COVER SONG RETRIEVAL: A COMPARATIVE STUDY OF SYSTEM COMPONENT CHOICES

Cynthia C.S. Liem, Alan Hanjalic
Department of Mediamatics, Delft University of Technology, The Netherlands
{c.c.s.liem,a.hanjalic}@tudelft.nl

ABSTRACT

The Cover Song Retrieval (CSR) problem has received considerable attention in the MIREX 2006-2008 evaluation sessions. While the reported performance figures provide a general idea about the strengths of the submitted systems, it is not clear what actually causes the reported performance of a certain system. In other words, the question arises whether some system component design choices are more critical for a system's performance results than others. In order to obtain a better understanding of the performance of current CSR approaches and to give recommendations for future research in the field of CSR, we designed and performed a comparative study involving system component design approaches from the best-performing systems in MIREX 2006 and 2007. The datasets used for evaluation were carefully chosen to cover the broad spectrum of the cover song domain, while still providing designated test cases. While the choice of the dissimilarity assessment method was found to cause the largest CSR performance boost and very good retrieval results were obtained on classical opus retrieval cases, results obtained on a new test case, involving recordings originating from different microphone sets, point out new challenges in optimizing the feature representation step.

1. INTRODUCTION

Cover Song Retrieval (CSR) generally refers to the problem of identifying different interpretations of the same musical work. Since 2006, this challenge has been included in the centralized yearly Music Information Retrieval (MIR) evaluation sessions known as the MIR EXchange (MIREX). Ever since, several systems for this task have been submitted and evaluated on a fixed, but undisclosed dataset. As the results obtained by these systems are expressed in the form of general performance numbers, no information is provided that could reveal the influence of specific CSR system component design choices and the composition of the evaluation dataset on the obtained retrieval results.

Although CSR appears to be more specific than e.g. music genre retrieval, cover songs still span a broad range of types, each with their own variants and invariants, posing specific challenges on the design of the CSR system. In order to validate design motivations and identify which system aspects are most critical for performance results, it is necessary to consider CSR systems as combinations of general system components and review performance with respect to these components. Additionally, the design of the evaluation dataset is critical for obtaining true insight into the performance of CSR systems.

In this paper, a comparative study is presented with special attention to the influence of individual system components and the composition of evaluation datasets on CSR system performance. We look at the two best-performing systems in MIREX 2006-2007, breaking them down into separate, generic components, which are recombined into alternative combinations. These are evaluated on 4 different datasets. Attention will hereby be paid to the validation of several 'semantically intuitive' choices in the systems. In this way, we aim at achieving better understanding of current CSR approaches, identifying which system components are most critical for the final performance results and which research directions deserve further attention in future CSR research.

2. PROBLEM DESCRIPTION

2.1 Definition of 'Cover Song'

While the term 'cover song' (or simply 'cover') used to suggest a pop music phenomenon, it has more recently been defined as 'a recording of a song or tune which has previously been recorded by someone else' [1]. This broad definition has typically been accepted in the MIR research field, accepting alternate takes of a song by the same artist to be covers as well. When considering the broad range of cover songs according to this definition, many musical aspects can be thought of that may vary among different covers. Several good suggestions for musical aspects that can be used in characterizing cover songs are given in [1].

2.2 Cover Song System Components

For the CSR problem discussed in this paper, a setting is assumed in which an example raw audio file is provided

[1] This also is seen in dictionaries, e.g. see http://dictionary.cambridge.org/define.asp?key=17817&dict=CALD, accessed May 2009.

as a query to a dataset, after which the audio files in the dataset are returned in an ordered way, according to their similarity score compared to the query. In this setting, CSR systems can be characterized using a general model, consisting of two main components:

- *Feature representation*, transforming a raw audio file into a representation suitable for further matching;
- *Dissimilarity assessment*, achieving the actual matching, applying a chosen dissimilarity measure.

Two more system aspects concerned with post-processing of the feature representation will further be considered:

- The typical approach of using short-time harmonic features for the feature representation produces very much data. In order to reduce this amount of data, an *averaging step* is adopted.
- In order to handle varying sound intensity levels, which can be caused both by the quality of the recording and by musical dynamics, a *normalization procedure* is usually applied to the chosen feature representation.

The musical variants expected in cover songs have influenced design choices for the mentioned system components. For the feature representation, chromagrams are commonly chosen[2]. These are considered to model melodic/harmonic progression over time without the need for exact transcriptions, while being robust to specific instrument timbres. Besides, multiple interpretations of a song will inevitably introduce tempo and timing variations, which also should be accounted for in CSR systems.

If system evaluations are done as a whole, it will not be clear from the results which of these component choices are most important to the final performance results. Additionally, validation of design choice motivations (such as the timbre-robustness of chromagrams) will be difficult.

2.3 Importance of Evaluation Dataset Composition

While during the system design, attention is paid to possible musical variants in cover songs, these do not appear to be considered with the same importance in system evaluation. Evaluation datasets typically are colorful cross-sections of private music collections, which are sought to contain as much musical variation as possible. However, the more variants in the dataset, the more difficult it will be to interpret an overall performance number. Given the broadness of the cover song spectrum, understanding of a system's performance can only be achieved if attention is paid to the types of cover song similarity test cases posed by a dataset.

A common problem in audio-based MIR research is the lack of public benchmark data. When different authors report performance numbers on different private music collections, comparison of their approaches cannot be made easily. The MIREX endeavour offered a centralized solution to this, comparing multiple algorithms on the same evaluation data. However, as details regarding the evaluation dataset composition are not revealed to the participants, only comparative information on total system performance is provided, while algorithm behavior on specific test cases once again remains unclear.

2.4 Contribution

To address the problems described above and gain more in-depth understanding of CSR performance in current approaches, in this paper, we describe a comparative study with two main focus points:

- to investigate the impact of choices in each individual general CSR system component listed above on the CSR performance;
- to relate the achieved performance results to specific test cases provided by the evaluation data.

The setup of this comparative study is explained in Section 3, while the results are reported and discussed in Section 4. We finish the paper in Section 5 with conclusions and recommendations for future work.

3. EVALUATION SETUP

In this section we first explain the systems we selected and implemented for our comparative study.

3.1 Basic Systems

3.1.1 Best CSR system in MIREX 2006

The system proposed by Ellis et al. in [2] was the best-performing system in the first MIREX CSR Task, held in 2006. We use the implementation that has been made available by the author [3], which is very similar to this original 2006 MIREX CSR submission.

Regarding the feature representation, chromagrams based on instantaneous frequency (CIF) are used. Features are averaged over beats, which appears to be a semantically intuitive choice, allowing robustness to tempo variances; a beat tracker is needed in order to achieve this. For normalization, each 12-bin chroma vector in the chromagram is normalized to unit norm.

For similarity assessment, cross-correlation (CC) is performed. In order to allow for different key transpositions, all 12 possible chroma transpositions are considered in this correlation step. Subsequently, a similarity score is achieved through the maximum peak correlation value found. This can be changed into a dissimilarity score by taking the reciprocal of this value.

3.1.2 Best CSR system in MIREX 2007

The system proposed by Serrà et al. in [4] was the best-performing system in the second MIREX CSR Task, held in 2007. This system showed a striking performance increase compared to all other systems; an improved version

[2] see for example the extended abstracts on `http://www.music-ir.org/mirex/[yearofsession]/index.php/Audio_Cover_Song_Identification_Results`, with 2006, 2007 and 2008 as possible session years (accessed May 2009).

also convincingly showed the best performance results in the 2008 MIREX CSR Task [5].

As no implementations of this system are publicly available, for the experiments described, the system has been reimplemented from the literature, using the information in [1, 4, 6]. The preprocessing steps (transient localization and spectral normalization) have still been omitted in our implementation, as it was not completely clear which procedures were exactly followed for these steps. For the same reasons, the system changes and parameter tunings mentioned in [5] could not be implemented, so our implemented system will show the most resemblance to the MIREX 2007 submission by the authors.

For feature representation, Harmonic Pitch Class Profiles (HPCPs) [6] are used. These are chromagrams (or pitch class profiles) in which each spectral peak contribution is weighted across multiple chroma bins. Additional contribution is weighted into the final representation by taking into account the first 8 harmonics of each spectral peak. Averaging is done over a fixed number of frames, as beat tracking was found to include additional errors that decreased performance (this also was noted in [7]). Normalization is performed by dividing a HPCP instance by the maximal value found in this instance, yielding a profile in which the maximum value is 1.

For matching, a procedure was devised called Dynamic Programming Local Alignment (DPLA), using binary similarity. For the two audio HPCP vectors to be matched, first an Optimal Transposition Index is computed. Subsequently, after applying the found optimal key transposition, a binary similarity matrix is constructed, based on remaining optimal transposition indices per HPCP short-time instance after the global transposition. Subsequently, in a way similar to string or DNA matching, a dynamic programming procedure with local constraints (for tempo fluctuations) is applied. The best path found will decide the similarity score, which is normalized to a dissimilarity score. More information on these procedures can be found in [1]. Parameter choices have been directly taken from [1]; as for the averaging factor, the choice was made to consider an averaging factor of 10 frames. Furthermore, only 12-bin HPCPs are considered instead of the suggested 36 bins, as 12 bins were used both in the Ellis et al. system and in later versions of the Serrà et al. system.

3.2 Considered Approaches

Using the systems described above, several possible general design choices can be extracted. The following choices have been verified in our algorithms:

- The general choice of feature representation: (1) *Chromagrams based on Instantaneous Frequency (CIF)*, (2) *Harmonic Pitch Class Profiles (HPCP)* and (3) *Pitch Class Profiles (PCP)*, which are constructed similarly to HPCPs, but omitting the additional harmonic weighting.
- The averaging factor for the feature representation: (1) *averaging over beats* and (2) *averaging over a fixed number of frames*.
- The matching procedure for dissimilarity assessment: (1) *cross-correlation (CC)* and (2) *Dynamic Programming Local Alignment (DPLA)*.

All possible combinations of these choices have been tested, with three possible normalization choices regarding feature representation: (1) *no normalization*, (2) *normalization to unit norm* and (3) *normalization by the maximum*.

3.3 Performance measures

We evaluate the systems using 2 evaluation measures, which also were adopted in the most recent MIREX evaluations [8]:

- (Arithmetic) Mean of average precisions (MAP);
- Mean rank of 1st correctly identified cover (MR1st).

The most recent MIREX evaluations employ two more evaluation measures focusing on the top-10 retrieval results. However, in our experiments, only MAP and MR1st will be suitable performance indicators: our datasets, which are discussed hereafter, contain cover sets of different sizes, as opposed to the MIREX dataset which contained 10 relevant cover versions per query song.

3.4 Datasets

4 datasets have been used in our experiments, which will be described now. The construction and choice for the datasets has largely been motivated by the need to provide clear and designated test cases. The choice was made to use 4 separate datasets in order to provide a clear-cut corpus per dataset. All audio tracks have been converted to the MP3 format. For each dataset, each audio file in the dataset was matched against all other files in the same dataset.

3.4.1 Covers80 dataset

This dataset, containing pop song covers, was made available by Ellis [3]. 166 recordings are included, encompassing 80 ‘cover sets’, which means the average number of versions is just 2.05. With the dataset being constructed rather randomly, musical variants within the dataset differ greatly and interpreting performance measures will be difficult. We decided to include results on this set anyway for reference reasons.

3.4.2 Beethoven piano sonatas

This dataset contains multiple interpretations of movements from 4 Beethoven piano sonatas. The data in this dataset originates from private music collections of the authors and the Beeld en Geluid (BeG) vinyl collection[3] in the European archive. The dataset contains 128 recordings, encompassing 13 ‘cover sets’. As piano sonatas are

[3] http://europarchive.org/collection.php?id=public_classical_music_BeG, accessed May 2009.

considered, all covers will consist of very similar instrument timbres and will be played in exactly the same key. Therefore, the set has very clear invariants and poses well-defined (although not too challenging) similarity tasks. A set of similar composition was used for a CSR system mentioned in [9], which showed near-perfect performance.

3.4.3 Songs

This dataset departs from recordings of classical art songs that were performed at the 1st International Student Lied Duo Competition, held in Enschede in April 2009. It contains study recordings from one of the participating duos, made at rehearsals and try-outs in preparation for the competition. Additionally, recordings of all the participants made during the official competition rounds are included. More specifically, included songs encompass compulsory songs, as well as songs that were performed by multiple different participants. Finally, the set was extended with extra song interpretations from private music collections, the BeG vinyl collection and the vinyl recordings from the King's Sound Archive [4] . In total, the dataset contains 205 recordings, encompassing 21 'cover sets'.

At the competition, recordings have been made with two different pairs of microphones at two different locations in the hall (on stage and in the hall). While recordings from these two pairs contain exactly the same musical interpretation, the recordings do show considerable acoustical differences. As this poses an interesting test case for the CSR algorithms, the takes from both microphone pairs have been included in the dataset.

Both this songs dataset and the Beethoven dataset consider multiple interpretations of exactly the same score. The problem of retrieving such interpretations has sometimes been considered as a subtask within CSR, known as opus retrieval. The difference between both sets is that the songs set shows much more variation in instrument timbre and musical keys, as the performing singers have different voice types.

3.4.4 Beatles

This dataset aims at being a slightly more specific dataset than the covers80 set with larger 'cover set' sizes, while still reflecting a similar corpus. The dataset contains original Beatles songs (including alternative takes and versions), as well as various covers taken from tribute CDs, including Baroque, R&B, Latin and easy listening styles. In total there are 197 audio files, encompassing 51 'cover sets'. On one of the CDs used, 4 covers were present of songs from individual Beatles members. These were included in our database without providing alternative versions. Typical CSR evaluation experiments contain even more of such 'outlier noise' files in evaluation datasets (e.g. MIREX), but in our experiments they explicitly have not been included extensively in order to focus on system behavior on actual covers.

4. RESULTS

Each of the possible combinations mentioned in Subsection 3.2 has been tested on all 4 datasets. The resulting MAP and MR1st scores are plotted twice, both in Figure 1 and Figure 2. While the performance scores are the same in both figures (expressed in data points at the same locations on the vertical axis), the used data markers indicate different system choices. In Figure 1, the data point markers indicate corresponding combinations of feature representations and dissimilarity assessment, while the data point markers in Figure 2 indicate corresponding combinations of normalization and averaging choices. In the figures, baseline results for random guessing are indicated as well, which were obtained by generating 50 random similarity matrices for each dataset and averaging the obtained results. Because of space limitations, only the results of the best-performing system combinations are numerically expressed in Table 1.

Dataset	MAP	MR1st
covers80	0.648	15.817
Beethoven	1	1
songs	0.986	1
Beatles	0.693	5.699

Table 1. Best performance scores for each of the datasets

The best results turn out to occur for the same combination consistently: the CIF feature representation, averaged over a fixed number of 10 frames, normalized to unit norm and with dissimilarity assessment based on DPLA. This means aspects from both studied original systems combine into an optimally performing system.

With respect to the feature representation choice, the CIF representation generally does not perform worse than HPCPs or PCPs. In the pop music datasets (covers80 and Beatles), it even performs clearly better than HPCPs and PCPs. Besides, as mentioned above, the CIF representation consistently occurs in the best-performing system component combinations for each of the 4 datasets. Regarding the difference between HPCPs and PCPs, the harmonic weighting in the HPCPs does not give convincing performance increases when compared to PCPs. While HPCPs were known to yield the highest correlation scores when compared to symbolic note information [6], this does not appear to be a convincing advantage in the CSR problem, which deals with approximate matches.

The notion in [1,4] that DPLA dissimilarity assessment yields much better results than CC is convincingly confirmed for all 4 datasets. This also holds for the statement in [1, 4, 7] that averaging over a fixed number of frames improves performance in comparison to averaging over tracked beats. While all better-performing system combinations contain normalized feature representations, the performance increase from the normalization step is much smaller than the increases caused by choosing DPLA dissimilarity and averaging over a fixed number of frames. Furthermore, there is no specific normalization

[4] `http://www.kcl.ac.uk/kis/schools/hums/music/ksa/ksa_sound.html`, accessed May 2009.

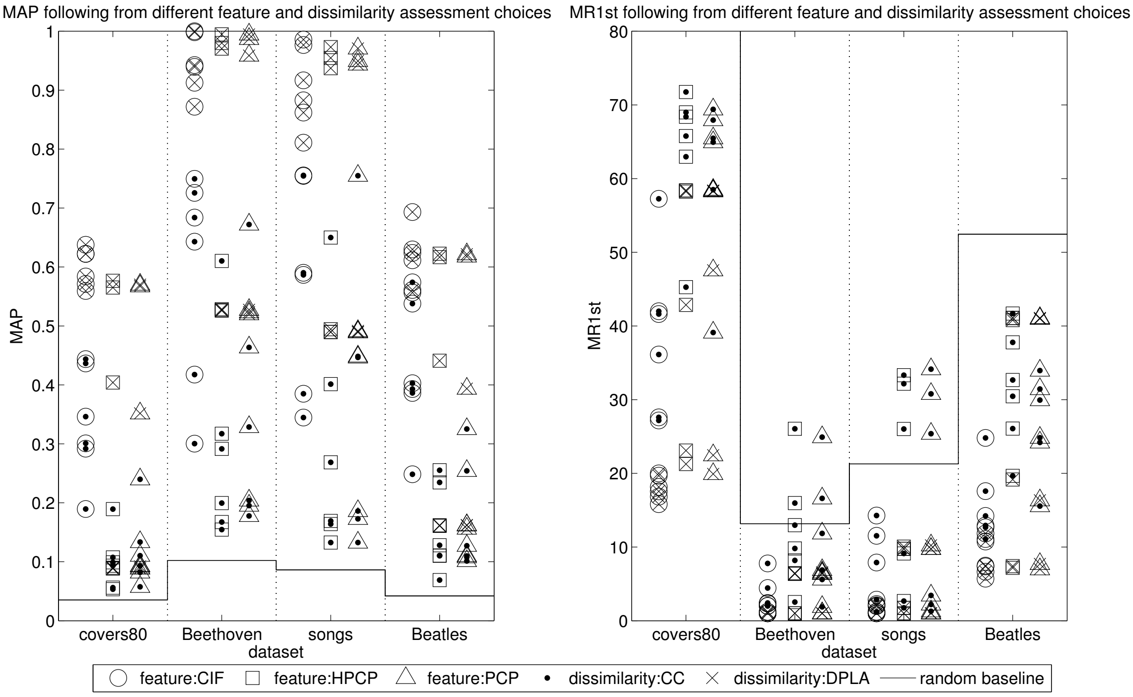

Figure 1. MAP and MR1st for the 4 datasets with feature and dissimilarity assessment choices indicated.

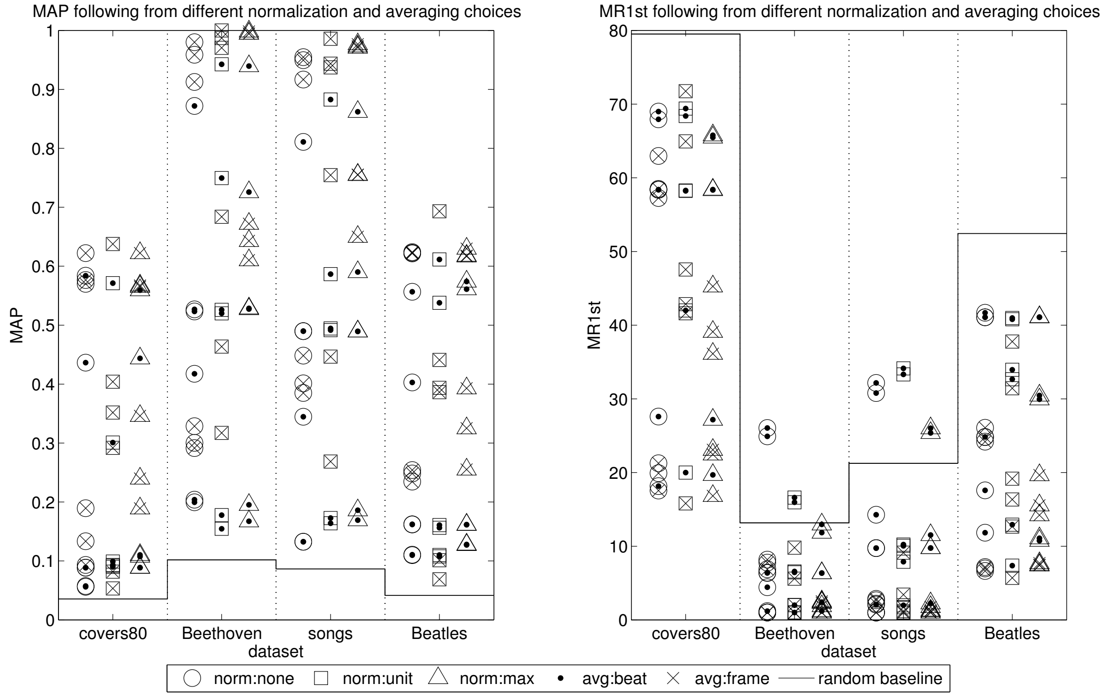

Figure 2. MAP and MR1st for the 4 datasets with normalization and averaging choices indicated.

choice performing convincingly better than other normalization choices.

As expected, the results for the classical opus retrieval cases (Beethoven and songs) are better than those on the pop music datasets. However, it is remarkable how close the performance on both classical datasets is, despite the much larger variations in timbre and key in the songs dataset. Errors in near-perfect results on the Beethoven set are caused by the historic vinyl recordings, which are degraded in quality compared to modern recordings. However, as shown in Table 1, the best-performing system combination was robust to the vinyl recording sound distortions, having perfect retrieval results on this set. In the Beatles database, if besides a query multiple alternative recordings of the same artists exist, these recordings are ranked very high in the retrieval results. However, the performance results worsen because of the inability of all implemented approaches to deal with the freer covers, such as the easy listening piano versions of the Beatles songs.

The similarity test cases posed by the songs dataset demonstrate some other interesting properties of the current CSR approaches. If a song is available in multiple interpretations from the same musicians, these interpretations are usually ranked higher than interpretations of other musicians. This might be due to timing aspects rather than timbral aspects, as interpretations of other singers of the same voice type as the singer in the query do not consistently rank higher than interpretations of other singers of other voice types or even the other gender. This validates the hypothesis that the followed approaches show timbre-robustness. The claimed key invariance of all approaches is also confirmed in our results, as songs sung in the same key as a query do not always rank higher than recordings of the song in other keys.

In the best system combinations for the songs database, if an alternate microphone recording of a given song is available, it is retrieved as the best-matching song. However, while such recording pairs undoubtedly contain exactly the same musical interpretation, the found dissimilarity scores of both pair members compared to a query of another interpretation are not identical. It even is not guaranteed that both pair members will be neighbors in the corresponding dissimilarity ranking to the query. This is an interesting notion that does not match our human notion of interpretation similarity.

5. CONCLUSION AND DISCUSSION

In this paper, more insight into the performance of current CSR approaches was sought through a comparative study, in which different combinations of CSR system components were evaluated on 4 carefully constructed datasets. The obtained results show that choices that semantically seemed intuitive do not necessarily yield better performance results: including harmonic weighting into a feature representation does not convincingly show performance improvement, while averaging the representations over beats actually makes the results worse.

Regarding the system components, the best feature representation found consistently in our experiments is the CIF representation, which is not the representation used in the best MIREX systems of 2007 and 2008. However, the dissimilarity assessment method used in those systems, binary similarity using DPLA, gives a large performance increase in comparison to using CC. This suggests that the dissimilarity measure has been the crucial factor in the success of the best MIREX CSR system submissions of 2007 and 2008. The remaining system aspect that was tested, the feature normalization, only gives a slight increase in performance.

Successful CSR system component combinations can deal very well with opus retrieval tasks, even if large timbre and key variance is present. However, ranking results for the different microphone recording pairs in the songs dataset show different ranks for identical musical interpretations which only differ in terms of the acoustical conditions. Therefore, the difference in the dissimilarity must be due to the feature representation, suggesting that further improvement is still possible here.

While the major changes from the best-performing system of MIREX 2007 to that of 2008 mainly focused on improving the dissimilarity assessment part [5], improvement possibilities in the other system components, especially the feature representation, are clearly not excluded. Further experiments are needed into alternatives that will be able to yield results that better approach our human notions of cover song similarity.

6. REFERENCES

[1] J. Serrà. Music similarity based on sequences of descriptors: Tonal features applied to audio cover song identification. Master's thesis, University Pompeu Fabra, Barcelona, Spain, September 2007.

[2] D.P.W. Ellis and G.E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume IV, pages 1429–1432, Honolulu, USA, April 2007.

[3] D. Ellis. The `covers80` cover song data set. Web resource, available: `http://labrosa.ee.columbia.edu/projects/coversongs/covers80`, 2007.

[4] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio, Speech and Language Proc.*, 16:1138–1151, August 2008.

[5] J. Serrà, E. Gómez, and P. Herrera. Improving binary similarity and local alignment for cover song detection. MIREX 2008 extended abstract, available: `http://www.music-ir.org/mirex/2008/abs/CS_Serra.pdf`, September 2008.

[6] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, University Pompeu Fabra, Barcelona, Spain, July 2006.

[7] J.P. Bello. Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats. In *Proc. of the Intl. Conf. on MIR (ISMIR)*, Vienna, Austria, September 2007.

[8] J.S. Downie, M. Bay, A.F. Ehmann, and M.C. Jones. Audio cover song identification: MIREX 2006-2007 results and analyses. In *Proc. of the Intl. Conf. on MIR (ISMIR)*, Philadelphia, USA, September 2008.

[9] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proc. of the IEEE*, 96(4):668–696, April 2008.