963**	36.587	35.688

Table 4. Comparison of mean average rank score on SecondHandSongs dataset.

Score	Raw chroma vectors	ChromaCD	ChordCD
Score (4.1)	**321.033**	382.686	402.829
Score (4.3)	362.050	341.092	407.792
Score (4.4)	374.675	414.561	426.181

6 Conclusion and Future Work

We have tested and compared three levels of feature extraction and five different scores for DTW method on covers80 and SecondHandSongs datasets. From experimental results we have found out the best DTW score, and obtained a competitive result for chroma vector distances and TPS distance. Overall, we conclude that simplifying the chroma information in the way that a distance is evaluated from consecutive chords or chroma vectors is (by itself) not a good approach for the CSI task, despite the interesting visualizations of the time series. But it can be helpful as one of the features used in a more complex system. Possibilities for the future study are: experimenting with TPS distance, comparison of all possible chord or chroma vector distances, but more importantly, finding other ways of employing these distance measures (e.g. using the distances as a similarity measure in DTW, rather than a feature). We think that for such difficult tasks as CSI, a good selection and combination of features is the way to obtain the best results.

Acknowledgments. This work has been partially funded by the Charles University, project GA UK No. 1580317, project SVV 260451, by grant of SGS No. SP2017/177 "Optimization of machine learning algorithms for the HPC platform", VŠB - Technical University of Ostrava, Czech Republic, by The Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070".

References

1. Bartsch, M.A., Wakefield, G.H.: To catch a chorus: using chroma-based representations for audio thumbnailing. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2001 (2001)
2. Bello, J.P.: Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats. In: Music Information Retrieval Evaluation eXchange, MIREX 2007 (2007)
3. Bertin-Mahieux, T., Ellis, D.P.W.: Large-scale cover song recognition using hashed chroma landmarks. In: IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, WASPAA 2011. IEEE (2011)
4. Bertin-Mahieux, T., Ellis, D.P.: Large-scale cover song recognition using the 2D fourier transform magnitude. In: Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012 (2012)
5. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011 (2011)
6. De Haas, W.B., Veltkamp, R., Wiering, F.: Tonal pitch step distance: a similarity measure for chord progressions. In: Proceedings of the 9th International Conference on Music Information Retrieval, ISMIR 2008 (2008)

7. Ellis, D.P.W., Poliner, G.E.: Identifying ‘Cover Songs’ with chroma features and dynamic programming beat tracking. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007 (2007)
8. Ellis, D.P.W.: Identifying ‘Cover Songs’ with beat-synchronous chroma features. In: Music Information Retrieval Evaluation eXchange, MIREX 2006 (2006)
9. Ellis, D.P.W., Cotton, C.V.: The 2007 LabROSA Cover Song Detection System. In: Music Information Retrieval Evaluation eXchange, MIREX 2007 (2007)
10. Foster, P., Dixon, S., Klapuri, A.: Identifying cover songs using information-theoretic measures of similarity. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(6), 993–1005 (2015)
11. Fujishima, T.: Realtime chord recognition of musical sound: a system using common lisp music. In: Proceedings of the International Computer Music Conference, ICMC 1999 (1999)
12. Gómez, E.: Tonal description of music audio signals, Ph.D. thesis, Universitat Pompeu Fabra (2006)
13. Khadkevich, M., Omologo, M.: Large-scale cover song identification using chord profiles. In: Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013 (2013)
14. Kocyan, T.: Adapting case-based reasoning for processing natural phenomena data, Ph.D. thesis, VŠB Technical University of Ostrava (2015)
15. Lee, K.: Identifying cover songs from audio using harmonic representation. In: Music Information Retrieval Evaluation eXchange, MIREX 2006 (2006)
16. Marsik, L., Pokorny, J., Ilcik, M.: Towards a harmonic complexity of musical pieces. In: Proceedings of the 14th Annual International Workshop on Databases, Texts, Specifications and Objects (DATESO 2014), CEUR Workshop Proceedings, vol. 1139 (2014). CEUR-WS.org
17. Mueen, A., Keogh, E.J.: Extracting optimal performance from dynamic time warping, KDD 2016 (2016)
18. Müller, M.: Information Retrieval for Music and Motion. Springer, Heidelberg (2007)
19. Osmalskyj, J., Piérard, S., Van Droogenbroeck, M., Embrechts, J.J.: Efficient database pruning for large-scale cover song recognition. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013 (2013)
20. Robine, M., Hanna, P., Ferraro, P., Allali, J.: Adaptation of string matching algorithms for identification of near-duplicate music documents. In: Proceedings of the International SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, SIGIR-PAN 2007 (2007)
21. Rocher, T., Robine, M., Hanna, P., Desainte-Catherine, M.: A survey of chord distances with comparison for chord analysis. In: Proceedings of the International Computer Music Conference, ICMC 2010 (2010)
22. Serrà, J., Gómez, E., Herrera, P., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Trans. Audio Speech Lang. Process. **16**, 1138–1152 (2008)
23. Serrà, J., Serra, X., Andrzejak, R.G.: Cross recurrence quantification for cover song identification. New J. Phys. **11**(9), 093017 (2009)
24. Tralie, C.J., Bendich, P.: Cover song identification with timbral shape sequences. In: Music Information Retrieval Evaluation eXchange, MIREX 2015 (2015)
25. Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., Keogh, E.: Indexing multidimensional time-series. VLDB J. **15**(1), 1–20 (2006)