Audio Cover Song Identification

Carlos Manuel Rodrigues Duarte
carlos.duarte@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa

October 2015

Abstract

Audio cover song identification is one of the main tasks in Music Information Retrieval and has many practical applications such as copyright infringement detection or studies regarding musical influence patterns. Audio cover song identification systems rely on the concept of musical similarity. To compute that similarity, it is necessary to understand the underlying musical facets such as timbre, rhythm and instrumentation, that characterize a song but, since that kind of information is not easy to identify, interpret and use, it is not a straightforward process. This document begins by giving information about the possible musical facets and how they influence the process of identifying a cover. The most common approaches to take advantage of those musical facets are addressed as well as how the similarity values between a pair of songs can be computed. There is also an explanation of how the system quality can be assessed. A system was chosen to serve as baseline and, based on recent work in the field, some experiments were made in order to try achieving an improvement in the results. In the best experiment an increase of 5% Mean Average Precision and 109 more covers being identified was obtained, using the similarity values of melody and voice descriptors fused together with the results given by the baseline.

Keywords: Chroma Features, Distance Fusion, Audio Cover Song Identification, Music Similarity, Music Information Retrieval

1. Introduction

Technology is rapidly evolving and nowadays it is possible to access digital libraries of music anywhere and anytime, and have personal libraries that can easily exceed the practical limits to listen to them [4]. These fast-pacing advances in technology also present new opportunities in the research area where patterns, tendencies and levels of influence can be measured in songs and artists. By having a way to compute a similarity measure between two songs, it is possible to provide new services such as automatic song recommendation and detection of copyright infringements.

These services can be achieved by identifying cover songs since, by their nature, cover songs rely on the concept of music similarity. In musical terms, a cover is a re-recording of an existing song that may, or may not, be performed by the original artist or have exactly the same features, but it has something that makes it recognizable once knowing the original. There exist many types of covers ranging from renowned bands that produce an existing music but in a style that corresponds to their identity, to unknown people who play music with the simple goal of trying to perform a song that they like.

The identification of cover songs is best made by humans. However, the amount of existing musical content makes manual identification of different versions of a song infeasible and, thus, an automatic solution must be used to achieve that even though it entails the issue of not knowing the exact way to represent human being's cognitive process. With that in mind, it is important to know which are the musical facets that characterize a song and what are the existing cover types in order to understand how they can be explored to make cover identification possible and the difficulty of computing an accurate similarity value between two songs.

Knowing the musical facets and how they can be used to extract meaningful information, a cover detection system can be constructed. The goal of this work will be to use an existing system, analyze its results, and develop a way to improve them. In this case, the improvements will be guided towards the identification of covers that are the closest possible, in terms of lyrics or instrumentation, to the original version.

This document will begin by giving background information about musical facets and how they affect the process of identifying musical covers. It

will review the most common approaches that audio cover song identification systems take in order to produce quality results and recent work in the area is addressed. A system was chosen to serve as baseline and based on the ideas of some recent work in the field, some experiments were conducted to improve the quality of its results. The improvement of the results was achieved using a heuristic for distance fusion between extracted melodies and the baseline, making possible the detection of covers that presented similar melodies and similar singing.

The next section reviews the underlying musical facets that condition the process of identifying a cover and addresses the common approaches taken for audio cover song identification. Section 3 is about recent related work. Section 4 describes all the experimental setups followed by the discussion of the results obtained. Conclusions are made and future work is discussed in section 5.

2. Musical Facets and Approaches for Cover Detection

The concept of cover is, in general, usually applied in the simplified sense of an artist reproducing the work of another artist, but it is not that straightforward. It is important to know what types of similarities exist between two songs and what they consist of. The type of cover can give us information, such as the changes that were applied or if the song was performed by the same artist or not. There are many types of covers such as remasterizations, instrumental versions, live performances, acapella, acoustic, and remix [18].

The type of cover can be useful to reveal what sort of resemblance we can expect between two songs. By knowing the most common possible types, one can expect a remasterization to be much more similar to the original song than a remixed version. That is due to the large quantity of possible variations that complicate the process of associating two songs together. Even a live performance can display enough variations to make the two digital audio signals different. Those variations can be relative to timbre, key, structure, tempo, lyrics, and even noise.

So far, the existing solutions for automatic audio cover identification rely on several approaches that try to make the most (or ignore) the information obtained from this musical facets and every year new techniques and approaches are created in the area of Music Information Retrieval (MIR). There is even a audio cover song identification competition held every year by an annual meeting named Music Information Retrieval Evaluation eXchange (MIREX).

2.1. Approaches for Cover Detection

The goal, for cover detection, is to compute a similarity value between different renditions of a song by identifying the musical facets that they share or, at least, the ones that present fewer variations. Those facets (e.g. timbre, key, tempo, structure) are subject to variations that make the process of computing that value more complex and forces the cover identification systems to be robust in that sense. The most common invariations that they try to achieve are related to tempo, key, and structure, since they are, in general, the most frequent changes and, together with feature extraction, they constitute the four basic blocks of functionality that may be present in a audio cover identification system.

Feature extraction
In this approach, there is an assumption that the main melody or the harmonic progression between two versions is preserved, independently of the main key used. That representation, or tonal sequence, is used as comparator in almost all cover identification algorithms. The representation can be the extraction of the main melody or at harmonic level with the extraction of the chroma features (also known as Pitch Class Profiles (PCP)).

Key Invariance
One of the most frequent changes between versions of one song is in its key. In order to achieve key invariance, one can perform all possible transpositions to find the most suitable. This is the optimal solution but also the slowest. To speed up the process, key estimation or Optimal Transposition Index (OTI) [19] can be used, but the optimal results will not be guaranteed.

Tempo Invariance
There are several ways to achieve tempo invariance such as beat tracking, compression and expansion techniques, and dynamic programming algorithms like Dynamic Time Warping (DTW).

Structure invariance
Summarizing a song into its most repeated or representative parts, or structural segmentation (i.e., split the song in sections), can be used. Dynamic programming alternatives, such as the Smith-Waterman algorithm, are also an alternative.

These approaches will create a representation for each song that will make possible for a comparison to be made between them. To compute the distance value between two songs, one can use the Euclidean Distance, the Cosine Distance, or even dynamic programming algorithms such as the DTW or Smith-Waterman, as previously mentioned. After computing the distance values, the quality of the system must be assessed. In order to do so, the evaluation metrics suggested by the MIREX task can be used. Those metrics are the total number

of correctly identified covers, the Average Precision (AP) at top 10, the Mean (arithmetic) of Average Precision (MAP), and the Mean Reciprocal Rank (MRR).

3. Related Work

Over the last years, in the area of cover song identification, there has been a considerable amount of new approaches and techniques that try to handle different issues. The typical goal is to try new algorithms or combinations of them in order to improve the results in comparison to previous systems, but the recent main focus by most researchers has been towards scalable strategies. The most common way to calculate the similarity between two different songs is through the use of alignment-based methods and they have shown to be able to produce good results[1] (75% MAP in MIREX'2009). However, these methods are computational expensive and, when applied to large databases, they can become impractical: the best performing algorithm [20] in MIREX'2008 implemented a modified version of the Smith-Waterman algorithm and took approximately 104 hours to compute the results for 1,000 songs[2]. If applied to the Million Song Dataset (MSD) dataset, the estimated time to conclude would be of 6 years [2].

Martin et al. [13] suggest the use of Basic Local Alignment Search Tool (BLAST), a bioinformatics sequence seaching algorithm, as an alternative to dynamic programming solutions. The data is indexed based in similarity between songs, and to compute the similarity value, the best subsequences are chosen, and then compared.

Khadkevich et al. [10] extract information about chords and store them using Locality-Sensitive Hashing (LSH). Bertin-Mahieux et al. [3] adopt the 2D Fourier Transform Magnitude for large-scale cover detection. This solution was further improved by Humphrey et al. [9], who modified the original work to use a sparse, high-dimensional data-driven component, and a supervised reduction of dimensions.

Balen et al. [2] extract high-level musical features that describe harmony, melody, and rhythm of a musical piece. The extracted descriptors are stored with LSH, which allows to retrive the most similar songs.

Lu and Cabrera [12] use hierarchical K-means clustering on chroma features to find audio words (centroids). A song will then be represented by its audio words. The similarity with other songs will be determined by how many audio words they share with the same location.

Outside the field of large-scale cover identification, several solutions regarding distance fusion have been suggested. Salamon et al. [17] extract the melodic line, the bassline, and HPCP 12-bins for each song. They explore the fusion of those features in order to discover which results in the best performance. The best results were obtained by fusing all the scores given by the three approaches. Distance fusion is also the main focus in the work of Degani et al. [6], where they propose a heuristic for distance fusion. Their proposal consists of normalizing all values to [0,1], computing a refined distance value, and produce a single matrix of results.

4. Experimental Setup

Taking into account the previous work and what has been done so far, an idea for an audio cover song identification system was created. The goal was to use a system as baseline and, given the results, determine what could be done in order to improve them. To evaluate the system, a dataset is required, and it would be best if it replicates the conditions of the MIREX task. To determine what improvements could be made, the results from the system would have to be analyzed manually and, given the patterns discovered, determine what is lacking in the baseline and figure out what are the best methods available to implement it.

4.1. Baseline

The system chosen to serve as baseline was developed by [7] that finished in first place in the MIREX'2006 Cover Song Identification Task[3]. This project is freely available online[4] and it is implemented in MATLAB.

4.2. Dataset

The dataset chosen was the CODA dataset [5]. CODA is a dataset created to replicate the ones used in the MIREX competition. Therefore, the dataset is composed by 30 original tracks, each with 10 covers, and another 670 with no relation with the originals. The songs are provided as MP3 files.

4.3. Experiments

Having the system and the dataset, results could be had. By checking the results in the Top10 of each query, it was possible to determine which were the covers that failed to be identified and the non-covers that were incorrectly identified. Some of those tracks were manually analysed and it was possible to observe some patterns. Most of the non-covers identified reflected the instrumentation used by the query. If, for example, the query song had mainly a classic guitar playing, the non-covers identified would most likely have a similar instrumenta-

[1] http://www.music-ir.org/mirex/wiki/2009:Audio_Cover_Song_Identification_Results

[2] http://www.music-ir.org/mirex/wiki/2008:Audio_Cover_Song_Identification_Results##Run_Times

[3] http://www.music-ir.org/mirex/wiki/2006:Audio_Cover_Song_Identification_Results

[4] http://labrosa.ee.columbia.edu/projects/coversongs/

tion with classic guitar. The covers that failed to be identified, on the other hand, showed that a large amount of them presented the exact same lyrics, rhythm, and the same melody, although most of them had different genre or instruments.

Considering these faults, some experiments were conducted that explored the use of the features present in the covers that failed to be detected. To merge the results from the experiments with the ones already had from the baseline, the metrics fusion method suggested by [6] was applied. The experiments conducted were related to rhythm, melody, and summarization, although the latter was made initially without checking the results from the baseline. The goal was to determine if summarizing a song would preserve its most important information so that the cover detection results would not be affected while being produced in a faster way.

4.3.1 Summaries

As previously mentioned, creating summaries of songs was explored to check if it was a good alternative to the original audio files. If the information held in the summaries was enough to identify the original song, they could be used as the dataset and provide results in a much faster way. The summaries were created using the work of [15] that applied unsupervised algorithms for speech and text to music audio files. The algorithms used for this matter were GRASSHOPPER, LexRank, Latent Semantic Analysis (LSA) and Maximum Marginal Relevance (MMR).

The configuration parameters used were: 0.5 for Frame Size and Hop Size for all; the Vocabulary Size was 50 for MMR and 25 for the rest; Sentence Size was 5 for LexRank and 10 for the rest. The Weighting parameter was BINARY for GRASSHOPPER and LSA while LexRank and MMR had DAMPENED_TF. The λ value used in MMR was 0.7. Each of these algorithms were produced in two manners: either as chroma features or as 20-bin vector MFCC's.

4.3.2 Rhythm

One of the observations made by manually analyzing the covers that failed to be in top 10 was that those tracks had similar rhythmic features with the original. Since that similarity was not detected, some sort of rhythmic features could be beneficial and improve the detection of those covers. This problem has already been addressed by some, such as [14] . In their work, an approach for rhythm extraction was presented that consisted in applying the Fourier Transform to the audio signal and then split it into low (0-100Hz) and high (12000-14000Hz) frequency bands. Twenty-six rhythmic features are then extracted.

The approach used was based on the rhythmic features extracted in [1] which is based of [14]. The main difference lies in the selection of 18 rhythmic features instead of 26. These features were extracted with MIRToolbox [11], a MATLAB framework with a large collection of musical feature extractors. The process was as follows:

- Compute the Fourier Transform for every 50ms frame, half-overlapping, for the low and high frequency bands. The frequency bands were the ones used in the original [14].
- Extract the rhythmic features suggested by [1].
- Merge the vectors of the two bands into a single one.
- Construct a matrix with all similarity values between all rhythmic features. The similarity value was obtained by computing the Euclidean distance.

4.3.3 Melody

To improve the detection of covers that present the same lyrics or similar melodic line, the work of [16] was considered. The task of extracting the melody consists of 1) calculate when the melody is present and when it is not, which can be referred to as voice detection and 2) estimating the correct pitch value of the melody when it is present.

The approach used is comprised of four stages: sinusoid extraction, salience function, contour creation, and melody selection. The goal of the first step is to analyze the audio signal and find out which frequencies are present at every point in time and enhance frequencies that are more perceptually sensitive to human listeners using an equal loudness filter. Next, the signal is split into small blocks that represent a specific moment in time and the Discrete Fourier Transform (DFT) is applied to each block, which gives us the spectral peaks, that are the most energetic frequencies at that moment and all the other frequencies are discarded.

The salience function stage, estimates how salient those selected frequencies are. In order to so, harmonic summation is used, which is searching for harmonic series of frequencies that would contribute to the perception of a correspondent pitch. The weighted sum of the energy of the harmonic frequencies is the salience of the pitch. The result is a representation of pitch salience over time, which is referred to as salience function.

The resulting salience function is used to track pitch contours. A pitch contour is a series of consecutive pitch values continuous in both time and frequency. The contours are tracked by taking the

Table 1: Results of all experiments

Features	Total @Top10	Average @Top10	MAP	MRR
Baseline (MIREX'2006)	761	2.310	-	0.490
Baseline (CODA)	1049	3.179	0.304	0.647
Baseline + GRASSHOPPER-Chroma	610	1.848	0.140	0.291
Baseline + Rhythm	1015	3.076	0.294	0.660
Baseline + Positive-only melody (PO-Melody)	1119	3.391	0.332	0.701
Baseline + All-values melody (AV-Melody)	1121	3.397	0.333	0.701
Baseline + PO-Melody + AV-Melody	**1158**	**3.509**	**0.348**	**0.732**

peaks of the salience function and use a set of cues based on auditory streaming to group them into contours. These contours can have a duration of a single note or even a short phrase.

The final stage is selecting the melody. This step does not consist of extracting all melody contours but rather filter out all non-melody contours and this is done by having a set of filtering rules. The result is a sequence of fundamental frequency (F0) values that correspond to the perceived pitch of the main melody.

The aforementioned approach is available through MELODIA[5], a Vamp[6] plugin, which can be used in a Vamp host program, such as the case of Sonic Annotator[7]. All the melodies were extracted for the CODA dataset in the form of a Comma-Separated Values (CSV) file, with two fields: a timestamp for every 2.9ms and the corresponding fundamental frequency value in Hertz. The frequency values can be relative to voiced or non-voiced segments whereas the latter is represented with zero or negative values.

The first processing step was quantising all the frequencies into a semitone representation (C, V, D, W, E, F, X, G, Y, A, Z) where V stands for C# , W for D#, X for F#, Y for G# and Z for A#. The length of the sequence was then shortened to reduce the computation time of the matching algorithm and to reduce the influence of rapid pitch changes which are likely to be performance specific. To do so, every 150 frames of 2.9ms were summarized by producing a pitch class histogram and selecting the bin with the highest occurrences. This strategy was based on the work of [17].

Matching the melodies was accomplished with a modified version of the Smith-Waterman algorithm suggested by [8], to work with melody representations. The modifications affected the weighted distance matrix and the gap penalty. The distance values of the weighted distance matrix were chosen based on how further the semitones are from each other. The gap penalty used was -8. The distance value returned for each match was used to construct the distance matrix between all melodies.

Two melody representations were tested. One without negative values and another with both positive and negative. In both cases, the zero values were not considered, since it could result in high matching values due to long periods of silence.

4.4. Results and Discussion

Overall, three kinds of experiments were made with different aspects and approaches. The results obtained are presented in Table 1 with the corresponding evaluation metrics. The results obtained by the baseline are also included to serve as a basis for comparison. The results obtained by fusing different approaches, using the proposed solution of Degani et al. [6], are also shown.

The results relative to the summarization and rhythm experiments proved to produce poor results, with the best score obtained by a summarization method being 75 successfully identified covers and 44 by the rhythm approach. In some way, this was expected because these approaches create a short representation, meaning that information useful for cover detection can be lost while selecting the values. An interesting observation was made when the results of these experiments were fused with the baseline results. Despite having had worse results by itself compared to most of the summarization methods, fusing the baseline with the rhythm resulted in a less significant loss of quality when compared to the best summarization method. This can mean that the rhythm representation is, in some way, connected to its original song but not in the best way. One can also deduce that the presence of similar rhythms on its own is not a good indicator of being a cover or not. Due to the poor quality of the results of these approaches, they were not included in further tests.

Melody-wise, as previously mentioned in section 4.3.3, two representations were made: one with only positive values and another considering positive and negative values. The results revealed that the positive-only representation is the best on its own. However, fusing with the baseline scores seems

[5]http://mtg.upf.edu/technologies/melodia

[6]Vamp is an audio processing plugin system that allows the use of plugins to extract information from audio data.

[7]http://www.vamp-plugins.org/sonic-annotator/

to favor, although slightly, the representation with all values. By combining these two representations and the baseline, the best result of the experiments was achieved, suggesting that the two melody representations interact well with each other. The difference between the two representations and why they concentrate on different characteristics is explained by the quantization made using histograms of 2.9ms. With longer length and more values on the positive and negative representation, the octave with most occurrences will most likely be different, therefore highlighting different characteristics than the positive-only one.

To further understand the influence of each representation, data relative to the difference in covers being detected among approaches was collected. That data is shown in Table 2 and consists of counting what are the common correctly identified covers, the shared non-covers that were incorrectly identified, the number of covers that were initially identified and then were lost, the non-covers that initially were not present in the Top10 and then were incorrectly identified there, and the total number of new covers successfully identified with the improved similarity matrix. The tests made were:

- T1: Baseline results vs. fusion of the baseline with positive-only melody.
- T2: Baseline results vs. fusion of the baseline with all-values melody.
- T3: Baseline results vs. fusion of the baseline with all-values melody and positive-only melody
- T4: Comparison between all-values melody and only-positive melody.

Table 2: Statistical analysis of the changes in cover detection

	T1	T2	T3	T4
Common covers	983	1003	947	236
Common non-covers	1582	1745	1207	461
New covers	136	118	211	-
Lost covers	66	46	102	-
New non-covers	41	1	79	-

Although T1 showed that the positive-only melody representation has a higher count of new covers identified, it also contributed with more false positives and had fewer correctly identified covers in common with the baseline. T2 in the other hand, despite having detected fewer new covers, revealed that the all-values representation preserves the quality of the results of the baseline, with a higher count of common covers and only one new false positive. To further prove that effect, the correlation coefficient values were estimated between the similarity matrices obtained from each approach. These results are included in Table 3. This confirmed that the matrix obtained from the baseline has a higher correlation coefficient with the all-values melody representation, explaining why the fusion between it and the baseline reveals a high count of common covers.

Table 3: Correlation coefficients results

	Baseline	Positive-only Melody	All-values Melody
Baseline	-	0.2463	0.3890
Final	0.8142	0.7170	0.7779

The combination of both melody representations and the baseline, as shown by T3, hinders the quality of the results provided by the baseline but, on the contrary, seems to strengthen the results of both melody approaches. Since the three matrices are fused together, all of them are considered equally important and, individually, the results from both melody representations are worse than the baseline. Therefore, it is expected that the number of common covers would decrease and the count of new true positives would increase, as well as the number of false positive. Such is the case, but the new covers vs. lost covers ratio is good enough to have better results compared to the baseline.

The correlation coefficient values for this 3-method fusion matrix and each of the composing matrices were also calculated in order to assess the contribution of each of them in the final result. The baseline results were the ones with higher correlation, as it would be expected, followed by the all-values melody representation and, with the lowest correlation coefficient, the positive-only melody representation.

With the objective of trying to understand the impact of fusing the results and what changes were made to those achieved with the baseline, some of the changes in the Top10s were analyzed. Specifically, some songs were listened to and compared to see if there was any pattern that could be identified. The songs that were listened were new covers being successfully identified, covers that were present in the baseline but were lost in the best method and songs that were included in the Top10 that are not covers.

The new covers being detected consist of songs that are, in general, very closely related and easily detected by a human judge. There were compatibility between songs of different genres and styles but possess a similar melodic line and/or lyrics. This means that the goal of using melody extraction was met. There was also cases of matches between pairs of songs in which one of them did not have any

lyrical elements or completely different instruments. There was an interesting case between a song that was entirely instrumental and the other one was in acapella style that were correctly identified as covers due to its similar melody.

The covers that were lost by the fusion process presented two patterns. One was related to the singing style of the performer, that deviated from the query song and the other consisted of different tempos, either in singing or instrumentation. Since the melody is extracted every 2.9ms, differences in tempo will result in different frequencies being detected.

There was no clear pattern in the songs incorrectly identified. Since there is a high number of false-positives being detected by the melodies alone, it is understandable that some would benefit from it and have their scores mistakenly boosted.

Having found an improvement relative to the results in the baseline, it can be assumed that features about melody and voice are useful to the task of identifying audio cover songs by using a fusion heuristic. Using a fusion heuristic allows the inclusion of different audio features, since it is a fast and scalable method that requires the matrices to be computed beforehand only once. One alternative version of the fusion heuristic suggested by [6] was made that had a minor change in how the distance was computed: instead of using the suggested refined distance value, the sum of the three distance values was computed; it was observed that the difference was very slim, with only three less covers being detected in total. While it can be argued that it is less complex and requires less computation effort, the original solution is still better and still fast to compute for, at least, three 1000x1000 matrices.

5. Conclusions

This paper reviewed the musical facets that condition the process of identifying an audio cover song and presented the most common approaches used to take advantage of those facets. A baseline was used and its results analyzed. In a general sense, it was able to correctly identify covers with similar instrumentation, and failed to detected those with the different instrumentation but with similar lyrics, rhythm, or melody. Experiments relative to summarization, rhythm, and melody, were realized in order to improve the identification of the covers that failed to be detected by the baseline. For the melody, two representations were made. One considering only voiced segments, and the other considering voiced and non-voiced. The matching of melody descriptors was accomplished using a modified Smith-Waterman algorithm to handle semitone representation.

While the summarization and rhythm experiments displayed no improvement, the fusion of the results from both melody descriptors with the results from the baseline, resulted in an increase of 5% in MAP and 109 more covers being detected. The new covers being detected displayed similar melodic lines and/or lyrics to the query song, while having a different instrumentation or genre. This proved that 1) the heuristic for distance fusion does enhance qualities from various methods, 2) melody is beneficial to cover song identification, and 3) voice related descriptors, such as the melody descriptors used, are useful to detect covers with similar lyrics.

Future work includes finding a way to improve the heuristic in order for it to not impair the good results of the baseline. Some of the covers were affected by the poor value returned by the melody and so they were not detected in the improved version. The melody descriptors were also affected by different tempos, and thus, making them tempo invariant would improve their quality.

Acknowledgements

I would like to thank my advisor Doctor David Martins de Matos for giving me the freedom to choose the way I wanted to work and for all the advices and guiding provided. I would like to thank L2F and INESC-ID for providing me with the means I needed to produce my work. I would like to thank Teresa Coelho for making the dataset that proved to be extremely useful for me to test all my work and also Francisco Raposo, for providing me summarization versions of that dataset in order for me to conduct my experiments. Last but not least, I would like to thank my friends and family for all the support and strength given to keep me focus on this journey. None of this would be possible without these people.

References

[1] P. G. Antunes, D. M. de Matos, R. Ribeiro, and I. Trancoso. Automatic fado music classification. *CoRR*, abs/1406.4447, 2014.

[2] J. V. Balen, D. Bountouridis, F. Wiering, and R. Veltkamp. Cognition-Inspired Descriptors for Scalable Cover Song Retrieval. In *ISMIR'14*, pages 379–384, 2014.

[3] T. Bertin-Mahieux and D. P. W. Ellis. Large-Scale Cover Song Recognition Using the 2D Fourier Transform Magnitude. In *ISMIR'12*, pages 241–246, 2012.

[4] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4):668–696, Apr. 2008.

[5] T. Coelho, C. Duarte, and D. Matos. Coda dataset description. 2015.

[6] A. Degani, M. Dalai, R. Leonardi, and P. Migliorati. A Heuristic for Distance Fusion in Cover Song Identification. In *WIAMIS'13*, pages 1–4, 2013.

[7] D. P. W. Ellis and G. E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, volume 4, Apr. 2007.

[8] J. D. Frey. *Finding Song Melody Similarities Using a DNA String Matching Algorithm.* PhD thesis, Kent State University, College of Arts and Sciences, 2008.

[9] E. J. Humphrey, O. Nieto, and J. P. Bello. Data Driven and Discriminative Projections for Large-Scale Cover Song Identification. In *ISMIR'13*, pages 149–154, 2013.

[10] M. Khadkevich and M. Omologo. Large-Scale Cover Song Identification Using Chord Profiles. In *ISMIR'13*, pages 233–238, 2013.

[11] O. Lartillot and P. Toiviainen. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 127–130, Vienna, Austria, September 23-27 2007.

[12] Y. Lu and J. E. Cabrera. Large scale similar song retrieval using beat-aligned chroma patch codebook with location verification. In *SIGMAP'12*, pages 208–214, 2012.

[13] B. Martin, D. G. Brown, P. Hanna, and P. Ferraro. BLAST for Audio Sequences Alignment: A Fast Scalable Cover Identification Tool. In *ISMIR'12*, number Ismir, pages 529–534, 2012.

[14] G. Mitri, V. Ciesielski, and A. L. Uitdenbogerd. Automatic music classification problems. In V. Estivill-Castro, editor, *Twenty-Seventh Australasian Computer Science Conference (ACSC2004)*, volume 26 of *CRPIT*, pages 315–322, Dunedin, New Zealand, 2004. ACS.

[15] F. Raposo. Influence of summarization on music classification tasks. Master's thesis, Instituto Superior Técnico, Lisboa, Portugal, 2014.

[16] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20:1759–1770, 08/2012 2012.

[17] J. Salamon, J. Serrà, and E. Gómez. Tonal representations for music retrieval: From version identification to query-by-humming. *International Journal of Multimedia Information Retrieval, special issue on Hybrid Music Information Retrieval*, 2:45–58, 2013.

[18] J. Serrà. *Identification of versions of the same musical composition by processing audio descriptions.* PhD thesis, Universitat Pompeu Fabra, Barcelona, 2011. TDX link: http://hdl.handle.net/10803/22674.

[19] J. Serrà, E. Gomez, and P. Herrera. Transposing Chroma Representations to a Common Key. In *IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects*, pages 45–48, 2008.

[20] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, 08/2008 2008.