THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.3E 2007. 12 pp. 69-73

Music Similarity Search Based on Music Emotion Classification

Hyoung-Gook Kim*, Jangheon Kim**
*Intelligent Multimedia Signal Processing, Kwangwoon University **Technical University Berlin, Germany
(Received May 9 2007; Revised Jun 21 2007; Accepted August 2 2007)

Abstract

This paper presents an efficient algorithm to retrieve similar music files from a large archive of digital music database. Users are able to navigate and discover new music files which sound similar to a given query music file by searching for the archive.

Since most of the methods for finding similar music files from a large database requires on computing the distance between a given query music file and every music file in the database, they are very time-consuming procedures.

By measuring the acoustic distance between the pre-classified music files with the same type of emotion, the proposed method significantly speeds up the search process and increases the precision in comparison with the brute-force method.
Keywords: *Music similarity search, Music emotion classification*

I. Introduction

With the rapid growth on Internet communication technologies and music compression algorithms in recent years, digital music archives are widely spread on the World Wide Web. Nowadays, some digital music download sites can provide more than 2,000,000 songs. Even the personal computer (PC) or some portable MP3 players equipped with micro hard disk driver can store more than 1,000 songs. Thus, users may need a way to search for songs which sound similar to a query song, thereby making it easy to navigate and discover new similar songs in a large archive of digital music. However, similarity in perception is an ill-defined concept and depends on various factors such as instruments, melody, timbre, rhythm, lyrics, style, and many more. In this paper we focus on searching for songs which "*sound similar*" to a given query song.

Finding music similarity is to automatically determine acoustic similarity between different songs and it concentrates on measuring the acoustic distance between them.

The related works on the topic of similarity query are presented in [1-7]. In [1], mel-frequecy cepstral coefficients(MFCCs) are adopted as the acoustic features and the distribution of MFCC is modeled by a Gaussian mixture model(GMM). Monte Carlo distance is used to calculate the distance between two songs based on their GMM. Logan et al. [2-3] also employ MFCC as features. A signature for each song is generated based on the k-means clustering of the MFCC. The signature can then be compared using the Earth Mover's distance. Berenzweig et al. [4-5] propose new features derived from MFCC features and use neural networks as anchor model pattern classifiers. Since the input to the classifier is a large vector consisting of 5 frames of MFCC vectors plus deltas, the neural network learns time-dependent information such as rhythm and tempo. Anchor space achieves comparable results contrasting with MFCC on the survey data [5]. In [6], Park et al. propose a music retrieval and browser system with the sequential forward

Corresponding author: Hyoung-Gook Kim (hkim@kw.ac.kr)
Dept. of Wireless Communication Engineering, Kwangwoon University, 447-1, Wolgye-Dong, Nowon-Gu, Seoul, 139-701

feature selection procedure and multi-feature clustering method performing on the 54 dimensions of feature. The feature vector is composed of the mean and standard derivative of spectral centroid/roll-off/flux, zero crossing rates, MFCC, and linear predictive coefficients. Welsh et al. [7] extracts frequency histograms, volume, noise, tempo, and tonal transition features from music audio data to reduce each song to a 1248-dimensional feature vectors. Query is performed using a k-nearest neighbor search in the feature space.

Usually the computational complexity in the similarity search is huge, while real-time response is required in most cases. If the system knows the user well in advance, the search can be restricted only to the user's interest space instead of the whole music repository. This will improve both accuracy and efficiency. By studying the related work, we found that many researchers introduce "same artist/same album/same genre" or similar measures as the ground truth to evaluate the performance of similarity search. Generally, two songs belonging to quite different genre categories or bringing listeners (or users) different emotion are hardly thought as similar. Thus, in this paper we present an efficient algorithm to retrieve similar music files based on music emotion classification against a large archive of digital music.

Because the music emotion classification and the music signature for similarity search are both unchangeable for a specified song, they only need to be calculated one time and then the information can be stored in the meta-data of compressed music archives in the proposed algorithm. Since these steps are very time-consuming procedure, adding the information into meta-data can significantly improve the efficiency of the algorithm. Oppositely, the result of similarity search depends on the songs in the database and must be obtained on-line, so the speed of similarity search is a critical parameter in this system. Furthermore, considering the increasing improvements on portable devices, the system is designed to be compatible to the modern high-end MP3 players and smart phones. Therefore, three important factors are investigated to obtain the performance of the proposed system: fast feature extraction methods, efficient music emotion classification approach, and rapid on-line music similarity search scheme.

Section II elaborates a system overview including music signature extraction and similarity measure designed for efficient similarity search. The experiments are described in details in Section III. And Section IV draws the conclusion of this paper and proposes some points of future work.

II. System Overview

The proposed music similarity search system is depicted in Figure 1. Two distinct phases, such as database generation and similarity search, are distinguished.

During the database generation phase, the music signature is directly extracted from the modified discrete cosine transform coefficients (MDCT), which are the output of partial MP3 decoder. The music signature extracted from whole music database is classified into predefined music emotion classes by the emotion classification based on AdaBoost algorithm. The detailed music emotion classification is described in [8]. The music signature with the emotion classification results is stored in the music signature database.

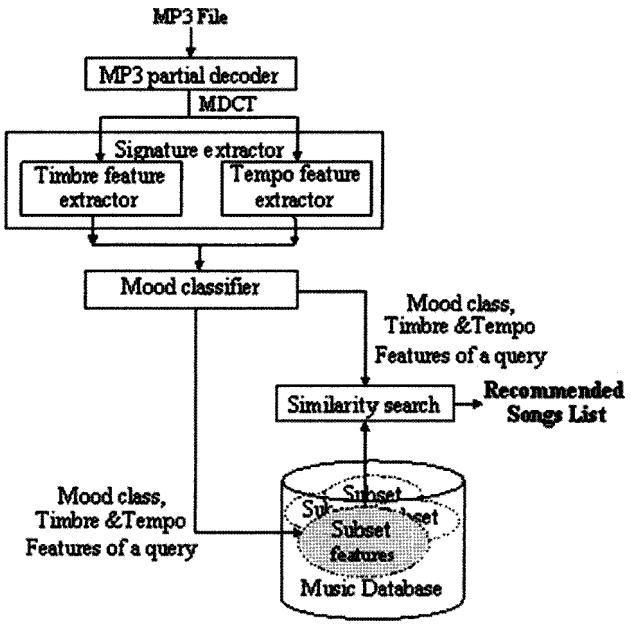

Fig. 1. Block diagram of similarity search system.

In the similarity search phase, the query music signature is extracted from the music piece. The automatic emotion classification finds the best-match emotion class for the query music. The distance between

the pre-classification music files with the same type of emotion is computed with a pre-designed similarity measure. Finally, all the distances are sorted and an orderly list of similar songs is given.

2.1. Music Feature Extraction

In order to fast search the similar songs in gigabytes data, the proposed music signature extraction must be very simple and small-size, as well as be able to capture some key characteristic in the music. Since one feature is usually related to one aspect of music, the music feature is a combination of multiple features.

For this, four timbre features and four tempo features are employed in the music signature. The four timbre features are MDCT-based subband centroid/bandwidth/flux/flatness, which have been described in [8].

The tempo features for music signature are based on the amplitude of band-pass filtered signals within the bass range (from 80Hz to 250Hz), which can be obtained by summarizing the magnitude of MDCT coefficients between the frequency range. Base range covers about 1.5 octaves. The power of main rhythmic instruments, e.g. bass drum, electronic bass, etc., distributes in this frequency domain.

The amplitude deviation between two successive frames and 3-order low-pass filter with 10Hz stop band are applied on the sub-band amplitude. Then 3-second continuous wavelet transform(WT) with eight different dyadic scales is performed on each amplitude deviation signal. Finally, the amplitude of WT $Y_\tau(j)$ is defined as *Tempo Spectrum*, is attained with 1-second time resolution, where j is the index of tempo spectrum at τ-th time interval. Based on the tempo spectrum $Y_\tau(j)$, four tempo features are computed within the tempo range from 30BPM (beats per minute) to 300BPM. In (1)-(4), J_0 and J correspond to the index of tempo spectrum on 30BPM and 300 BPM respectively.

- Tempo spectral centroid:

$$\tilde{C}_\tau = \frac{\sum_{j=J_0}^{J}\left(|Y_\tau(j)|^2 \times j\right)}{\sum_{j=J_0}^{J}|Y_\tau(j)|^2} \tag{1}$$

- Tempo spectral bandwidth:

$$\tilde{B}_\tau = \sqrt{\frac{\sum_{j=J_0}^{J}\left(|Y_\tau(j)|^2 \times \left(j-\tilde{C}_\tau\right)^2\right)}{\sum_{j=J_0}^{J}|Y_\tau(j)|^2}} \tag{2}$$

- Tempo spectral flux:

$$\tilde{F}_\tau = \sum_{j=J_0}^{J}\left(|Y_\tau(j)|-|Y_{\tau-1}(j)|\right)^2 \tag{3}$$

- Tempo spectral flatness:

$$\tilde{L}_\tau = 10\times\log\left(\frac{\sqrt[(J-J_0+1)]{\prod_{j=J_0}^{J}|Y_\tau(j)|^2}}{\left(\sum_{j=J_0}^{J}|Y_\tau(j)|^2\right)/(J-J_0+1)}\right) \tag{4}$$

Among a lot of frames of timbre features and tempo features, a simple feature vector as the music signature is used. The maximum, the average, and the standard derivative of these features across the middle 30-seconds part of song to reduce the complexity of features are chosen. Finally, a 24-dimensional feature vector $V_q = \{v_q(m)\}, m = 1,\cdots,24$ for q-th song is used as music signature in similarity search.

2.2. Similarity Measure

Similarity measure is defined as the distance between two songs' music signature V_p and V_q. The basic definition idea is that two songs sound similar and the distance between two vectors is closer. As the proposed music signature is composed of the maximum, the average, and the standard derivative of eight features, the scaling factor of each component must be quite different. Therefore, all the components in the feature vector are normalized to a spherical space by the mean and the standard derivative computed from a large music database with variety music emotion categories as shown in (5).

$$\hat{v}_q(m) = \left(v_q(m)-\mu(m)\right)/\sigma(m) \tag{5}$$

where $\hat{v}_q(m)$ is the m-th normalized feature component in the music signature of the q-th song; $\mu(m)$ and $\sigma(m)$ are the mean and the standard derivative computed from a 20,000-song database; $m = 1,\cdots,24$. In this paper, we use a Euclidean distance between two normalized vectors to measure the similarity of two songs on account of its computational economy. For p-th song and q-th song, the similarity measure is defined as (6).

$$\varepsilon = \sqrt{\sum_{m=1}^{24}\left(\hat{v}_p(m)-\hat{v}_q(m)\right)^2} \tag{6}$$

2.3. Similarity Search Scheme

Since most of the methods for finding similar music file from a large database requires on computing the distance between a given query music file and every music file in the database, they are very time-consuming procedures.

While the distance measure provides good results in small database, it does not yield useful results in the large search space (large database). To reduce the search space against large music database we apply the emotion classification to the similarity search in this paper. The basic idea comes from the criterion: "*two similar songs must be in the same music emotion*". Using the acoustic-based distance measure between the pre-classification music file with the same type of emotion, the proposed method significantly speeds up the search process and increases the precision.

In this paper all the music songs in the database are divided into 4 emotion categories (subsets) such as *sad, calm, pleasant,* and *excited.* But some songs are hard to distinguish the definitive emotion categories, as an example, sad vs. calm and pleasant vs. excited.

Since there are two confusion sets in theses four emotion sets, some songs may be false classified into one of the confusion subsets and the precision of the similar music search is decreased. To achieve high precision results against the errors confused by the emotion classification we computes the distance as following procedure: 1) If the query music file is classified to sad emotion class, the distance between the query signature and every signature with *sad* emotion subset in the music database is computed. 2) The confusion set of the *sad* emotion subsets is the *calm* emotion subset. Thus, the distance between the query signature and every signature with the *calm* emotion subset in the music database is computed. 3) Finally, all the distances are sorted and an orderly list of similar songs is given.

The advantages of this method are two-fold:

- It can improve the search speed significantly versus brute-force method.
- Music signature used in similarity search is a highly compressed feature, whereas the music emotion classification results come from multiple frames matching. The latter is more precise and can effectively remove the serious false positive errors.

III. Experiment

The proposed similarity search on musical data is dependent on music emotion classification results. As the criterion of only selecting relatively consistent and widely accepted music emotions, we choose four kinds of music emotions categories: 1) *sad,* 2) *calm,* 3) *pleasant,* and 4) *excited.* Corresponding to each music emotion, there are 500 songs in the dataset, totally 2000 (4×500) songs. Seven independent trained listeners label each song. Only the song, which is consistently agreed as the same emotion category by seven listeners, is accepted by the dataset.

With AdaBoost algorithm combining timbre and tempo features, the music emotion classification precision can reach 93.2%. Using tempo features based on wavelet transform instead of fast Fourier transform we obtained the improved classification results although DB size gets larger compared to [8].

The confusion matrix of the result of the proposed method is shown in Table 1.

Table 1. Music Emotion Classification Confusion Matrix.

Emotion types	Sad	Calm	Pleasant	Excited
Sad	438	62	0	0
Calm	13	481	6	0
Pleasant	4	0	484	12
Excited	0	0	39	461

From the confusion matrix of Table I, there are two confusion sets in these four music emotion categories: sad vs. calm and pleasant vs. excited, which are also hard to be decided for listeners' perception. Especially for the 500 sad songs, there are 62 songs false recognized as calm songs. After checking the errors, we find that some songs express the sad emotion only by the content of lyrics, but the proposed features only can represent the timbre and tempo information of music songs. In this case, it is impossible to distinguish sad songs from calm songs.

2000 songs are chosen as the testing set and labeled as 100 similar sets. Each similar set comprises 20 songs which have the same artist and the same emotion, and perceptually sound similar. By performing the similarity search method on each song in the testing set, we can count the quantity of the songs which are in the same similar set with the query song. The average recall

numbers counted from top-10, top-20, top-50, top-100 recommendation lists are shown in Table 2. "Pre-class" represents the proposed method which is based on the results of emotion classification. "Brute-force" is the comparative brute-force method which searches the similar songs among the whole database. In addition, the speed ratio of our proposed method versus the brute-force searching method is also given in Table 2.

Table 2. Similarity Query Results with/without Pre-classification.

Rank list	Average Recall Number		Average Precision Number		Speed ratio (Times)
	Pre-class	Brute-force	Pre-class	Brute-force	
Top-10	5.71	5.53	7.56	7.23	2.00
Top-20	5.92	5.45	12.43	12.15	2.01
Top-50	8.52	8.03	17.39	17.12	2.03
Top-100	9.72	9.47	19.42	19.32	2.06

The averaged recall number is the sum of all recall values divided by the number of the recall case used, while the recall is the number of relevant retrieved music files divided by the number of relevant music files in the database. The precision is the number of relevant retrieved music files divided by the number of retrieved music files. Obviously, the high precision of music emotion classifier guarantees the high performance of the proposed similarity search method. The distance between the pre-classified music files with the same type of emotion is smaller than the distance between the music files with the different type of emotion. With the emotion classification scheme, the searching speed, the average recall rate, and the average precision rate are significantly improved.

For the application of top-10 similarity query on the 20,000-songs database, averagely 40,000 times of comparison and data exchanging operations are needed. Performing it on a high-end modern PDA (CPU: Intel Xcale 300MHz, RAM: 40MB), each query can be finished with 0.1 second.

IV. Conclusion and Future Work

In this paper, we propose an efficient music similarity search system based on the results of music emotion classification. A new tempo feature based on MDCT and wavelet transformation is combined with conventional timbre features in order to significantly improve the precision of music classification and similarity results. Furthermore, high precision of music classification results in better recall rate and higher search speed than the traditional brute-force searching scheme.

Our future work is to integrate emotional contents and lyrics in similarity search. Of geometric interest is the investigation of non-Euclidean distance metrics as well as the application of efficient neighbor-finding algorithms. Thus, we can obtain a more compact and diverse music signature for different music clusters.

References

1. J. Aucouturier, and F. Pachet, "Improving timbre similarity: How high's the sky?," Journal of Negative Results in Speech and Audio Sciences, Apr. 2004.
2. B. Logan, and A. Salomon, "A music similarity function based on signal analysis," in Proc. of the 2001 IEEE International Conf. on Multimedia and Expo (ICME' 01), 745-748, Aug. 2001.
3. B.Logan, "Music recommendation from song set" in Proc. of 5th International Conf. on Music Information Retrieval 2004 (ISMIR2004), Oct. 10-15, 2004.
4. A. Berenzweig, D. Ellis, and S. Lawrence, "Anchor Space for Classification and Similarity Measurement of Music," in Proc. of the 2003 IEEE International Conf. on Multimedia and Expo (ICME' 03), 1, 29-32, 2003.
5. A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, "A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures," in Proc. of 4th International Conf. on Music Information Retrieval 2003 (ISMIR2003), Oct. 26-30, 2003.
6. K. S. Park, W. J. Yoon, K. K. Lee, S. H. Oh, and K. M. Kim, "MRTB framework: a robust content-based music retrieval and browsing," IEEE Trans. on Consumer Electronics, 51 (1), 117-122, Feb. 2005.
7. M. Welsh, N. Borisov, J. Hill, R. Behren, and A. Woo, "Querying large collections of music for similarity," UC Berkeley Technical Report UCB/CSD-00-1096, U.C. Berkeley Computer Science Division, Nov. 1999.
8. H-G. Kim, K.-W. Eom, "Automatic emotion classification of music signals using MDCT-driven timbre and tempo features," The Journal of the Acoustical Society of Korea, 2006.

[Profile]

• **Hyoung-Gook Kim,**

The journal of the Acoustical Society of Korea, vol.26 (2E)
Present: Professor in the Wireless Communication Engineering, Kwangwoon University, Korea

• **Jangheon Kim**

2000: B.S. degree in Yonsei University, Korea
2002: M.S. degree in Yonsei University, Korea
2003: Research Scientist in Heinrich Hertz Institute (HHI), Germany
2004~present: Ph.D. Student in Technical University Berlin, Germany