# A Survey of Content-based Music Similarity

**Luyang LI**
University of Rochester
luyang.li@rochester.edu

**ABSTRACT**

The growth of digital music and consumable devices industry has resulted in the existence of hundreds of millions of music pieces. It has caused several problems like organizing the music library or recommending music to be extremely urging and hard, which all address the same question: how to find similar music?

This paper explores the methods that use computer algorithms to assess music similarity, where the degree of similarity is based solely on the musical contents. By extracting several kinds of features from low-level features like timbre to some higher-level features which have musical meanings, the similarity between songs is assessed by comparing the quantitative distance between features.

## 1. INTRODUCTION

As the word 'content-based' says, the general idea is about using some non-meta content features to determine the similarity between music. By making use of features from low-levels ones like zero-crossing rate, MFCC distance to some higher-level features like some meaningful music structures, the similarity between two certain pieces of music could be determined without knowing the metadata like genre, composer and etc.

In the existing commercial system, the similarity of music is mostly determined by User-content matrix (Collaborative Filtering, CF), meta-data of music tags (artist, genre and etc.) and some user-generated tags (last.fm). However, all of the three methods mentioned above have their shortcomings.

For CF method, it's the most widely used method and can be accurate in recommending the songs that is similar in music style or tastes. However, a serious problem is that the CF method is always narrowing the user's taste. For example, if the user consecutively gave positive feedback to three or more rock song, the recommendation system may endlessly recommend rock music despite the user might also have an interest in country music. By using the content-based music similarity, we would be able to find the similarities beyond the shared interest pattern, thus leading the user out of his usual taste pattern to explore new music.



For metadata and tag data, they can be used to help improve the CF result and also solve some problem like cold start. However, these data is often not accurate or too vague to represent anything useful. For example, many pieces of music may not even have any tags available and even if it does, we can't say that all music with the tag of rock is reliably similar.

Starting from often cited early paper [1], in which B. Logan and A. Saloman used MFCC as the extracted feature and vector distance as the similarity, the methods has been grown into a large and systematic study.

On one hand, different kinds of features were selected to do the job, including timbre features like MFCC, zero crossing rate, rhythm feature like tempo and melody feature like pitch and harmonic content.

On the other hand, instead of computing the simple feature vector distance, more advanced machine learning skills are also used in computing the content similarities, such as Gaussian Mixture Model, Support Vector Machine and etc.

Since the topic is addressing more and more attention, some competition were held to address the best algorithm and set up an experimental standard for evaluating the method. One notable contest is the MIREX contest [7] on Audio Music Similarity and Retrieval which was held annually since 2006, except for 2008.

One thing notable is that starting from a simple and general idea, the topic "content-based music similarity" has been extended to many sub-problems like cover song detection, query-by-singing and etc. However in this paper, only the general topic on studying the similarity of music will be discussed.

In this paper, a discussion of the definition of "music similarity" will be introduced in section 2. Then several kinds of features often used in the topic will be included in section 3. In section 4, some methods especially those earlier methods which involve non-model techniques will be covered and discussed. Moreover, more model-based approaches will be introduced in section 5. In section 6, some majorly used experimental setup and standards will be introduced algorithms will be compared. Finally, a conclusion will be drawn and future work will suggested in section 7.

## 2. MUSIC SIMILARITY

Although people can tell if two music audio are similar or not, the concept or even the mathematical definition of

music similarity is still vague and needed to be studied. To address this problem, many works has been done and researchers tried to define the ground truth data of "similar music" from two aspect.

### 2.1 Subjective Similarity

In some database, human evaluation was used to define the subjective similarity between music audio.

One good example is the database described in [2], a system call Evalutron 6000, whose data has been followed in the MIREX contest in audio music similarity and retrieval.

In the database, online users or volunteers are given with two random piece of music of 30 seconds and are asked to evaluate the similarity between two songs. The evaluation result is in form of both concrete classes from options including "similar", "somewhat similar" and "not similar" and a continuous indication on scale of 0-10. After the evaluation, cover-songs and songs by the same artist will be filtered out for the reason that evaluators maybe judge the result based on factors other than the music content.

One thing noticeable is that a subjective similarity matrix could also be drawn from a collaborative filtering system, which is often the similarity in commercial system. But this kind of result is barely seen in the experiments on this topic, largely because the connection between similarity based on user's interest and similarity based on the music content is not explicit yet.

### 2.2 Objective Similarity

Due to the unavailability of related work and under study of the concept of music similarity, in early days, in the research field the concept of music similarity is often defined as the similarity of a combination of contextual information including singer, genre, album and etc.

In [3] and [6], music with textual tags was used in the experiments. The similarity is based on a bi-value result that if the query song and retrieved song belong to the same genre. And within genre, the further similarity is decided by the album (for modern music) and instrumentation style like orchestra, string and etc. (for classic music).

In some more reliable subjective similarity definition like the definition used as part of MIREX in [8] [11] [12] [13], the subjective similarity is a combination of genre, artist, album and some manually pre-define tables. In these standards, a real-value similarity matrix will be calculated from the available meta-data.

## 3. FEATURE EXTRACTION

One important part of designing an effective algorithm is to choose the proper features used for computing the similarity and many features have been found or designed to address the problem.

### 3.1 Timbre

Timbre is often defined as the perceptual features that distinguish different sounds with the same loudness and pitch. The physical characteristics of sound that determine the timbre are often extracted from spectrum and envelope.

Some famous and common timbre features generally falls into following categories [14]:

Temporal features: features computed from the audio signal frame (zero-crossing rate, linear prediction coefficients, etc.).

Energy features: features referring to the energy content of the signal (Root Mean Square energy of the signal frame, energy of the noisy part of the power spectrum, etc.).

Spectral shape features: features describing the shape of the power spectrum of a signal frame: centroid, slope, roll-off frequency, variation, Mel-Frequency Cepstral Coefficients (MFCCs).

Perceptual features: features computed using a model of the human earring process (relative specific loudness, sharpness, spread).

One feature that needs special attention is MFCC introduced in [4]. After taking the logarithm of the amplitude spectrum based on Short Time Fourier Transform (STFT) for each frame, the frequency bins are grouped and smoothed according to Mel-frequency scaling. MFCC features are generated by de-correlating the Mel-spectral vectors using discrete cosine transform. Having been proved to be effective, the MFCC has been the most widely used feature in related field.

Since timbre features have been playing an important role, much study has been done on it. According to the study of Aucouturier and Pachet [5], timbre similarity has a grass ceiling, so we need to take other kinds of features into consideration

### 3.2 Rhythm

A precise definition or rhythm does not exist. Most authors refer to the idea of temporal regularity [6]. As a matter of fact, the perceived regularity is distinctive of rhythm and distinguishes it from non-rhythm. More generically, the word rhythm may be used to refer to all of the temporal aspects of a musical work.

One classic rhythm pack is describe in [6] and [15], which consist of several features extracted from the autocorrelation of a short window of music signal including:

- Fr1: Ratio of the power of the highest peak to the total sum.
- Fr2: Ratio of the power of the second-highest peak to the total sum.
- Fr3: Ratio of Fr1 and Fr2.
- Fr4: Period of the first peak in BPM.

– Fr5: Period of the second peak in BPM.

– Fr6: Total sum of the power for all frames in the window.

Other rhythm features include Fluctuation Patterns which measure periodicities of the loudness in various frequency bands, considering a number of psychoacoustic findings; Onset which indicate an sudden increase amplitude, often representing the starting point of an musical event or note.

### 3.3 Melody and Harmony

Harmony may be defined as the use and study of pitch simultaneity and chords, actual or implied, in music. On the contrary, melody is a succession of pitched events perceived as a single entity. Harmony is sometimes referred to as the vertical element of music with melody being the horizontal element.

Some famous and common rhythm features include pitch [13] and harmonic contents

### 3.4 Other Notable Features

Other than the features in the three categories mentioned above, some high-level or mid-level features have also been used in some research like the bass-line feature used in [6] and block set features used in [8]. Although these mid-level or high-level features are often computed from low-level features discussed above, these features are often considered separately for the reason that they are well-defined and compact.

## 4. THE NON-MODEL BASED APPROACHES

In early days, due to the unavailability of advanced machine learning study and some usable pre-label data set, the methods used to calculate the features tend to be unsupervised and simple.

In these methods, the algorithms often use a two-step strategy. First, a single or several different features were selected and represented in the form of vectors. Then the distance, either Euclidean Distance between vectors or Kullback-Leibler Divergence between vector distributions, were calculated as the similarity between songs.

The advantage of these algorithms is that they are computationally effective and simple if the features selected were not complex at the same time of having an acceptable performance.

The disadvantage is although the performance is acceptable in some cases, it's still not favorable enough and the performance seemed to be reaching the glass ceiling.

In the work done by B. Logan and A. Saloman [1], a simple similarity is calculated by making use of a single feature of MFCC. First the song will be divided into different frames and the MFCC component of which were computed. Then these frames were clustered using K-means clustering to for a signature of song. Finally the distance is decided by the Earth Mover Distance between the signatures of each song.

In the work done by Tao Li and Mitsunori Ogihara [3], a combination of features was selected, including MFCC, Spectral Centroid, Spectral Rolloff, Spectral Flux and especially Daubechies Wavelet Coefficients Histogram. Then the features were stored in a one-dimensional vector. Finally, the Euclidean distances between vectors were computed as the distance between songs.

In the work done by Tetsuro Kitahara, Yusuke Tsuchihashi, and Haruhiro Katayose [6], the whole algorithm generally followed the similar progress of [3], except that other than timbre and rhythm features, a bass-line feature is also used in the method. To extract the bass-line feature, they used 21 features falls into two categories: pitch variability and pitch motion.

In the work done by George Tzanetakis[13], a large combination of feature were selected. These features covers a whole range of available features from timbre features including MFCC, chroma features like pitch to rhythmic features including onset energy and beat histogram. Then after normalizing all the features, the similarity was computed from the Euclidean distance between feature vectors. One thing noticeable is that it's part of an open-source MIR toolbox which has been used in commercial organizations.

In the work done by Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees and Gerhard Widmer [11], the features selected were main consist of rhythmic features and some timbre features were also added to improve the performance. For timbre and rhythm features, distance between feature vectors were calculated separately and added together after normalization. The basic idea of this method is to simulate the music similarity in a rhythm similarity way and it's one of the best performed unsupervised method by far.

## 5. THE MODEL-BASED APPROCHES

In later work, some more advanced methods with some machine learning techniques have been put into used.

In these methods, the specific method varies but generally follows the following steps: First often a large combination of features was selected and computed for each audio object. Then some pre-labeled audio objects, with labels including genres, artists or sometimes even similarity, were randomly selected as the training dataset to train machine learning models. In the training segment, the models would "figure out" the right combination of weights for every feature selected. Then un-labeled audio objects were used as a query and the similarity was given.

The advantage of these methods is that the performance is relatively better than the unsupervised one and could be further improved by improved dataset. But one drawback is that the algorithms are often more complicated and thus may not be of great computational cost.

The most commonly-used model is Gaussian Mixture Model [9] [12] and Support Vector Machine [8].

For Gaussian Mixture Model, it estimates a probability density as the weighted sum of M simpler Gaussian densities, called components or states of the mixture.

For Support Vector Machine (SVM), the basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.[16]

In the work done by J J Aucouturier and F Pachet [9], first only MFCC was selected and computed as the feature descriptor. Then a trained Gaussian Mixture Model is used to model the distribution of each song's MFCC. Then the similarity was computed from the probability that the MFCC of one song could be generated from the model of another song. The author also made attempt to reduce the computational cost of the method.

In the work done by Christophe Charbuillet, Damien Tardieu and Geoffroy Peeters [12], a new combinations of features including both timbre and rhythm features were selected. Then a modified version of GMM called Universal Background Model is used to apply the similar progress. One thing noticeable is that it is one of the best performed model-based algorithm in the MIREX contest.

In work done by Klaus Seyerlehner, Markus Schedl, Peter Knees and Reinhard Sonnleitner [8], a whole new set of features were selected. In this paper, a combination of block-based mid-level features instead of low-level features was used as the descriptor. Then tag prediction algorithm was applied by using Support Vector Machine. Finally the similarity is decided by the combination of a feature vector distance and tag distance.

## 6. EXPERIEMNTAL SETUP AND RESULTS

### 6.1 Early Experimental Standard

In early days, a typical dataset often consists of audio clips with meta-data of a number varies from 300 [3][6] to 8000 [1]. These audios cover different singers, albums and music genres. Sometimes some researchers use audio samples crawled from online music retailers like Amazon.

In these experiments, a common experimental strategy is as following:

1. Several songs are selected as the query songs.
2. For each query song, a list of top-k most similar song is returned as the result.
3. Check the album or music genre of songs in the returned list.

For the result, the performance is often measured by the rate that there at least exist one song in the top-k list that belongs to the same genre or album as the query song, which are often called accuracy.

In these experiments, the music similarity problem is often treated as "genre classification". One reason of this is that subjective similarity data is often unavailable in early days until the MIREX tried to solve this problem. Another reason is that, as mentioned before, at the beginning some problem that are considered as separate or at least sub-problem now were considered somewhat the same as the topic of content-based music similarity and retrieval.

In the paper [1], there are an average of 3.4/6.5 songs in the top 5/10 similar song list that belongs to the same genre as the query song.

In the paper [3], the accuracy for top three matches is 63.3% and 90.0% for top nine matches.

In the paper [6], the accuracy for jazz/classical music has been improved to as high as 92% although the accuracy for the genre pop is still not very satisfying for the reason that pop songs often covers a lot of music styles

### 6.2 The MIREX Contest

To address the problem mentioned above, the MIREX (stands for Music Information Retrieval Evaluation eXchange) contest set the task of "Content-based Music Similarity and Retrieval" as an annually contest since 2006 and introduced new standard and dataset for the research in this area.

In MIREX, a dataset of 7000 songs are chosen from organizers of "uspop", "uscrap" and "american" "classical" and "sundry", which covers a broad range of western music from classical genre like baroque to modern genre like metal and rock'n'roll. For each participant, the algorithm should return a 7000x7000 distance matrix where the distance value indicates the similarity level.

Then, 50 songs were randomly selected from the 10 genre groups (5 per genre) as queries and the first 5 most highly ranked songs out of the 7000 were extracted for each query. Then, for each query, the returned results (candidates) from all participants were grouped and were evaluated in two ways: both subjectively and objectively.

For subjective performance, the similarity between query and candidates were evaluated by listeners. In details, each individual query/candidate set was evaluated by a single grader. For each query/candidate pair, graders provided two scores:

One is called BROAD score with three values: 0 for Not Similar, 1 for Somewhat Similar and 2 for Very Similar. Another one is called FINE score, it is in scale of 0 to 100 and larger number indicates more similarity.

For objective performance, several statistical numbers will be calculated from the generated distance matrix including average percentage of genre matches in top 5 results (precision), average of percentage of available genre, artist and album matches in top-k results (recall), percentage of files that were never in the top-k results and percentage of files that were always in the top-k results.

By far, the work introduced in [12] has the best performance in the MIREX contest with a FINE score of

58.586 and BROAD score of 1.296, which could be considered as being able to tell subjectively "somewhat similar" songs.

The work introduced in [11] is a leading non-model based algorithm in MIREX contest with a FINE score of 55.080 and BROAD score of 1.228, which is a little bit lower than the best model-based method.

Another leading algorithm is introduced in [8] with a FINE score of 58.128 and BROAD score of 1.292, which almost matches the performance of [12]

One thing worth noticing is that the work introduced in [8], [11] and [12] are done in 2010, 2010 and 2009, respectively. For the following years, the original authors tried to improve their methods, only to receive a less favorable performance.

As the representative of industrialized methods, the algorithm proposed in [13] received a FINE score of 45.842 and a BROAD score of 0.940.

### 6.3 Other Experimental Setup and Database

Other than the two kinds of database introduced above, some other works has also been done to help to evaluate the algorithms and to provide the ground truth.

In [10], a project call CAL500 is introduced, which involve a massively manually tagged song dataset. In CAL, 500 "western popular" songs from 500 unique artists are selected. Then a vocabulary of 174 "musically relevant" semantic keywords is chosen as the vocabulary set. Finally volunteers were asked to label the songs in the database using the words from the vocabulary set. Although the dataset is primarily targeted at serving as a ground truth of music annotation system, the available tags could also be used to compute the similarities between songs.

There are also some online music data service organizations or companies that allow researchers to get the similarity matrix between songs in this database. Examples of this include Last.fm and Music Brainz, which allows users to download a similarity matrix which is computed based on shared tags. Another example is the service provided by Echonest, which has an online API that allows user to query with certain songs and get back a list of similar songs.

## 7. CONCLUSIONS AND FUTUREWORK

Several conclusions could be drawn from the work that has been done, including:

Some general and more objective tasks like genre classification have also been able to be achieved with desirable performance.

Although the result of the work in this field is still barely seen in commercial system, the performance of algorithms in some aspect has already been acceptable for commercial use. Also some open-source projects has been trying apply the algorithm to address practical problems.

However, there is still a long way to go in this field:

As for now, the features used in the algorithms are still largely some low-level features which involve more spectral description rather than music meaning or explanation. More music meaningful structure will be needed to be taken into consideration.

It means that in some way the scientist is trying to "test" the content similarity pattern using massive experiments and machine learning skills instead of trying to solve the problem from a solid theoretic background, which result in large amount of meaningless work and a more and more obvious glass ceiling of performance of current algorithm.

Another problem related to the problem discussed above is that the effectiveness of each feature is needed to be tested. Nowadays, researchers tend to use a combination of massive kinds of features without knowing the inner connections and real effect. Nowadays researchers are trying to avoid the problem by massive experimental testing and machine learning algorithms or just ignore it.

Last but not least, the bridge connecting content-based similarity and user-interest-based similarity should be explored. Although nowadays subjective similarities has been put into consideration, the content-similarity is still very far from the result got from some user-based algorithm like collaborative filtering, which is often more effective in music recommendation. For the reason that one important potential usage for content-based music similarity research is to help to improve the performance of current recommendation system, more attention still need to be addressed on exploring the connection between content similarity and shared interest.

## 8. REFERENCES


[1] B. Logan and A. Saloman, "A music similarity function based on signal analysis" in *Proc. IEEE Int. Conf. Multimedia and Expo*, Tokyo, Japan, 2001

[2] Online Protocol, "Music Similarity Grading System: Collecting Ground-Truth Similarity Information for Music Information Retrieval Evaluations" in MIREX 2006

[3] Tao Li and Mitsunori Ogihara: "Content-based Music Similarity Search and Emotion Detection", *Acoustics, Speech, and Signal Processing*, 2004.

[4] S.B. Davis, and P. Mermelstein: "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366

[5] Aucouturier and Pachet: "Timbre Similarity: How high is the sky?" JNRSAS 2004

[6] Tetsuro Kitahara, Yusuke Tsuchihashi, and Haruhiro

Katayose, “Music Genre Classification and Similarity Calculation Using Bass-line Features”, *Tenth IEEE International Symposium on Multimedia*

[7] Downie, J. Stephen, Andreas F. Ehmann, Mert Bay and M. Cameron Jones. “The Music Information Retrieval Evaluation eXchange: Some Observations and Insights” *Advances in Music Information Retrieval* Vol. 274, pp. 93-115

[8] Klaus Seyerlehner, Markus Schedl, Peter Knees and Reinhard Sonnleitner, “A Refined Block-level Feature Set For Classification, Similarity And Tag Prediction”, MIREX 2013

[9] J J Aucouturier and F Pachet “Finding songs that sound the same” In *Proc. of IEEE Benelux Workshop on Model based Processing and Coding of Audio*, 2002, pp. 1–8

[10] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet “Semantic Annotation and Retrieval of Music and Sound Effects” *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 16, no. 2, Feb. 2008

[11] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees and Gerhard Widmer, “On Rhythm And General Music Similarity”, ISMIR 2009

[12] Christophe Charbuillet, Damien Tardieu and Geoffroy Peeters “GMM Supervector for Content Based Music Similarity” *In Proc. Of 14th Int. Conference on Digital Audio Effects*

[13] George Tzanetakis: “MARSYAS Submissions To MIREX 2012”, MIREX 2012

[14] N. Scaringella, G. Zoia and D. Mlynek: “Automatic Genre Classification of Music Content” in *IEEE Signal Processing Magazion Vol 23*

[15] G. Tzanetakis and P. Cook: “Musicla Genre Classification of Audio Signals” in *IEEE Trans. Speech Audio Process*, 10(5):292-302, 2002

[16] Webpage: http://en.wikipedia.org/wiki/Support_vector_machine