ISSN 1870-4069

Cover Song Recognition Using Machine Learning Techniques

Andree Silva-Reyes, Fabiola Martínez-Licona, Alma Martínez Licona

Universidad Autónoma Metropolitana, Electrical Engineering Depto.,
Mexico

andree_sr@hotmail.com, {fmml, aaml}@xanum.uam.mx

Abstract. The task of recognizing a song as a cover version of another is relatively easy for the human being, when the song is known. However, making a machine to do this job is complex because of the number of variables involved in the development of a cover; these include variations in tempo, instrumentation, gender, and duration with respect to the original version. A methodology that aims to identify covers from the application and analysis of machine learning techniques, sparse codification, signal processing and second order statistics, in order to obtain the best configuration, is proposed. Acoustic features such as pitches and timbres, as well as beat information of the cover songs were obtained from the Million Song, a metadata database oriented to music information retrieval. Along the experimentation it was able to try different analysis configurations on the metadata and to appreciate the effects on the comparisons between original and cover versions. According to the results, a system that integrates a frequency processing on the pitches with beat alignment, a sparse codification and a clustering technique was obtained with correct cover identification similar to the state of the art results. It was also possible to get information about learning techniques combinations with different metrics that allows future experiments to improve the results.

Keywords: cover song recognition, clustering, sparse codification.

1 Introduction

The intersection among music, machine learning and signal processing has let to address a wide range of task such as automatic definition of melodies, chords and instruments, identification and characterization of long term times and structures or recognition of musical genres and covers [1]. There are organizations like the International Information Society for Recovery of Music Information (ISMIR), or the Music Information Retrieval Evaluation Exchange (MIREX), that have promoted the use of these fields to the access, organization and understanding of musical information, focusing on the research and development of computational systems that aim to solve these series of tasks.

A musical version, or cover, is defined as a new interpretation, live or in studio, of a song previously recorded by another artist [2]. This implies that a musical cover may have shifts in the rhythms, tempo, instrumentation ranges, gender or duration with respect to the original version. As an example the song Summertime, originally performed by Abbie Mitchel in 1935, has up to 1200 musical covers to date according to the project Second Hand Songs[1]. Some of these are in general similar to the original song and some others are quite different; the Million Song Dataset[2] reports versions in the musical genres of jazz, rock-pop, rhythm & blues and even country. A Cover Identification System (CIS), is an automatic system that ideally determines if a song is a cover version of some musical piece located in a database. This problem has been addressed by applying several methods based in two stages: the first stage consists on the extraction and the analysis of the most important characteristics of the song such as its melodic representation, harmonic progression or pitch; the second stage aims to the measurement of the similarity degree between the features extracted from each piece of music.

Some previous work on the subject has aimed to solve these steps by proposing different methods. Lee showed an extraction method based on Hidden Markov Models applied to sequence of chords for each song and followed by a similarity degree measurement between chord sequences, using dynamic time warping [3]; the problem with this technique is that it needs a huge amount of time and computational resources. In [4] Jensen et.al calculated a Chroma-gram, a matrix from the Chroma vector sequences, that was not sensitive neither to instrumentation nor time changes, to obtain the minimum distance of the matrices using the Frobenius norm. Ravuri and Ellis [5], proposed to obtain the Chroma-gram and calculate three characteristics per song to classify by means of vector support machines (SVM), or multilayer perceptron (MLP), while Chuan proposed to calculate a Chroma-gram that saves the partial harmonics of the melody and maintain the volume invariance, to make a framework that measures similarity by means of a binary classifier [6].

In [7], a method inspired on the creation of digital fingerprints used to minimize the execution times in the search for covers is proposed by Bertin-Mahieux and Ellis. This research makes use for the first time of the database MSD (Million Song Dataset) which consists of characteristics and metadata for a million songs under the Creative Commons (CC) license; in [8] an introduction to MSD, as well as its creation process and possible uses are presented by the same authors plus Lamere. The search process in a database may be accelerated by dividing a song into small fragments that may be used as hashes; in [9] Grosche and Maller applied this technique, but with larger segments for each song in order to minimize el number of searches. In [10] Bertin-Mahieux and Ellis used the 2D Fourier Transform to procure a representation of the Chroma patches to obtain an efficient nearest neighbor algorithm; this scheme makes

[1] http://www.secondhandsongs.com/statistics

[2] https://labrosa.ee.columbia.edu/millionsong/

the nearest neighbors to have more probabilities of being related to the same song. A couple of modifications of this method are presented in [11] to improve the classification; the authors used data dispersion to enhance their separability followed by a dimensionality reduction.

Two methods that help making queries in a faster way on a large data base are proposed for cover identification: Basic Alignment Search Tool (BLAST), which is a bio-sequence indexation technique that Martin et. al used to increase the efficiency [12], and a data base prune method that Osmalskyj et. al used to reduce the set where the search is made [13]. The Locally-Sensitive Hashing (LSH), method was used by Khadkevich and Omologo in [14] to obtain similar chord songs and then to apply a progression method to refine the search ranking. On the other hand, Salamon et al focus on extracting tonal representations (melody, bass line and harmonic progression), by using state of the art algorithms, and a dynamic programming algorithm to measure the degree of similarity in [15]. From these works it is concluded that harmonic representations are more reliable for cover identification although the tonal representations better improves the recognition accuracy.

Van Balen et al use three descriptors to content-based music recovery: pitch bi-histogram, Chroma correlation coefficients and harmonization features [16]. Serrà explains the steps involved in a CIS: feature extraction, key invariance, tempo invariance, structure invariance and similarity calculation [17], while in [16], the authors group these steps into a two phases system: calculation of harmonic features and pitch for each song and comparison of the similarity.

This paper presents the results obtained from a CIS based on machine learning techniques. This is based on the state-of-the-art feature processing plus the introduction of a sparse codification in order to obtain a reduced size feature vector. The components to separate the main melody from the accompaniment were selected, the machine learning architecture for cover recognition was determined and the first test adjustments were carried out. Followed, a final test and the performance assessment were developed.

2 Methods

The CIS presented here was based on the work of Bertin-Mahieux and Ellis [10]. Although some other research is aimed to classify musical covers, these authors used the MSD database by applying a very precise methodology that allows obtaining a comparative framework with our results. In general, the tasks done in [10], are to get the Chroma features of MSD, to align the Chroma features with the beat, to apply the power law on the resulting matrix, to generate patches and to calculate the 2D Fourier transform, to calculate the median and finally to apply PCA. Figure 1, shows the methodology used here on each one of the songs.

In general, the characteristics of the songs were obtained from an adaptation of the Chroma and timbre data taken directly from the database. A spectral analysis was applied to the resulting matrices in order to get a 2-D Fourier Transform (2D-FFT), a Wavelet Transform (WT) and patterns generated from the algorithm named Sparse

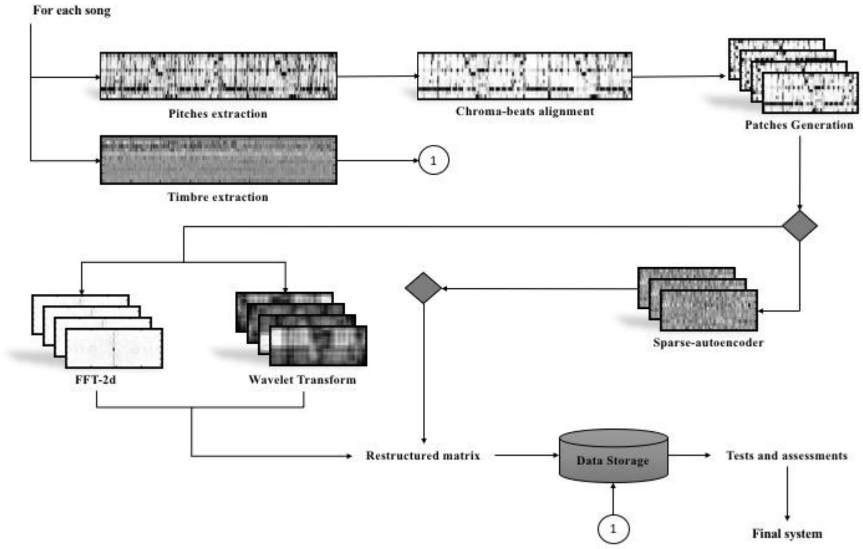

Fig. 1. Methodology applied to the individual song based on [10].

Autoencoder. WT and Sparse Autoencoder were added to the techniques showed in [10]. Cover recognition was made from metrics used to assess the ensemble clusters of each song.

2.1 Database

The Million Song Dataset was used to extract the songs and their covers. From this dataset it was extracted a subset named Second Hand Songs (SHS), that is oriented to original and cover songs[3] and is divided into training and test subsets [18].

2.2 Feature Extraction

Each musical piece included in MSD consists of a set of 55 attributes including song descriptors such as artist name, song title, labels with similar artists, and acoustic descriptors of the song such as pitches, timbre, volume and beat.

The first step was to determine the most representative attributes and the way to manipulate this information. This was done by selecting the most relevant features of the MSD songs listed on Table 1 through different classification methods including, Naive Bayes, Multilayer Perceptron and Support Vector Machine. Initially all attributes in the database were used, and they were subsequently removed one by one

[3] http://labrosa.ee.columbia.edu/millionsong/secondhand

in each new experiment in order to determine those that impacted the most on the correct classification of the songs. As a result, and in consistency with [7] and [19], pitches and timbre attributes were selected as features for the CIS development.

The pitches are the consequence of a perceptive property that allows ordering the sounds on a frequency-related scale [20]. In MSD pitch information is given by a set of Chroma vectors of length 12 that represents the notes with their semitones according to the Chromatic scale: C, C #, D, D #, E, F, F #, G, G #, A, A #, B. Each note is assigned a normalized numerical value, [0 1], which describes the relative dominance of each pitch on the time segment previously defined and named “musical event”.

Timbre is the quality of a musical note or sound that distinguishes different types of musical instruments or voices [21]. In MSD timbres are derived from the spectral-temporal surface of a segment and it is represented by a vector of length 12, where the four first coefficients are related to the average segment volume, sound brightness, sound flatness and sound strength, while the remaining have no concrete meaning so they were left out of the final characteristic vector.

2.3 Chroma-Beats Alignment

Beat is a basic unit for time measurement in music. Time variability over the different cover songs are controlled when they are aligned with the beat, so this alignment was applied on the pitches. The Chroma-beat vector was made taking the weighted average of the Chroma vectors included on the time segments of each beat. In order to increase the contrast in the Chroma values, the Chroma-beat vector was raised to 1.96 power according to [10].

Table 1. Song titles used to determine the set of attributes for CIS.

Song title	Number of Covers
Summertime	47
Silent Night	43
White Christmas	37
Body and Soul	19

2.4 Patches Generation

All the Chroma vectors belonging to the same song conform the Chroma matrix of this song. This matrix is segmented into a set of patches that were formed by taking 75 Chroma vectors and shifting by one Chroma vector, in order to obtain a number of [12x75], patches which value depends on the duration of the song. Next, the 2D-FFT and the WT as well as the Sparse Autoencoder, were applied to the patches data to get the song descriptors that will be used in the comparison among the covers.

2.5 Sparse Autoencoder

As described in [22], an autoencoder is a neural network that uses an unsupervised learning algorithm based on backpropagation in order to obtain the output values that most closely resemble the input ones, thus attempting to reproduce their own inputs. The architecture is similar to these of MLP, having input, output and hidden layers but the output layer has the same number of nodes as the input layer, to try to reproduce the values that enter the network, and the number of nodes of the hidden layer is less than the nodes of the input layer so the network is forced to turn the data representation into a smaller version. This feature makes Sparse Autoencoder (SA), to be used for dimensionality reduction purposes, so it allowed managing the amount of pitches data of each cover song, something “not covered” in [10].

The SA algorithm was applied on each song and on the set of songs with the same title in order to get one pattern that represents the individual song and other for the ensemble title. The set of patches for the song was restructured into a [900xN] array, where the [12x75], patches were converted to a [900x1], vector and N was the number of patches of the song. SA was applied to this array to achieve a [900xM], matrix, where M was the number of patterns calculated from the algorithm; after different testing configurations this value was set to 30, so the [900x30], matrix was restructured into a 30 [12x75], patterns that, according to the SA theory, represent the characteristics of the musical piece. The same procedure was applied to the ensemble of songs, so it was obtained a set of 30 [12x75], patterns per title.

2.6 2-D FFT and Wavelet Transforms

The set of patches obtained from each song were taken and transformed one by one by the 2-D FFT. Next, these Fourier transformed patches were integrated into an array of [Nx900], where each row represents a restructured patch from [12x75] to a [900x1] vector, and N continues to be the number of patches of the song. The same procedure was performed for the ensemble of songs with the same title. The set of 30 [12x75], patterns obtained from the application of the SA on the data was transformed with the 2-D FFT and restructured to a [30x900], M matrix

Wavelet Transform (WT), allows a signal to be represented as small time-frequency scale components and is suitable for the kind of signals that presents variations or abrupt discontinuities [23].

WT was applied, in addition to the 2-D FFT, to both the patches and the SA patterns of each cover song and the ensembles following the same procedure as in 2D-FFT. The wavelets used were Morlet, Sinc, Gauss, Meyer and Daubechies.

2.7 Cover Song Recognition Experiments

Different configurations were tested in order to find out the best approximation to a cover song identification system. The following experiments were carried out for this purpose:

EI. Cover song identification by comparison with a general ensemble bounded by a threshold. This experiment was divided in two stages. In the first stage a general descriptive ensemble of songs with the same title was obtained and the maximum value was calculated to use it as the threshold for the comparison of songs. The second stage consisted of making comparisons to determine whether a piece of music was a cover song or not. It was used the aligned array and the patch matrix, working with them separately. The general ensemble was obtained from the set of songs with the same title. The statistical measurements of mean, median and standard deviation (SD), were calculated from the set of patches and the aligned arrays of each song belonging the same group, and the descriptive ensemble was formed by the average of these statistical measures. The threshold was calculated by measuring the distances (subtraction, Mahalanobis distance and Euclidean distance), between each song and the general ensemble. The tests allowed to determine the degree of similarity between the general ensemble and other cover songs with the same title that were not included in its creation.

EII. Cover song identification by comparison with a general pattern bounded by a threshold. The same process as used in the previous section was applied to the set of SA patterns of each song. A stage of obtaining the general patterns of the group of songs with the same title was carried out with the statistical measurements of mean, median and standard deviation (SD), and the local entropy; next the average of these statistical measurements were computed. The distance measurements of subtraction, Mahalanobis and Euclidean, were used to obtain the threshold distance between each song patterns and the general pattern of the song.

EIII. Cover song identification using supervised (MLP) and supervised (K-nn) machine learning techniques and a representation of the data using 2D-FFT and WT. In pattern recognition k-nn is a non-parametric method used for regression and classification, based on measures of similarity or distance functions. The median of the [900xN] 2D-FFT matrix patches from the reference song and the other songs to compare, was calculated resulting in a [900x1], vector per song. Then, the k-nn algorithm was applied using the Euclidian, SEuclidian, City-block, Minkowski, Chebyshev, Cosine, Correlation and Spearman distance measurements. The same steps were repeated on the 2D-FFT of the patterns, the TW of the patches and the TW of the patterns. Comparisons were made in order to determine which characteristics allowed a better clustering of the data leading to the configuration that allowed to have a cluster of cover songs.

The procedure for the application of MLP was quite similar. The median of the matrices 2D-FFT and WT of both the patches and the patterns of the reference and other songs was calculated and then different configurations of MLP neural networks were tested in order to find the best cover song classification scheme.

EIV. Cover song identification using self-organized maps (SOM) on timbre features. SOM is a computational method for visualization and analysis of high-dimensional data [24]. They automatically organize the data so similar entries are mapped by being close to each other. SOM were applied on the first four coefficients of the timbre matrix directly obtained from MSD. The objective was to observe the way

Table 2. Best results for % of correct non-cover song identification by comparison with a general ensemble bounded by a threshold (EI) tests.

Subtraction (%)	Mahalanobis (%)	Euclidean (%)	Data Used	Statistical Meas
6.2	0.3	1.5	Aligned matrix	Mean
6.2	0.2	1.4	Patches	Mean
5.7	0	2.5	Patches	Median

Table 3. Best results for % of correct cover song identification by comparison with a general ensemble bounded by a threshold (EI) tests.

Subtraction (%)	Mahalanobis (%)	Euclidean (%)	Data used	Statistical Meas
56.8	29.1	42.9	Aligned matrix	Mean
54.8	25.6	45.8	Aligned matrix	Median
53.4	27.5	42.4	Patches	Mean

in which timbre coefficients were grouped in the case of having songs of the same title, and to identify if there were differences between cover and non-cover songs.

3 Results

The data set for the analysis and comparison was made from those SHS examples with at least 15 cover songs; the total number of cover songs was 417, corresponding to 19 different titles. As for the comparison tests, it was determined that each one was formed by five songs and an identifier of the test group: two covers of the same song and three other different songs. All songs were randomly selected, having a total of 1000 tests, which were used to assess the precision of the algorithms of each experiment.

The best results achieved in the cover song identification by comparison with a general ensemble bounded by a threshold experiment (EI), are shown in Table 2, for the non-cover song tests and in Table 3 for the cover song tests.

The best results obtained in cover song identification by comparison with a general pattern bounded by a threshold (EII), are shown in Table 4 for the non-cover song tests and in Table 5 for the cover song tests.

The Cover song identification using the supervised (K-nn), machine learning techniques and a representation of the data using 2D-FFT and WT (EIII), results are shown in Table 6. The best EIII correct results using the supervised MLP are shown in Table 7.

To perform the cover song identification using self-organized maps (SOM), on timbre features (EIV), the algorithm was trained with a cover and a non-cover songs of the group and the test was made with the remaining cover and non-cover songs. The overall result of correct assignation was 48.4% on the 1000 tests.

Table 4. Best results for % of correct non-cover song identification by comparison with a general pattern bounded by a threshold (EII) tests.

Cosine Distance (%)	**Cityblock Distance(%)**	**Euclidean Distance(%)**	**Data used**
0	2.4	2.4	Entropy SD
1	0.7	0.5	Pattern SD
0.1	0.3	0.1	Pattern median

Table 5. Best results for % of correct cover song identification by comparison with a general pattern bounded by a threshold (EII) tests.

Cosine Distance (%)	**Cityblock Distance(%)**	**Euclidean Distance(%)**	**Data used**
36.1	57.8	57.4	Entropy SD
31.7	34.8	34.6	Entropy median
31.9	34.6	34.9	Entropy median

Table 6. Results for cover song identification (%) tests using K-nn, (EIII).

Euclidean	**Minkowski**	**Cosine**	**Correlation**	**Matrix**	**Wavelet**
61.5	61.5	64	63.3	2D-FFT pattern	
60.2	60.2	63	56.5	2D-FFT patches	
42.4	42.4	40	32.9	WT pattern	Sinc
44.1	44.1	40.8	31.6	WT pattern	Gauss
45.3	45.3	40.5	34.6	WT patches	Morlet

Table 7. Results for cover song identification tests using MLP, (EIII).

% Correct classification	**Matrix**
49.5	2D-FFT pattern
47.5	2D-FFT patches
36.6	WT patches
36.2	WT pattern

4 Discussion and Conclusions

The results show several issues concerning to the analysis tools and learning techniques, as next presented.

1. *Cover song identification by comparison with a general ensemble bounded by a threshold*. Results in Table 2, show that it was not possible to separate the set of non-cover songs from the rest of the songs since most of them were below the threshold of the ensemble. For cover songs the best result was achieved with the mean of the aligned matrix using subtraction as a distance metric as shown in Table 3.
2. *Cover song identification by comparison with a general pattern bounded by a threshold*. As in the previous case, the results in Table 4 show that the threshold was not able to distinguish the non-cover songs from the rest while Table 5 shows that the best results for cover songs was achieved with the standard deviation of the data entropy using CityBlock distance as a metric. The pattern obtained by applying the Sparse Autoencoder allowed to reduce the data dimension and also to have another kind of representation for the song that might be of interest.
3. *Cover song identification using supervised (MLP), supervised (K-nn) machine learning techniques and a representation of the data using 2D-FFT and WT*. Tables 6 and 7 show that the supervised K-nn performed better and Fourier transform showed more relevant data than WT. Music as a signal can present a lot of events that change in time and frequency and the wavelet transform seemed to be a suitable signal processing tool to show them. Nevertheless the WT processing in metadata, as those of the MSD, did not achieve good results so this transform would be more useful if it is applied on raw data.
4. *Cover song identification using self-organized maps (SOM) on timbre features*. Timbres did not provide an improvement on the results. On the other hand, the SOM maps revealed some difficulties for the identification of the cover song group, but also showed keys to be identified that might be used in other schemes in order to increase the results.

Some other relevant aspects to consider are related to the data. MSD database contains a lot of records and a musical content variety that let the algorithms to be tested in a huge amount of data.

This is convenient since music databases exponentially grow every time. Concerning the features used in the experiments, timbres and pitches were the most relevant in terms of impact on the cover classification as is shown on a previous binary classification tests using Naive Bayes.

According to the results the best configuration achieved is shown in Figure 2. Likewise, the best result obtained was 64%, of correct cover song identification, which is close to the reported results (66-82%) in [10], although they are not the same tests.

In order to verify the effectiveness of the presented methodology, the experimentation performed in [10], was reproduced; 500 binary queries were carried out under the following scheme: given a song A that served as a query and two songs B and C, determine whether B or C was the cover of A. The results obtained in the reference were 82%, of covers identified correctly without applying PCA and 82.2%, with PCA with 50 major components on their data. In order to make the comparison the same scheme was used with other list of 500 queries taken from the same database. The results obtained were 75%, of covers identified correctly without applying PCA

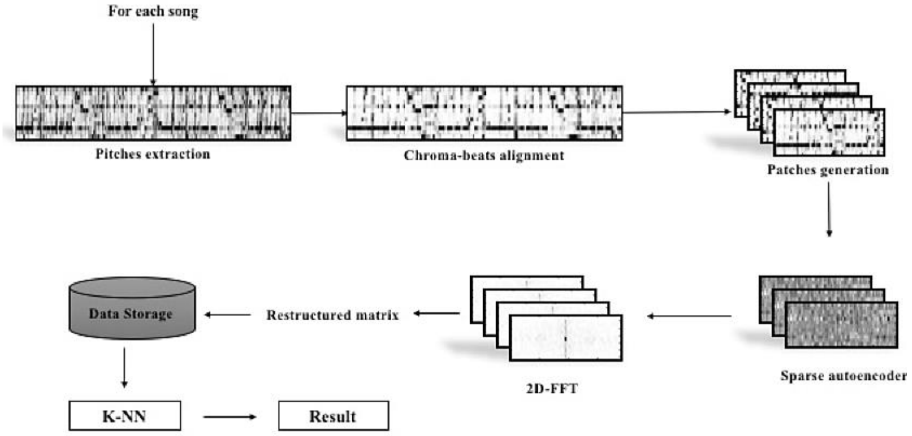

Fig. 2. Best configuration obtained of the CIS.

and 78.8% with PCA with 50 major components. By applying to the same data the CIS configuration shown in Figure 2, the best result was 73.8% (of correctly identified covers).

Further work is addressed on using raw audio signals for the analysis and the dealing with the unbalance issues that this task presents, the improvement on the parameter selection and preprocessing, and the experiment with other machine learning techniques such as Support Vector Machines or deep learning neural networks and apply inference systems like decision trees.

Acknowledgement. The first author would like to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico, for a grant given to him by them.

References

1. Humphrey, E., Bello, J. P., Le-Cun, Y.: Moving Beyond Feature Design: Deep Architectures and Automatic Learning. Music Informatics, Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR), pp. 403–408 (2012)
2. Oxford Dictionary: https://en.oxforddictionaries.com/definition/cover (2017)
3. Lee, K.: Identifying cover songs from audio using harmonic representation. MIREX task on Audio Cover Song Identification (2006)
4. Jensen, J. H., Christensen, M. G., Ellis, D. P. W., Jensen, S. H.: A Tempo-Insensitive Distance Measure for Cover Song Identification Based On Chroma Features. Proceedings of IEEE International Conference on Acoustics, (ICASSP), Speech and Signal Processing, pp. 2209–2212 (2008)

5. Ravuri, S., Ellis, D. P. W.: Cover Song Detection: From High Scores to General Classification. Proceedings of IEEE International Conference on Acoustics, (ICASSP), Speech and Signal Processing, pp. 65–68 (2010)
6. Chuan, X.: Cover song identification using an enhanced Chroma over a binary classifier based similarity measurement framework. International Conference on Systems and Informatics (ICSAI), pp. 2170–2176 (2012)
7. Bertin-Mahieux, T., Ellis, D. P. W.: Large-scale cover song recognition using hashed Chroma landmarks. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, (WASPAA), pp. 117–120 (2011)
8. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., Lamere, P.: The million song dataset. Proceedings of the 12th International Society for Music Information Retrieval, Conference (ISMIR) (2011)
9. Grosche, P., Müller, M.: Toward characteristic audio shingles for efficient cross-version music retrieval. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 473–476 (2012)
10. Bertin-Mahieux, T., Ellis, D. P. W.: Large-Scale Cover Song Recognition Using the 2D Fourier Transform Magnitude. Proceedings of the 13th International Society for Music Information Retrieval, Conference (ISMIR), pp. 241–246 (2012)
11. Humphrey, E., Nieto, O., Bello, J.: Data driven and discriminative projections for large-scale cover song identification. Proceedings of the 14th International Society for Music Information Retrieval, Conference (ISMIR), pp. 149–154 (2013)
12. Martin, B., Brown, D., Hanna, P., Ferraro, P.: BLAST for Audio Sequences Alignment: A Fast Scalable Cover Identification Tool. Proceedings of the 13th International Society for Music Information Retrieval, Conference (ISMIR), pp. 529–534 (2012)
13. Osmalskyj, J., Piérard, S., Van Droogenbroeck, M., Embrechts, J. J.: Efficient Database Pruning for Large-Scale Cover Song Recognition. Proceedings of IEEE International Conference on Acoustics, (ICASSP), Speech and Signal Processing, pp. 714–718 (2013)
14. Khadkevich, M., Omologo, M.: Large-scale cover song identification using chord profiles. Proceedings of the 14th International Society for Music Information Retrieval, Conference (ISMIR), pp. 233–238 (2013)
15. Salamon, J., Serrà, J., Gómez, E.: 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(1), pp. 45–58 (2013)
16. Balen, J. V., Bountouridis, D., Wiering, F., Veltkamp, R. C.: Cognition-inspired Descriptors for Scalable Cover Song Retrieval. Proceedings of the 15th International Society for Music Information Retrieval, Conference (ISMIR), pp. 379–384 (2014)
17. Serrà, J., Gómez, E., Herrera, P.: Audio cover song identification and similarity: background, approaches, evaluation and beyond. Z. W. Ras, A. A. Wieczorkowska (eds.) Adv. In Music Information Retrieval, Studies in Computational Intelligence, 16, pp. 307–332 (2010)
18. Second Hand Songs dataset: http://labrosa.ee.columbia.edu/millionsong/secondhand (2017)
19. Patterson, R. D., Gaudrain, E., Walters, T. C.: The Perception of Family and Register in Musical Tones. Mari Riess Jones, Richard R. Fay, Arthur N. Popper. (eds.) Music Perception, pp. 37– 38 (2010)
20. Klapuri, A., Davy, M.: Signal processing methods for music transcription (2006)

21. Tristan, J., CSO, DesRoches, D.: Lead Audio Engineer. Analyzer Documentation, The Echo Nest Corporation (2016)
22. http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf (2014)
23. Andrew Ng: Sparse Autoencoder. CS294A Lecture notes, http://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf (2017)
24. Meyer, Y., Bartram, J. F.: Wavelets and applications. The Journal of the Acoustical Society of America, 92(5), pp. 3023–3023 (1992)
25. Rojas, R.: Neural Networks (1996)