Content-based retrieval of music using mel frequency cepstral coefficient (MFCC)

Xin Luo*, Xuezheng Liu, Ran Tao, Youqun Shi

School of Computer Science and Technology, Donghua University, Songjiang Districe, Shanghai, 201620, China

Received 1 March 2014, www.cmnt.lv

Abstract

In the last few years, with the growing of multimedia in Internet, MP3 music become one of the most popular types. Some of the MP3 music collections available are approaching the scale of million tracks and this has posed a major challenge for searching, retrieving, and organizing music content. In this paper, we proposed a method to retrieve the MP3 lossy compression format music by using MFCC features. The Kullback-Leibler Divergence and Earth Mover's Distance (EMD) are used to compute music similarity. Experiments show that the retrieving probability of our design can achieve high recall values of 95% out of a total of 1951 tracks in the database.

Keywords: content-based music retrieval, MP3, MFCC, kullback-leibler divergence, EMD

1 Introduction

In the last few years, with the growing of multimedia in Internet, MP3 music become one of the most popular types. Some of the MP3 music collections available are approaching the scale of million tracks. Therefore, it's become increasingly difficult when people classify and manage these massive musical date only by his own hand. We need to query the music information quickly, accurately and efficiently, the traditional approach that query by the text has been far from satisfying the user's search request for such resources. Besides the traditional way, which is using keyword-based mode to retrieve the song information to obtain the desired song, people also want to query it just by the characteristics of the music itself. The CBMR (Content-based Music Retrieval) is an important technique in recent years in handling massive multimedia data in a network environment, together with image retrieval and video retrieval it is today's hotspot in content-based retrieval research field [1].

The work of CBRM research began in the 1990s, the initial research is mostly about the humming retrieval system. Ghiaset [2] carried out pioneering research in Single-track MIDI music humming retrieval system, they use time-domain autocorrelation algorithm to extract pitch information, then by using a fast approximate string matching algorithm they implemented monophonic music retrieval. Tomonariet [3] using span and pitch information as search clues, adopt both dynamic threshold determination and coarse-to-fine matching for matching. Kosugi [4] proposed using pitch direction and distribution at the same time to improve the system performance. They developed a Sound Compass system, it contains 10086 audios, and people must humming coordinate with musical beats while using this system. Seungminet [5, 6] improved the pitch-tracking algorithm, they added the fundamental frequency index function to the traditional tracking algorithm, and by using genetic algorithm-based relevance feedback schemes they signify-cantly improved the retrieval accuracy.

In recent years the audio fingerprint-based fast retrieval technique, which is using binary number to represent the audio features, has becoming the key technology of audio retrieval engine [7-9]. Audio fingerprint extraction first proposed by Philps Institute in Holland, they proposed a global information based audio fingerprints extraction method. This method segment the audio spectrum into many frames, each frame represented by a 0 or 1, thus the whole audio spectrum can be expressed by a binary numbers sequence, the advantage of this method is the whole audio spectrum can be recorded. Another way is proposed by Shazam Company, this UK company try to find some feature points from the audio spectrum, they usually are peak points, and make them be Peak-Pairs, then use the sequence of the Peak-Pairs as the audio fingerprint of this frame. This method centralize the effective messages and got a good noise immunity.

Query by Humming allows the users to find a song by humming part of the tune, use the melody and rhythm to retrieval, make it convenient for the user to find a song compare to the traditional way which use the song title, artist and other text information to retrieval. However, the key part of this way is humming, this make it little meaningful to the user. Audio fingerprints is a better way to retrieval, classify and sorting music for users, but its shortcomings are the feature information is not typical and the noise immunity is also not good enough. Since the general users prefer MP3 format and MP3 as a lossy compression format is designed to significantly reduce the data size of an audio, when we use the audio fingerprint to retrieval a MP3 music it's hard to achieve high accuracy, but for most user's listening experience the reproduction acoustics and the original sound almost in the same level.

This passage proposed a MP3 retrieval system which is using MFCC (Mel-frequency Cepstral Coefficients) to characterization the audio signal. MFCC is based on the human

***Corresponding author** e-mail: xluo@dhu.edu.cn

auditory characteristics, it's a non-liner relationship with the Hz frequency. Through the 1951 Chinese and other language songs retrieval experiment we proved the effectiveness of the proposed method.

2 Proposed audio retrieval

Our proposed way is use MFCC as the audio features. MFCC is the determinants of the spectrum modeling, therefore it can reflects the tone of an audio for a certain degree. We extract MFCC vector sets from each audio and classify them as cluster sets, then calculate the distance between each cluster, use the nearest result as the similar retrieval result.

2.1 RETRIEVAL SYSTEM FRAMEWORK

In this section we review the audio retrieval system flow. The search processes, as shown in Figure 1, can be summarized in the following phases: first, extract MFCC features from all songs in MP3 music database and establish a MFCC feature database. The length of the MFCC feature is the 30 seconds in the middle of the front half of the corresponding song. Next, compress the extracted MFCC features by LBG Design Algorithm, then use k-means cluster them. When to retrieve a song, first extract the MFCC feature MF from query music, and then calculate the similarity distance between MF and the MFCC feature database using KL divergence function or EMD distance function, finally sort the distance in ascending order, output the top 100 songs as final retrieval results.

Details about the feature extraction and distance calculation are given in Figure 1.

2.2 MFCC FEATURE EXTRACTION

Mel frequency cepstrum coefficient just like the cepstrum is the feature, which is used to represent the channel characteristics. Cepstral analysis is mainly to get the audio spectral envelope by Fourier transformation, but, the human auditory perception experiments showed that human auditory perception only focus on certain areas, rather than the whole spectral envelope. The human auditory system is a special non-linear system, MFCC consider the characteristics of human hearing, first, it mapped a linear spectrum to Mel nonlinear spectrum based on auditory perception, and then convert it to cepstrum. Logan et al [10] through music modeling verified the effectiveness of MFCC feature in speech and music field.

The MFCC extraction algorithm performs as follow:

Step 1: Pre-emphasis the voice signal with a pre-emphasis filter, magnify the high frequency part. By doing so, make the channel characteristics more clearly. The pre-emphasis filter is determined as:

$$y(n) = x(n) - px(n-1), \tag{1}$$

$x(n)$ is the waveform of the voice, p is the pre-emphasis coefficient.

Step 2: To reduce the edge effect, apply Hamming window to each frame whose waveform have already attached pre-emphasis filter, calculate the amplitude of the vector by FFT(Fast Fourier Transform).

Step 3: Use Mel Filter Bank compress the Amplitude Spectrum. The filter bank consists of 33 triangular shaped band-pass filters. The focus here is to generate Mel-Frequency. The dimension of Mel is the horizontal axis which is used to reflect the human auditory characteristics, its unit is mel. The lower the frequency, the narrower the interval, the opposite is also true, the higher, the wider. The human ear in the subtle low frequency sound could feel the pitch different, but in the high frequency sound, its goes harder.

Step 4: Transform the compressed values by DCT to remove the correlation between the signals in different dimension, map the signal into low dimensional space.

Step 5: Using the low dimensional composition of the obtained cepstrum as MFCC feature values.

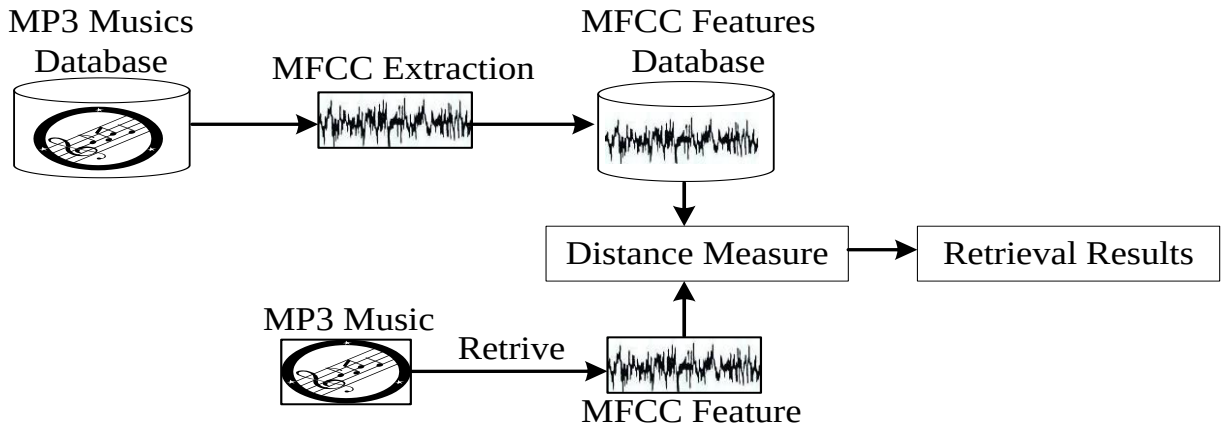

FIGURE 1 Flow of proposed method

This way, a music can be described by a series of spectrum vectors, and each vector is the MFCC feature vector of each frame.

The block diagrams for calculating MFCCs is given as Figure 2.

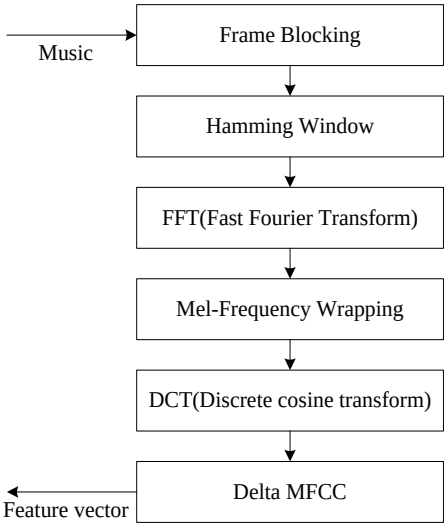

FIGURE 2 MFCC feature extraction processing

2.3 METHOD FOR SIMILARITY DISTANCE CALCULATION

In pure feature vector, the distance between each feature values can be expressed by Euclidean distance. But in audio retrieval, the feature value is not the vector but the normal distribution, that's why we cannot use a simple Euclidean distance.

2.3.1 KL Divergence

KL Divergence, a distance measure index of normal distribution, is the abbreviation of Kullback-Leibler Divergence, also known as Relative Entropy. It measures the difference between two probability distributions in the same sample space. According to the KLD, the definition of Kullback-Leibler Divergence between Normal distribution *N*1 and *N*2 (respectively, the mean vector, variance – covariance matrix) is:

$$D_{KL}(N_1 \| N_2) = \frac{1}{2}\left(tr(\sum_2^{-1}\sum_1) + (\mu_2 - \mu_1)^T \sum_2^{-1}(\mu_2 - \mu_1) - \ln\left(\frac{|\sum_1|}{|\sum_2|}\right) - k \right), \quad (2)$$

where *k* is the dimension of the mean vector. Usually the Kullback-Leibler Divergence is not symmetry, the value of $D_{KL}(N_1 \| N_2)$ and $D_{KL}(N_2 \| N_1)$ are not the same. However, this will not achieve our goals, so we use the Kullback-Leibler Divergence in symmetry case, that is:

$$0.5(D_{KL}(N_1 \| N_2) + D_{KL}(N_2 \| N_1)), \quad (3)$$

in this way we can ensure its symmetry.

2.3.2 The Earth Mover's Distance (EMD)

In computer science, the Earth Mover's Distance (EMD) is a measure of the distance between two multi-dimensional distributions in some feature space where a distance measure between single features. These distributions can be summarized with clustering algorithms, which reduce the feature space in a fixed number of bins. Each cluster c_j is associated with a weight w_j that indicates the size of the cluster (e.g. the occurrences of features in each cluster). The EMD describes the cost that must be paid to transform one distribution (considered as a mass of earth spread in space) into the other (considered as a collection of holes in the same space). The EMD measures the least amount of work needed to fill the holes with earth, considering that a unit of work corresponds to transporting a unit of earth by a unit of distance (ground distance). The EMD evaluation is based on the solution of the transportation problem, which consists in finding the least expensive flow from one distribution to another according to some constraints.

In a more formal way we can express the transportation problem as follows[11].

Let $P = \{(p_1, w_{p1}), ..., (p_m, w_{pm})\}$ be the first distribution with m clusters, where p_i is the cluster representative and w_{ij} is the weight of the cluster. $Q = \{(q_1, w_{q1}), ..., (q_n, w_{qn})\}$ the second distribution with n clusters; and $D = [d_{ij}]$ the ground distance matrix where d_{ij} is the ground distance between clusters p_i and q_j.

The flow $F = [f_{ij}]$ that minimizes the overall cost is computed by Eq. (4) where f_{ij} is the flow between p_i and q_j.

$$WORK(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}, \quad (4)$$

and it is subject to the following constraints:

$$f_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n, \quad (5)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{pi}, 1 \leq i \leq m, \quad (6)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{qj}, 1 \leq j \leq n, \quad (7)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}\right). \quad (8)$$

The first constraint indicates that the items can be moved only from *P* to *Q* and not vice versa. The next two constraints are related to the amount of mass which can be sent from the elements in *P* (it must not exceed the weight values) and to the amount which can be received by elements in *Q* (again limited by the weights). The last constraint forces to move the maximum amount of mass as possible. After solving the transportation problem and computing the total flow *F*, the EMD is defined as the work normalized by the total flow:

$$EMD(P,Q)=\frac{WORK(P,Q)}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}}. \quad (9)$$

The EMD is a robust method to compare multidimensional distributions of features. It is a true metric if the ground distance is metric and if the total weights of the two signatures are equals

3 Experiment and Evaluation

3.1 CONDITION FOR MFCC FEATURE VALUES EXTRATION

Extraction feature values from the entire song will lead the data amount being too large, to reduce calculation, Pampalket [12] proposed choosing the centred 30 seconds of each song as a proposing target. What's more, according to the characteristics of the song, usually the first half of a song can reflect its characteristic most, and for a considerable part of the song, the second half of them is usually repeating the first half. For this reason, in our experiment we use the MFCC feature extracted from the middle 30 seconds in the first half of a song as its processing target.

Table 1 shows the 30 seconds MFCC feature vectors of song A, sampling frequency 16Hz, shift width 160 samples.

One row in Table 1 represent 20 dimensional vectors of a frame, and a 30 seconds audio include 3000 frames, that's 6000 dimensional vectors in total, the calculation can be pretty big. To reduce this, we use LBG Design Algorithm which proposed by Lindeet [13] to compress it. After clustered 6000 dimensional vectors by this algorithm, we classify the similar vectors to the corresponding class, and assume all classes are normal distribution to calculate its mean vector and covariance matrix, using the result as the new feature value, therefore the feature value of audio *A* can be described as:

$$A=\left\{\left(\mu_{A1},\Sigma_{A1},w_{A1}\right),...,\left(\mu_{Am},\Sigma_{Am},w_{Am}\right)\right\}, \quad (10)$$

where *m* is the number of classes.

Using *k*-means as cluster algorithm. By clustering, we can get the reduced-dimensional vector:

$$A_{MFCC}=CA_i\sigma_i, \quad (11)$$

A_i is the mean vector of each class, σ_i is the covariance matrix, *C* is the number of classes. In this experiment *C* taken 16, 32 and 64 respectively. By using this equation, when the feature vector of audio A using 16 as the number of classes, a 30 seconds MFCC feature vector can be compressed to 6720 from 60000 dimensions.

3.2 DATASETS

To verify retrieval results, our experiments collected 1951 MP3 audios and songs in Chinese and other languages. We manually classified these audio files by its types and kept them in different folders, details are shown in Table 2.

3.3 RESULT AND EVALUATION

Our experiment evaluate the retrieval system from two aspects.

Let us do a brief review of this experiment at first. First choose 10 audios from each class, there are 6 classes, so its 60 audios in total, use these audios for querying and calculate the effectiveness of the retrieval system. While retrieving, extract the MFCC feature MF from the middle 30 seconds in the first half of the query music, calculate the EMD distance between MF and the MFCC feature database, outputs the top 100 songs as final results.

TABLE 1 MFCC Features of Song A

–5.36	–0.20	–0.92	...	–1.40	–1.78	1.40	–0.66	22.44
–1.17	–2.34	5.57	...	3.82	2.54	–0.87	1.42	22.75
–0.77	–5.59	5.56	...	4.30	4.36	–2.55	1.47	22.78
0.17	–6.48	7.63	...	3.76	5.45	–3.11	4.00	22.57
...	...	...	...	...	...	...	...	...

TABLE 2 Mp3 music datasets of experiment

Folder	Audio amount	Size
Encouragement	161	659.4MB
Fervor	334	1.4GB
Gladness	341	1.3GB
Quiet	301	1.2GB
Romantic	332	1.3GB
Sentimental	482	2.1GB
Total	1951	7.96GB

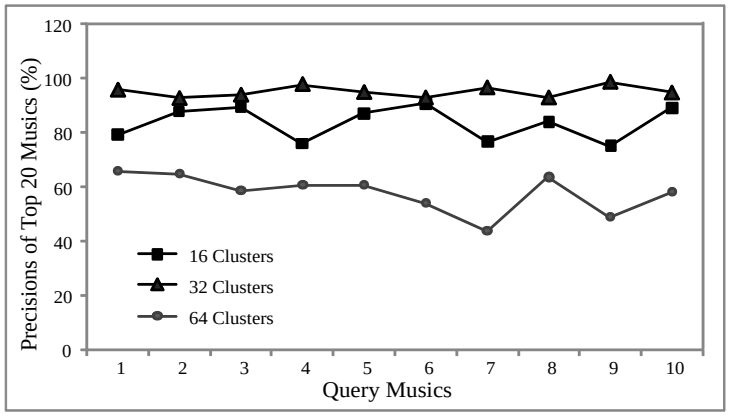

FIGURE3 The retrieval accuracy in using datasets with different classes cases.

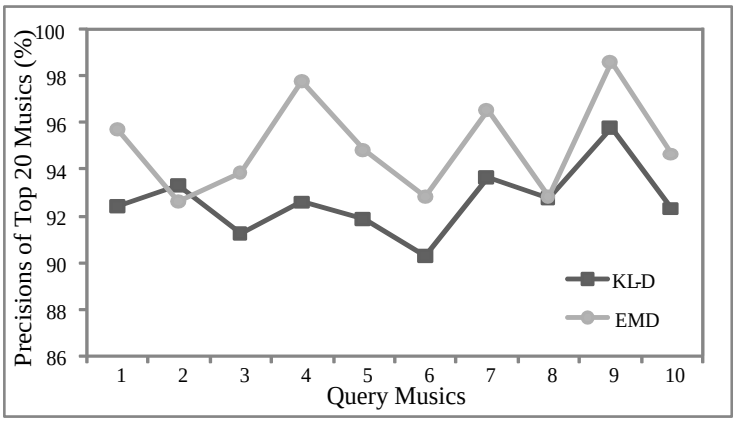

FIGURE 4 The retrieval accuracy by using different distance calculation method

We chose 10 most representative music to do this experiments, each retrieved 3 MFCC datasets (including 16, 32 or 64 classes) respectively. According to the retrieval results, select top 20s to compare with the corresponding class, calculate the precision to compare their accuracy. Our retrieval results show that in a MFCC datasets have 16 classes the accuracy is 83.28%, the accuracy is 95.05% in a 32 classes datasets and 57.85% in a 64 classes datasets. Therefore, the MFCC datasets which have 32 classes can get a better retrieval precision, the larger the number of classes, the lower the accuracy of query and the longer the calculation time. The results are shown in Figure 3.

And after that, for these 10 query music, we use KL divergence and EMD distance respectively to retrieve the MFCC datasets in 32 classes. Results shows as Figure 4.

From the experimental results, we can see that EMD is better than KL divergence under the same conditions.

4 Conclusion and prospects

This passage proposed an audio retrieval system, which is based on MP3 audio compression format, the audio feature values use MFCC and the retrieval distance use KL-D and EMD respectively. The proposed method are mostly focused on the first half of an audio, because the latter part usually repeat the previous section. Therefore we extract MFCC features from the middle 30 seconds of the first half to construct the feature vector datasets, and use LBG design algorithm compress the extracted features, then reduce dimension by using cluster algorithm.

In order to test and verify the effectiveness of our experiments, we collected 1951 different types MP3 music and songs in all kinds of languages. We tested MFCC feature datasets in 16, 32 and 64 classes. Experiment results show that when the class number is 32 the retrieval result is the best. When the number of class goes to 64, the retrieval amount becoming bigger, time consuming more, even the accuracy goes worse.

The proposed method only applies to MFCC feature, and because MFCC is based on human auditory characteristics, it is worked not very well on the low frequency area, which is sensitive to human ear. Our future work will focus on how to extract audio features more efficient. In addition, because the music features are usually got a large amount, in the future we will do more works on how to enhance the retrieval speed and reduce the calculation.

Finally, this experiment using a small MP3 music database, the future we will do further experiments in high-speed retrieval effectiveness aspect on the use of a large-scale database.

Acknowledgements

This research was partially supported by “the Fundamental Research Funds for the Central Universities (No. 13D11205)”.

Reference

[1] Casey M A, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M 2008 *Proceedings of the IEEE* **96**(4) 668-96
[2] Ghias A, Logan J, Chamberlin D, Smith B C 1995 Query by humming: musical information retrieval in an audio database *Proceedings of the third ACM international conference on Multimedia* ACM 231-6
[3] Sonoda T, Goto M, Muraoka Y 2002 A WWW-based melody retrieval system *Electronics and Communications in Haman (Part II: Electronics)* **85**(9) 63-74
[4] Kosugi N, Nishihara Y, Sakata T, Yamamuro M, Kushima K 2000 A practical query-by-humming system for a large music database *Proceedings of the eighth ACM international conference on Multimedia* ACM 333-42
[5] Rho S, Hwang E 2006 FMF: Query adaptive melody retrieval system *Journal of Systems and Software* **79**(1) 43-56
[6] Rho S, Han B-j, Hwang E, Kim M 2008 MUSEMBLE: A novel music retrieval system with automatic voice query transcription and reformulation *Journal of Systems and Software* **81**(7) 1065-80
[7] Haitsma J, Kalker T 2002 A highly robust audio fingerprinting system *ISMIR 2002* 107-15
[8] Bellettini C, Mazzini G 2010 A Framework for Robust Audio Fingerprinting Journal of Communications **5**(5) 409-24
[9] Xiao Q, Saito N, Matsumoto K, Luo X, Yokota Y, Kita K 2013 Index Compression for Audio Fingerprinting Systems Based on Compressed Suffix Array International *Journal of Information and Education Technology* **3**(4) 455-60
[10] Logan B 2000 Mel Frequency Cepstral Coefficients for Music Modeling *ISMIR*
[11] Rubner Y, Tomasi C, Guibas L J 2000 The earth mover's distance as a metric for image retrieval *International Journal of Computer Vision* **40**(2) 99-121
[12] Pampalk E 2006 Computational Models of Music Similarity and their Application in Music Information Retrieval *Vienna University*
[13] Linde Y, Buzo A, Gray RM 1980 *IEEE Transactions on Communications* **28**(1) 84-95

Authors

Xin Luo, born in October, 1972, Shanghai, China

Current position, grades: assistant professor at the School of Computer science and technology at Donghua University (Shanghai).
University studies: PhD at the Faculty of Engineering at University of Tokushima, Japan (2007).
Publications: 1 book, 20 Papers.
Scientific interests: multimedia information retrieval and pattern recognition.

Xuezheng Liu, born in November, 1989, Qingdao, Shandong, China

Current position, grades: postgraduate at Donghua University.
University studies: BS degree in information security at Donghua University.
Scientific interests: artificial intelligence and cloud computing, music information retrieval.

Youqun Shi, born in September, 1964, Xuzhou city, Jiangsu Provence, China

Current position, grades: professor at Donghua University.
University studies: PhD in Application of Computer Technology, China University of Miming & Technology.
Scientific interests: component oriented programming, soft as a service.
Publications: 1 book, 15 papers.

Ran Tao, born in October, 1975,Shanghai, China

Current position, grades: Master of Computer Technology, Senior Engineer and Master supervisor at Donghua University
University studies: Master degree at Donghua University in Shanghai in 2007
Scientific interests: wisdom education and E-commerce in cloud computing
Publications: 4 patents, 10 papers, 2 books