MUSIC TYPE CLASSIFICATION BY SPECTRAL CONTRAST FEATURE[1]

Dan-Ning Jiang, Lie Lu**, Hong-Jiang Zhang**, Jian-Hua Tao*, Lian-Hong Cai**

*Department of Computer Science and Technology, Tsinghua University, China
jdn00@mails.tsinghua.edu.cn, jhtao@tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

**Microsoft Research, Asia
{llu, hjzhang}@microsoft.com

ABSTRACT

Automatic music type classification is very helpful for the management of digital music database. In this paper, Octave-based Spectral Contrast feature is proposed to represent the spectral characteristics of a music clip. It represented the relative spectral distribution instead of average spectral envelope. Experiments showed that Octave-based Spectral Contrast feature performed well in music type classification. Another comparison experiment demonstrated that Octave-based Spectral Contrast feature has a better discrimination among different music types than Mel-Frequency Cepstral Coefficients (MFCC), which is often used in previous music type classification systems.

1. INTRODUCTION

Music is very popular in modern life, and the amount of digital music increases rapidly nowadays. How to manage a large digital music database has arisen as a crucial problem. Automatic music type classification could be very helpful for the music database management. Although music type is not a very clear concept, music still could be divided into two major categories: classical music and popular music. Classical music, which is opposed to popular music, is intended to include all kinds of "serious" music, while popular music means "music of the populace" [6]. For each major music category, it could be further divided into some small classes due to different periods or different music style. In our music type classification system, classical music is further classified into baroque music and romantic music, which correspond to the baroque era and romantic era in western music history respectively [6]; popular music is further classified into three types, which include pop songs, jazz, and rock. Thus, five types are classified in our system.

There are many music characteristics could be used to discriminate different music type, such as the musical structure, tempo, rhythm, melody, chord, and so on. However, it is extremely difficult to extract them precisely by signal processing methods for most digital music. Therefore, many previous researches turned to spectral characteristics, which are found useful for discriminating different music types and easy to be extracted. Matityaho [1] applied multi-layer neural network on the average amplitude of Fourier transform coefficients to separate classical and pop music. Han [2] used the nearest mean classifier to classify music into classical music, jazz, and popular music with some simple spectral features. Soltau [3] used HMM and ETM-NN method to extract the temporal structure from the sequence of cepstral coefficients, and implemented a music type classification system for rock, pop, techno and classic. Pye [4] used Gaussian Mixture Model (GMM) and Mel-Frequency Cepstral Coefficients (MFCC) to obtain a best classification result in his system, which includes six types of blues, easy listening, classical, opera, dance and indie rock. However, while developing different models to improve the performance of music type recognition system, most of these work used average spectral envelope (such as MFCC) to represent the spectral characteristics of music. This kind of features averages the spectrum in each sub-band and reflects the average spectral characteristics, but it could not represent the relative spectral characteristics in each sub-band, which seem more important to discriminate different types of music.

In this paper, Octave-based Spectral Contrast feature is proposed to represent the relative spectral characteristics of music. Octave-based Spectral Contrast feature considers the strength of spectral peaks and spectral valleys in each sub-band separately, so that it could represent the relative spectral characteristics, and then roughly reflect the distribution of harmonic and non-harmonic components. Experiments showed that Octave-based Spectral Contrast feature had good discrimination in music type classification and performed better than MFCC feature.

The rest of this paper is organized as follows. Section 2 discusses the representation of Octave-based Spectral Contrast feature in detail. Our classification scheme is described in Section 3. In Section 4, experiments are performed to evaluate the proposed feature.

2. OCTAVE-BASED SPECTRAL CONTRAST REPRESENTATION

Octave-based Spectral Contrast considers the spectral peak, spectral valley and their difference in each sub-band. For most music, the strong spectral peaks roughly correspond with harmonic components; while non-harmonic components, or

[1] The work was performed at Microsoft Research Asia

noises, often appear at spectral valleys. Thus, Spectral Contrast feature could roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum. Previous features, such as MFCC, average the spectral distribution in each sub-band, and thus lose the relative spectral information. Considering two spectra that have different spectral distribution may have similar average spectral characteristics, the average spectral distribution is not sufficient to represent the spectral characteristics of music. However, Spectral Contrast keeps more information and may have a better discrimination in music type classification.

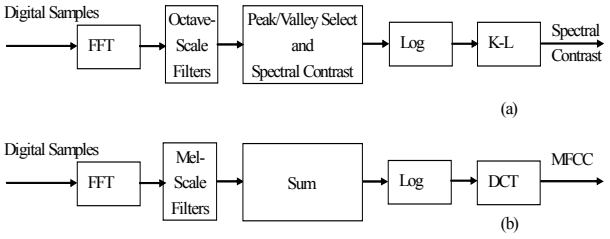

Fig. 1. The comparison of (a) Octave-based Spectral Contrast and (b) MFCC

Fig. 1 (a) illustrates the estimation procedure of Octave-based Spectral Contrast feature. FFT is first performed on the digital samples to obtain the spectrum. Then, the frequency domain is divided into sub-bands by several Octave-scale filters. The strength of spectral peaks, valleys, and their difference are estimated in each sub-band. After being translated into Log domain, the raw Spectral Contrast feature is mapped to an orthogonal space and eliminated the relativity among different dimensions by Karhunen-Loeve transform.

The above procedure is to estimate Octave-based Spectral Contrast feature from one frame. For a music clip or a music piece, the mean vector and standard deviation vector of all of its frames are used to represent its spectral characteristics.

The estimation procedure of MFCC is also listed in Fig. 1 (b) to compare with that of Octave-based Spectral Contrast feature. There are some differences between the two procedures:

(1) The filter bank is different. Octave-based Spectral Contrast feature uses octave-scale filters, while MFCC uses Mel-scale filters. Although Mel-scale is suitable for general auditory model, octave-scale filter is more suitable for music processing. In our implementation, the frequency domain is divided into six Octave-scale sub-bands, which are 0hz~200hz, 200hz~400hz, 400hz~ 800hz, 800hz~1600hz, 1600hz~3200hz, and 3200hz~8000hz (the sample rate is 16khz). Since the Spectral Contrast feature is based on Octave-scale filters, the feature is named as Octave-based Spectral Contrast. It will be simplified as Spectral Contrast for convenience in the left of this paper.

(2) Spectral Contrast extracts the strength of spectral peaks, valleys, and their difference in each sub-band, while MFCC sums the FFT amplitudes. Thus, Spectral Contrast feature represents the relative spectral characteristics, but MFCC only involves the average spectral information. Spectral Contrast includes more spectral information than MFCC.

(3) At the last step, Spectral Contrast feature uses a K-L transform while MFCC uses a DCT transform. They are equivalent from the view of eliminating relativity. It should be noticed that the orthogonal base vectors for K-L transform are got from the training data set.

2.1. Raw Spectral Contrast Feature Estimation

In features extraction, the music piece is first segmented into frames by 200ms analysis window with 100ms overlapping. For each frame, FFT is performed to get the spectral components and then it is divided into six octave-based sub-bands. Finally, Spectral Contrast is estimated from each octave sub-band.

The raw Spectral Contrast feature estimates the strength of spectral peaks, valleys and their difference in each sub-band. In our scheme, in order to ensure the steadiness of the feature, the strength of spectral peaks and spectral valleys are estimated by the average value in the small neighborhood around maximum and minimum value respectively, instead of the exact maximum and minimum value themselves. Thus, neighborhood factor α is introduced to describe the small neighborhood. Detailed expressions are as follows:

Suppose the FFT vector of k-th sub-band is $\{x_{k,1}, x_{k,2}, \cdots, x_{k,N}\}$. After sorting it in a descending order, the new vector can be represented as $\{x'_{k,1}, x'_{k,2}, \cdots, x'_{k,N}\}$, where $x'_{k,1} > x'_{k,2} > \cdots > x'_{k,N}$.

Then the strength of spectral peaks and spectral valleys are estimated as:

$$Peak_k = \log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\} \quad (1)$$

$$Valley_k = \log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,N-i+1}\} \quad (2)$$

And their difference is:

$$SC_k = Peak_k - Valley_k \quad (3)$$

where N is total number in k-th sub-band, $k \in [1,6]$.

Different values of α from 0.02 to 0.2 are tested in experiments. It shows that varying α in this range does not influence the performance significantly. In real implementation, α is set to be 0.02.

$\{SC_k, Valley_k\}$ ($k \in [1,6]$) is used as the 12-dimension raw Spectral Contrast feature. Although Spectral Contrast means the difference of strength between the spectral peaks and valleys, the strength of spectral valleys are also contained in the feature to keep more spectral information.

2.2. Karhunen-Loeve Transform

It is obvious that there exist some relativity among the different dimensions of raw feature. To solve this problem, Karhunen-Loeve transform is performed on the raw feature to remove the relativity. After K-L transform, the feature vector is mapped into an orthogonal space, and the covariance matrix also becomes diagonal in the new feature space. These properties of K-L

transform make the classifying procedure easier and lead to good classification performance even with simple classifier.

In our experiments, the matrix that generates the orthogonal base vectors is estimated from the covariance matrix of each class. It is represented as:

$$S_w = \sum_{i=1}^{5} P_i \Sigma_i \tag{4}$$

where S_w is generate matrix, P_i and Σ_i are the prior probability and the covariance matrix of the i-th class respectively. In experiments, P_i is set to be 0.2, which corresponds to equal probability distribution for each class; Σ_i is estimated from the training set of the i-th music type class. The orthogonal base vectors are the eigenvectors of the generate matrix S_w. Then the transformation is done as below:

$$x' = Ux \tag{5}$$

$$U = [u_1, u_2, \cdots, u_j, \cdots, u_D]^T \tag{6}$$

where x is the raw feature vector, x' is the Spectral Contrast feature vector after K-L transform, D is the dimension of the feature space, and u_j is the j-th orthogonal base vector.

3. CLASSIFICATION SCHEME

In general, human could discriminate a music type in several seconds, such as 10 seconds. Therefore, our classification scheme is first based on 10-second music clips. Then the classification experiments on whole music are also performed.

Gaussian Mixture Model (GMM) with 16 components is applied in our approach and Expectation Maximization (EM) algorithm is used to estimate the parameters of GMM model for each music type.

Let x be the feature vector of a 10-second music clip, then the probability density (mixture density) of this music clip belonging to class-i is defined as:

$$p(x \mid G_i) = \sum_{j=1}^{16} w_{ij} \mathsf{N}(x, u_{ij}, C_{ij}) \tag{7}$$

where G_i is the GMM model of the i-th class; w_{ij}, u_{ij} and C_{ij} are the weight, mean vector, and covariance matrix of the j-th component in G_i, respectively.

The classification is easy to proceed. As usual way, each clip in the testing set is classified into the class that has the largest probability density according to Bayesian criterion.

The performance can be increased when the whole music is used as classification unit instead of 10-second clip. In order to classify a whole piece of music, the music is first divided into several 10-second clips. Final classification result of the music is determined by combing the probabilities of every clip.

Suppose there are N independent 10-second clips in a whole piece of music, and the feature set is $X = \{x_1, x_2, \cdots, x_N\}$, then the probability density of the whole music in class-*i* can be calculated as following:

$$P(X \mid G_i) = \prod_{j=1}^{N} p(x_j \mid G_i) \tag{8}$$

The classification is then determined by the maximum probability density.

In real implementation, one 10-second music clip is extracted from every 30 seconds in each piece of music in order to decrease the computation complexity.

4. EXPERIMENTS

4.1. Database for Experiments

There are about 1500 pieces of music in our database for experiments, and five music types are included: baroque music, romantic music, pop songs, jazz, and rock. Most of the baroque pieces in the database are literatures of Bach and Handel, who are the most important composers in the baroque era. The romantic database is composed of literatures of Chopin, Schubert, Liszt, Beethoven, and other composers in the romantic era. Pop songs are those singed by some popular singers, which includes nine men and sixteen women. Jazz and rock in the database also include literatures of many different composers. In each music type database, different possible musical form and musical instruments are included.

All the music data in the database are 16kHz, 16 bits, mono wave files. About 6250 10-second clips, which are randomly selected from the 1500 pieces of music, compose the classification database, where 5000 is for training and 1250 for testing. For each music type, there are about 1000 clips in the training set, and about 250 clips in the testing set. 10-second clips from the same music piece would not appear both in the training set and testing set. In the classification experiments on whole music, the training data is the same as those for 10-sencond music clips, while the testing data is composed by the music piece whose clips are presented in the original testing data set.

4.2. Classification Results

An experiment is first performed on 10-second clips by using Spectral Contrast. The mean and standard deviation of Spectral Contrast composes a 24-dimension feature for a music clip. The classification performance is pretty good, and the average accuracy reaches 82.3%. The detailed classification results are listed in Table 1.

	Baroque	Romantic	Pop	Jazz	Rock
Baroque	83.2%	12.8%	0.4%	3.6%	0.0%
Romantic	12.9%	84.2%	0.8%	1.2%	0.8%
Pop	1.6%	2.4%	78.4%	11.6%	6.0%
Jazz	2.0%	0.4%	15.2%	78.4%	4.0%
Rock	0.4%	0.8%	6.0%	5.6%	87.2%

Table 1. The detailed classification results on 10-second clips

From Table 1, it could be seen that the classification error rate between the baroque and romantic music is high, while few clips of these two types are classified into the other three classes by mistakes. This is because that the baroque and romantic music both belong to classical music and thus their spectral characteristics are similar. The same phenomena could be seen from pop songs, jazz and rock.

We also performed an experiment on classification of whole music piece. The detailed results are shown in Table 2.

	Baroque	Romantic	Pop	Jazz	Rock
Baroque	86.7%	10.0%	0.0%	3.3%	0.0%
Romantic	7.3%	90.9%	0.00%	1.8%	0.00%
Pop	0.0%	0.0%	92.3%	6.2%	1.5%
Jazz	1.7%	0.0%	5.2%	91.4%	1.7%
Rock	0.0%	0.0%	4.5%	3.0%	92.5%

Table 2. The detailed classification results on whole music piece

From Table 2, the average classification accuracy on whole music piece is up to 90.8%, which is much higher than 82.3% on 10-second clips. The classification error rate of each music class decreases much.

4.3. Comparison with MFCC

Mel-Frequency Cepstral Coefficients (MFCC) are widely used in audio classification [5] and music classification [3][4]. It has been proven that MFCC performs well in these tasks. It is also reported that adding an energy term with MFCC features could greatly improve the performance of music type classification [4]. So, in this comparison experiment, we will compare the performance among the following three feature sets: Spectral Contrast, MFCC with Energy term, and MFCC without Energy term. The comparison experiments are only implemented on our testing set of 10-second clips.

As Spectral Contrast, 12-order MFCC features are extracted from each frame. Then, the mean and the standard derivation of the MFCC, which compose a 24-dimension feature set, are estimated to represent the music clip. When energy term is considered, the feature is 26-dimension.

Table 3 listed the average classification accuracy when using different feature set.

Feature Set	Classification Accuracy
MFCC	74.1%
MFCC + Energy	78.0%
Spectral Contrast	82.3%

Table 3. The average classification accuracy when using MFCC, MFCC + Energy, and Spectral Contrast

Table 3 shows that Spectral Contrast performs better than the other two feature sets in music type classification. The classification accuracy reaches to 82.3% with Spectral Contrast, which is 8.2% and 4.3% higher than MFCC and MFCC with energy term, respectively.

Fig. 2 illustrates the detailed comparison results. From Fig. 2, it can be seen that the classification error rate decreases 20%-30% for each music type when using Spectral Contrast instead of MFCC. It proves that the new proposed Spectral Contrast feature can improve the music classification performance satisfactorily.

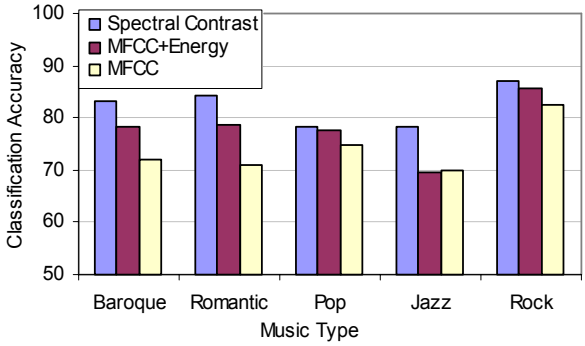

Fig. 2. The details of the comparison results of Spectral Contrast and MFCC features

5. CONCLUSION

This paper presented a set of new feature named Octave-based Spectral Contrast. Spectral Contrast deals with the strength of spectral peaks, valleys, and their difference separately in each sub-band, and represents the relative spectral characteristics. Based on Spectral Contrast feature, an automatic music type classification system is implemented to classify music into five classes, which include baroque music, romantic music, pop songs, jazz, and rock. An average accuracy of 82.3% is achieved for classification on 10-second music clips, and 90.8% is achieved on whole music pieces. Our comparison experiment also showed that the proposed Spectral Contrast feature had a better performance than MFCC feature in music type classification.

6. REFERENCES

[1] B. Matityaho and M. Furst. "Neural Network Based Model for Classification of Music Type", in Proc. of 18th Conv. Electrical and Electronic Engineers in Israel, pp. 4.3.4/1-5, 1995.

[2] K. P. Han, Y. S. Pank, S. G. Jeon, G.C. Lee, and Y. H. Ha, "Genre Classification System of TV Sound Signals Based on a Spectrogram Analysis", IEEE Transactions on Consumer Electronics, VOL. 55, No. 1, pp. 33-42, 1998.

[3] H. Soltau, T. Schultz, Martin Westphal, and Alex Waibel. "Recognition of Music Types". ICASSP 1998, Vol. II, pp. 1137-1140, 1998.

[4] D. Pye. "Content-Based Methods for the Management of Digital Music". ICASSP 2000, Vol. IV, pp.2437-2440, 2000.

[5] D. Li, I. K. Sethi, N. Dimitrova, T. McGee, "Classification of General Audio Data for Content-Based Retrieval", Pattern Recognition Letters, VOL. 22, No. 5, pp. 533-544, 2001.

[6] S. Sadie as Editor, "The Cambridge Music Guide", Cambridge University Press, 1985.