LARGE-SCALE COVER SONG RETRIEVAL SYSTEM DEVELOPED USING MACHINE LEARNING APPROACHES

Tzu-Hsiang Huang
Communication Engineering Department
National Central University
Taoyuan, Taiwan

Pao-Chi Chang
Communication Engineering Department
National Central University
Taoyuan, Taiwan

***Abstract*—Large-scale cover song retrieval systems should be able to calculate song-to-song similarity and accommodate differences in timing, key, and tempo. Simple vector distance measure is not adequately powerful to perform cover song recognition, and expensive solutions such as dynamic time warping do not scale to millions of instances, making cover song retrieval inappropriate for commercial-scale applications. In this work, we used the content-based music features of songs as input and transformed them into vectors by using the 2D Fourier transform approach. Furthermore, we explored different machine learning approaches to reinforce the pattern of these vectors. By projecting the songs into a semantic vector space, we can use the efficient nearest neighbor algorithm to compare the similarity of songs and retrieve the most similar songs from the large-scale database. The proposed system is not only efficient enough to perform scalable content-based music retrieval but can also develop the potential of machine learning approaches, making similar music recognition applications faster and more accurate.**

Keywords- large-scale music information retrieval; 2D Fourier transform; machine learning; million song dataset

I. INTRODUCTION

In the big data era, methods to track and manage music similar to a specific work are in demand. Traditional retrieval methods search according to text-based clues, which leads to confusion. Present-day systems and research use content-based information retrieval to solve many problems. In this work, we developed a content-based music information retrieval (MIR) system, which promises many potential applications. The content-based retrieval system can find not only the exact copy of a given music track but also the novel versions of the original work, for example, "cover songs." This has several applications. For example, when copyrighted music has been identified, the copyright holder can ensure the correct handling of songwriting royalties. In addition, listeners can be offered music recommendations according to content similarity. Moreover, researchers can analyze patterns among similar pieces of music for digital signal processing.

Studies have investigated small-scale databases that contain few tracks because of the scarcity of generally available databases. Most cover song recognition algorithms are based on comparisons between chroma patterns or related features and use time-consuming similarity computation methods such as dynamic time warping (DTW). The release of the Million Song Dataset (MSD), which contains metadata and audio features for 1 000 000 songs, has spurred the investigation of large-scale music information retrieval techniques [7]. Finding cover songs from among 1 000 000 tracks is closer in scale to a commercial application than is identifying songs from a small database.

Previous cover song identification systems on large-scale datasets are based on indexing and fast retrieval. Bertin-Mahiuex and Ellis proposed using "chroma jump codes" [3]. They also proposed using 2D-Fourier transform to compress the beat-aligned chromagram and retrieve the song in fixed-length vector space [4]. Maurizio Omologo proposed using "chord profiles" [5] which is a type of high-level summarization of musical songs.

In recent years, machine learning has received more attention from the MIR community. Content-based MIR techniques typically feature two-stage architecture: first, features are extracted from music audio signals and transformed into a more meaningful representation. After being processed, these features are used as input to a classifier to perform the MIR task.

In this work, we used 2D Fourier transform magnitude [4],[6] to compress chroma patches of a song and pooled the output into a fixed-length vector. We then used machine learning approaches to reinforce the pattern of the vector. Because every vector represents a song in a semantic vector space, we can use nearest neighbor classification to efficiently search for songs similar to a music track.

The rest of this work is organized as follows: In Section II, we introduce the content-based music retrieval, including the cover song recognition task, chroma feature, and 2D Fourier transform. In Section III, we briefly introduce the machine learning approaches. In Section IV, we explain our music retrieval system and describe how we transformed a music signal into a semantic vector. In Section V, we detail the experiments and compare the performance of our system with related works. We present the conclusions in Section VI.

II. CONTENT-BASED MUSIC RETRIEVAL

MIR has many applications, such as genre classification, artist recognition, music recommendation, and automatic

978-1-5090-4571-6/16 $31.00 © 2016 IEEE
DOI 10.1109/ISM.2016.94

tagging. Most content-based MIR techniques use the spectrogram-based feature as input to perform tasks without needing handcrafted labels.

We assumed that cover songs are similar to the original songs. Our feature representation is the magnitude of the 2D Fourier transform of beat-aligned chroma patches. In the following section, we introduce the cover song recognition task, explain how our features are computed and describe their musical invariant properties.

A. Cover Song Recognition

Cover song recognition has been widely studied in recent years and at the Music Information Retrieval Evaluation eXchange (MIREX) since 2007. An overview of cover song recognition can be found in [1]. Most covers are similar to the original song in melody but different in key and timing structure.

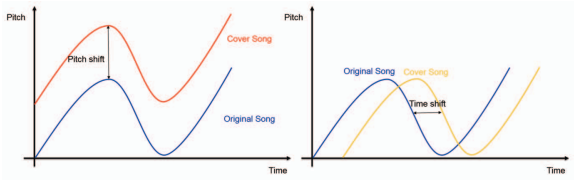

Fig. 1. Illustration of musical transposition

Fig. 1 illustrates the pitch and time shift of the original and cover. According to [1], cover song recognition algorithms generally comprise five steps (Fig. 2).

Fig. 2. Cover song recognition general block diagram.

B. Chroma feature

Chroma features are pitch-class profiles (PCP) that are derived from the spectrogram and provide a coarse approximation of the music score. A chromagram is similar to a constant-Q spectrogram except that pitch content is folded into a single octave of 12 discrete bins, each corresponding to a particular semitone. Fig. 3 shows the chromagram of an excerpt from *My Heart Will Go On*; the vertical axis represents the pitch and the horizontal axis represents the frame unit. We can also emphasize the energy of each frame by combining the chroma feature and loudness.

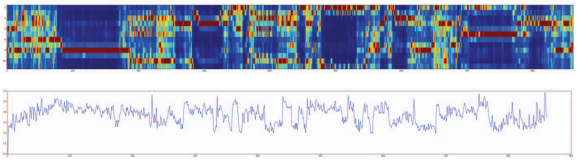

Fig. 3. Chromagram and loudness of a song

We used the Ellis-style beat tracker[1] [10] to obtain the segmentation of the song into beats. Averaging every segment chroma over beat times results in a beat-synchronous chroma feature. Empirically, we can raise the highest values relative to the lowest by using a power-law expansion.

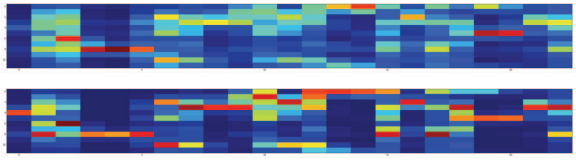

Fig. 4. Beat-aligned chroma features. (Top) a beat-synchronous chroma feature. (Bottom) the power-law expansion feature. 2D Fourier transform

Taking the 2D Fourier transform is a common technique in digital image processing, where it is useful for compacting energy. Marolt first used 2D Fourier transform in an MIR task [6]. Fourier transform has a transposition property, in which the shift along the time axis does not affect the magnitude (proof follows).

Assuming $y(t) = x(t - t_0)$, $\mathrm{Y}(\omega)$ and $\mathrm{X}(\omega)$ are the spectrum of $y(t)$ and $x(t)$, respectively:

$$\begin{aligned} \mathrm{Y}(\omega) &= \int_{-\infty}^{\infty} y(t) \cdot e^{-i\omega t}\, dt \\ &= \int_{-\infty}^{\infty} x(t - t0) \cdot e^{-i\omega t}\, dt \\ &= e^{-i\omega t0} \cdot \int_{-\infty}^{\infty} x(r) \cdot e^{-i\omega r}\, dr \\ &= e^{-i\omega t0} \cdot \mathrm{X}(\omega) \end{aligned} \tag{1}$$

The equations illustrate that time shift changes only the phase and does not influence the magnitude of the spectrum. Retaining the magnitude component and discarding the phase component provides invariance both to transposition in the pitch axes and beat axis. Fig. 5 furnishes an example of the transformation from chroma matrix to 2D Fourier magnitude coefficient (2D-FMC) matrix.

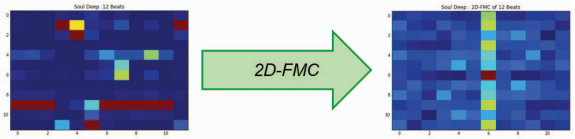

Fig. 5. Magnitude of 2D Fourier transform of chroma matrix.

III. MACHINE LEARNING APPROACHES

Machine learning provides broad application opportunities to different multimedia communities, such as computer vision and speech processing. Even some classical machine learning approaches can significantly improve the performance of MIR tasks.

A. Pooling

The representation that we obtained using 2D Fourier transform is not yet suitable as input to a classifier. We wanted to transform a matrix into a vector to render it compact and suitable for comparison using simple metrics such as Euclidean distance or cosine distance. We can then acquire a feature that is capable of performing scalable retrieval tasks. Although there are many kinds of pooling functions such as mean, median, and maximum, we observed that mean pooling obtained the most favorable result in our system.

[1] https://github.com/librosa/librosa

B. Feature Learning

After pooling, a music signal can be represented as a vector. Recent results in feature learning indicate that simple algorithms such as K-means can be very effective, sometimes surpassing more complicated approaches based on restricted Boltzmann machines or auto-encoders [8]. Some classical machine learning approaches such as principal component analysis (PCA) and linear discriminant analysis (LDA) can be used to reduce the dimensions of these vectors and reinforce the feature pattern. K-means algorithm can automatically group the vectors into k clusters. We can use a trained K-means model to obtain the bag-of-words representation. In the K-means transformed vector, each dimension represents the distance from the input vector to each centroid of cluster.

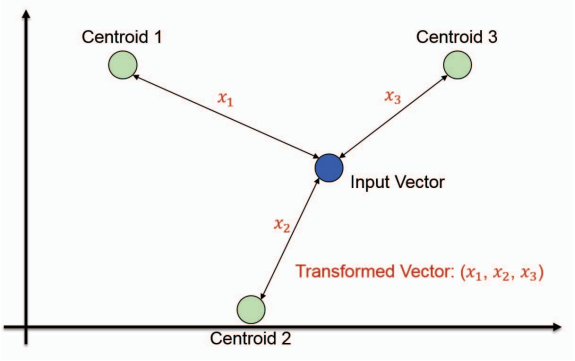

Fig. 6. Illustration of K-means transformation.

PCA is a type of dimensionality reduction method, which maintains the principal components according to the variance of each dimension. Vectors can be more compressed and discriminant after PCA. Both K-means and PCA are unsupervised learning, which means that the models can learn the pattern without any label. The other dimensionality reduction method is LDA, which is a supervised learning. LDA can learn the sematic relationship that minimizes the intraclass variance and maximizes the interclass discrimination through the ground truth labels.

Fig. 7 provides a visualization of musical vectors' distribution and reveals that the same color points are covers from ground truth. Closer points are more similar to each other.

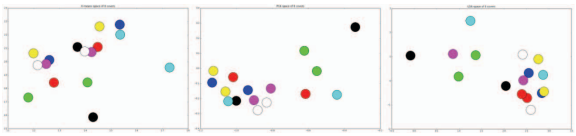

Fig. 7. Examples of transformed vectors' distribution in 2D-Euclidean space: From left, K-means transformation, PCA transformation, and LDA transformation.

C. Nearest Neighbor Classification

The nearest neighbor classification is one of the simplest classifiers in machine learning. The nearest neighbor algorithm has been successful in many classification and regression problems. The nearest neighbor algorithm classifies the points according to their vector distance from each other. The nearest points of a specific point are labeled the neighbors. The classifier retains the training data and outputs the top-k nearest neighbors as a result. Fig. 8 furnishes an example of a Euclidean space that has five points. We can calculate the top-3 neighbors of the white point as the white point itself, the green point, and the red point, in that order; their Euclidean distances are 0, 2.236, and 4.242, respectively.

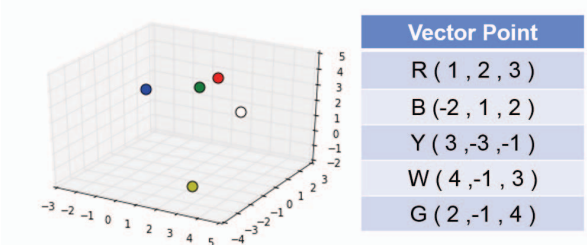

Fig. 8. An example of vector point space.

IV. PROPOSED SYSTEM

The block diagram of the proposed system is presented in Fig. 9. The following text provide step-by-step explanations of the proposed system, from audio signal to the final retrieval list.

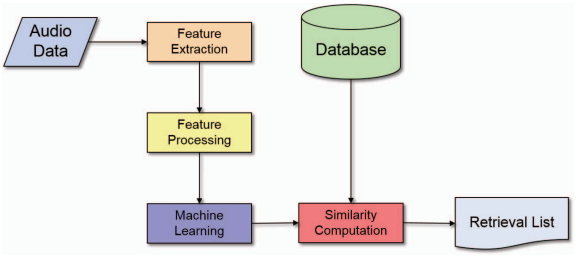

Fig. 9. Diagram of the proposed system.

A. Feature Extraction

The proposed system extracted the features from the audio signal by using tools provided by the librosa[1]. It obtained the chroma feature, loudness, and beat information for tracks to ensure the same representation for any input. The MSD had already provided common features.

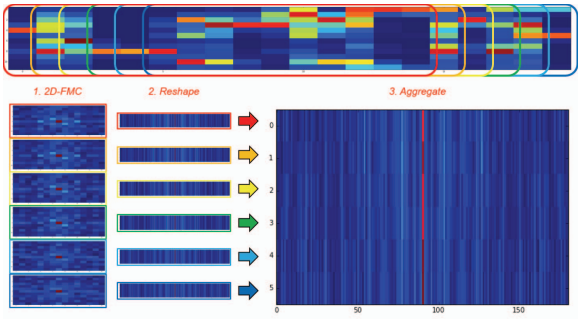

Fig. 10. Illustration of 2D-FMC Aggregation.

B. Feature Processing

To get more meaningful representation of a song, each feature underwent a process:

First, the system combined the features to obtain the beat-aligned chroma features as described in Section IIB. It then used a fixed-length window to obtain the 2D-FMC matrices convolutionally and reshaped these matrices to aggregate them. Finally, we could obtain the same length matrix for any different length song, whose size is the product of the window size and the number of semitones (12 in our system). This is a 2D-FMC aggregation matrix.

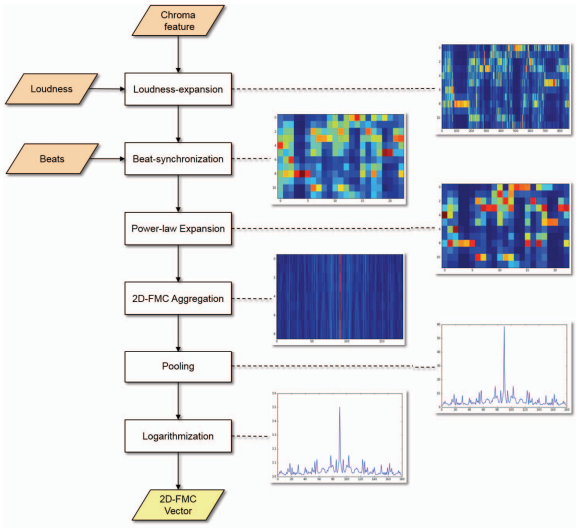

Fig. 11. Diagram of the feature processing.

We can use a pooling function, such as mean and median, to obtain the 2D-FMC vector from the aggregation matrix. After pooling, the matrix is transformed into a vector; the number of dimensions is the same as the length of the matrix. Empirically, taking the logarithm can enhance the resolution of most of the dimensions that improve the retrieval result.

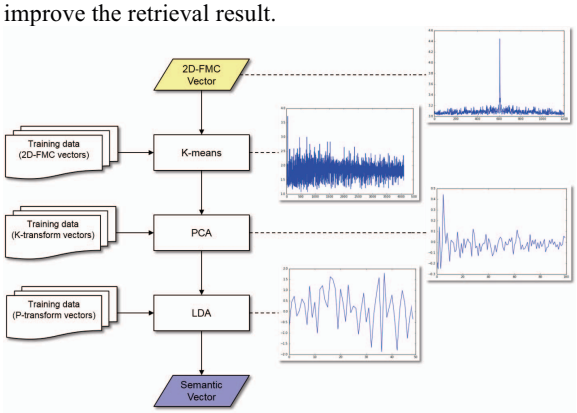

Fig. 12. Diagram of the machine learning combination model; this model combines K-means, PCA, and LDA in order, which provides the most favorable result for our experiment.

C. Machine Learning

Although the 2D-FMC vectors are powerful enough to perform the cover song recognition task, the performance can be further improved by reinforcing the vector patterns. We investigated different combinations of three machine learning models. We used the SecondHandSongs (SHS) training set as training data; it contains 12 960 tracks. We then used the trained model to transform 2D-FMC vectors into semantic vectors that have lower dimensions and a clearer pattern. We use scikit-learn[2] implementation [9].

D. Similarity Computation

After projecting all the songs into a sematic vector space, we used simple metrics to measure the similarity between songs. The Euclidean distance was first considered. Fig. 12 illustrates the Euclidean distance metric; when the distance is smaller, the similarity is greater.

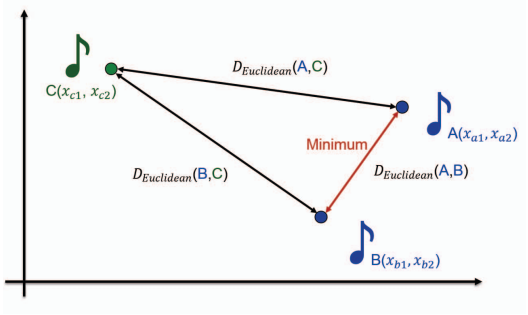

Fig. 13. Illustration of a Euclidean similarity measurement. Vectors A and B are separated by the shortest distance; therefore, they are the most similar songs when using the Euclidean distance metric.

We also explored other vector distance metrics, including Cosine distance and Manhattan distance. Empirically, the advantage of Cosine distance is that using an angle between two vectors to measure similarity is the more effective way to identify similar songs. Fig. 13 indicates that although some music vectors have the same Euclidean distance for the input, we can still find the closest candidate by the Cosine distance metric and obtain a more accurate result.

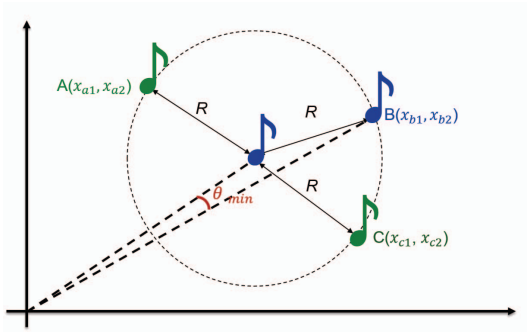

Fig. 14. Comparison of Euclidean and Cosine similarity.

V. EXPERIMENTS

To evaluate our retrieval system, we chose mean average precision (MAP) and average rank (AR) as our measures [3],[4]. MAP is computed as the mean of the average precision over a set of queries. The MAP reflects not only the accuracy but also the order of correct documents in a ranked list. AR is computed as the average position of relevant documents. For MAP, higher is preferable. For

AR, lower is preferable. To avoid misleading results caused by the overfitting of model training, we trained our models using the SHS training set (12 960 tracks) and reported the result on the SHS testing set (5236 tracks). Finally, we compared our system with related work [2],[3],[4],[5] on the full MSD[3] (1 000 000 tracks).

A. SecondHandSong dataset

The SHS dataset is a subset of the MSD, which is designed for the cover song recognition task. The SHS dataset consists of two subsets: the training set (12 960 tracks) and the testing set (5236 tracks). The SHS training set has 4128 cliques and the SHS testing set has 1726 cliques; "clique" here means groups of versions of a single underlying musical work.

B. Retrieval experiment

There are numerous parameters in our system, including the window sizes of 2D Fourier transform, the pooling function, and the coefficients of logarithm. We have also investigated various combinations of machine learning models. The dimensional number of 2D-FMC vector is 1200, of which the 2D Fourier window size is 100. All the parameters were chosen empirically. We explored the window size range 50–100. We obtained more favorable results using a mean pooling function. Table 1 provides the optimal results of different combinations.

Feature	MAP	AR
2D-FMC (1200)	0.130285	1054.442
K-means (4096)	0.051222	1383.014
PCA (100)	0.150482	**955.829**
LDA (100)	0.095285	1528.450
K+P (100)	0.050138	1383.818
K+L (100)	0.042946	1997.573
K+P+L (100)	**0.207817**	1037.962
P+K (100)	0.052065	1257.401
P+L (100)	0.164885	1251.320
P+K+L (100)	0.061158	1954.219

Table 1. Results for the SHS testing set (5236 tracks); K denotes K-means, P denotes PCA, L denotes LDA. The number in parentheses means the dimensional number of the transformed vector.

Distance	MAP	AR
Euclidean	0.207817	1037.962
Cosine	**0.242044**	**764.041**
Manhattan	0.193269	1112.511

Table 2. Results for the SHS testing set (5236 tracks); K denotes K-means, P denotes PCA, L denotes LDA.

Different similarity metrics were explored. Table 2 shows that simply replacing the Euclidean distance with Cosine distance can significantly improve the performance. The experiments revealed that the KPL (K-means, PCA, and LDA, in order) combination exhibits the optimum performance and that the Cosine similarity is the most useful method for identifying similar songs. K-means transform can represent any input vector by the Euclidean distance relationship with the cluster centroids in the SHS training set. Although the performance increased as the number centroids increased, the performance of K-meant the transformed vector remained less favorable than that of the original 2D-FMC vector. PCA is an appropriate choice for reducing the dimension, even though only the single PCA can enhance the MAP. The performance of LDA was not as ideal as were the results for the SHS training set because supervised learning was affected by the training data. Although LDA can maximize the interclass discrimination, too many dimensions can result in a less optimal performance. We can properly combine the advantages of each machine learning approach by first using the K-means to obtain the bag-of-words representation, then reducing the dimensions by PCA, and finally using LDA to maximize the discrimination. We can effect a significant improvement by using the combinational model. The combination of PCA and LDA can engender more favorable results than any single approach. The combinational order of each approach is crucial. We assumed that dimensional number parameters that perform optimally in each approach would also perform similarly in combinations. The dimensional number of the proposed KPL approach was enhanced from 1200 to 4096 by K-means transformation. The number was then reduced from 4096 to 100 by PCA transformation. Finally, we retained both 100 and 50 dimensions through LDA because each approach has its advantages and disadvantages. We compared our proposed system with related works [2],[3],[4],[5] on the MSD (1 000 000 tracks). We queried the SHS testing set (5236 tracks) and sought their covers on the MSD except itself (1 vs. 999 999), because the most similar song for any query is the query song itself for which the vector distances is 0. The results are presented in Table 3.

Method	MAP	AR
Random	~0.000014	500 000
Pitch Histogram [2]	0.00162	268 063
Jcodes [3]	0.00213	308 370
2D-FTM(50) [4]	0.02954	173 117
2D-FTM(200) [4]	0.01999	180 304
Chord Profiles [5]	0.03709	**114 951**
Proposed KPL(50)	0.05918	128 322
Proposed KPL(100)	**0.07062**	144 217

Table 3. Results for the MSD (1 000 000 tracks)

[2] http://scikit-learn.org/stable/
[3] http://labrosa.ee.columbia.edu/millionsong/

Fig. 15 indicates that our system exhibits superlative performance among the others and that its MAP can reach 0.07062.

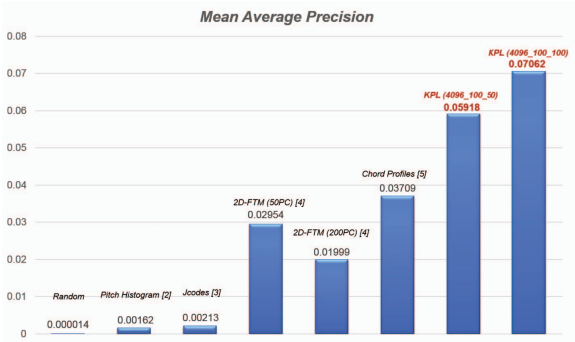

Fig. 15. MAP Results for the MSD.

VI. CONCLUSION

In this work, we proposed a large-scale cover song retrieval system. We used 2D Fourier transform to compress the music information features and the combinational machine learning model to reinforce the vectors pattern. We have experimented different combinations and found that a proper order of three classical machine learning approaches can outperform any single approach. In addition to applying the nearest neighbor algorithm to efficiently retrieve similar music, we also explored different similarity metrics and discovered that the Cosine similarity provides the optimal result in our experiment. Our results indicate that our system is a promising start for large-scale music retrieval tasks that use machine learning approaches. In the future, we can continue to investigate different machine learning models such as deep neural networks to enhance accuracy.

REFERENCES

[1] J. Serra, E. Gómez, and P. Herrera, "Audio Cover Song Identification And Similarity: Background, Approaches, Evaluation, And Beyond," Advances in *Music Information Retrieval*. Springer Berlin Heidelberg, 2010. 307-332.

[2] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch Histograms In Audio And Symbolic Music Information Retrieval," *Journal of New Music Research* 32.2 (2003): 143-152.

[3] T. Bertin-Mahieux and D.P.W. Ellis, "Large-Scale Cover Song Recognition Using Hashed Chroma Landmarks," In Proceedings of *WASPAA*, New Platz, NY, 2011.

[4] T. Bertin-Mahieux, and D.P.W. Ellis, "Large-Scale Cover Song Recognition Using the 2D Fourier Transform Magnitude," in *ISMIR*, Porto, Portugal, October 2012.

[5] M. Khadkevich, and M. Omologo, "Large-Scale Cover Song Identification Using Chord Profiles," in *ISMIR*, Curitiba, Brazil, November 2013.

[6] M. Marolt, "A Mid-Level Representation For Melody-Based Retrieval In Audio Collections," *IEEE Transactions on Multimedia* 10.8 (2008): 1617-1625.

[7] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *ISMIR*, Miami, FL, October 2011.

[8] S. Dieleman, and B. Schrauwen, "Multiscale Approaches To Music Audio Feature Learning," in *ISMIR*, Curitiba, Brazil, November 2013.

[9] F. Pedregosa, et al., "Scikit-Learn: Machine Learning In Python," *Journal of Machine Learning Research* 12.Oct (2011): 2825-2830.

[10] D.P.W. Ellis, "Beat Tracking By Dynamic Programming," *Journal of New Music Research* 36.1 (2007): 51-60.