MUSICAL GENRE CLASSIFICATION BY INSTRUMENTAL FEATURES

Jiajun Zhu, Xiangyang Xue, Hong Lu

Dept. of Computer Science and Engineering, Fudan University, Shanghai, China

{zhujiajun, xyxue, honglu@fudan.edu.cn }

Abstract

Automatic musical genre classification is very useful for many musical applications. In this paper, the features of instrument distribution and instrument-based notes are proposed to represent the high-level characteristics of music. Experimental results show that the proposed features have a good performance in musical genre classification. Comparison between our proposed features with the commonly used features --- Mel-Frequency Cepstral Coefficients (MFCC) and MFCC with energy term illustrates that our proposed features perform better in discriminating some musical genres, such as Pop, Jazz, and Rock.

1 Introduction

The rapid development of Internet and technologies for multimedia compression such as MPEG, have greatly increased the amount of digital music. How to manage large digital music database has become a very crucial problem. Automatic musical genre classification is very useful for Music Information Retrieval (MIR) [Chai and Vercoe (2001), Pye (2000)]. While such tasks can be easily accomplished by human beings, the difficulties in description of different music types remain a big challenge for the computers.

There are many musical characteristics, such as tempo, musical structure, melody, and rhythm that can be used to discriminate different music types. According to the knowledge about music, it is easy to discriminate rock with classic music by the kinds of instruments played in their performances because drums appear in almost all rock music while classic music is often played by piano and violin. Melodies and tempos are also very useful in discrimination between jazz and classical music, since although jazz and classical music can be both played in solo piano, they are composite using different techniques by artists. Unfortunately, though the problem of music trancription has been studied for more than 30 years, there is still no efficient and satisfying signal processing method to precisely extract those perceptual features from most unstructured music formats such as mp3 and wav.

Previous work on music genre classification can be divided into 2 categories: one focuses on extracting high-level music characteristics from structured music such as MIDI files and studying different models in modeling melodies and musical structures. In 1997, a machine learning approach is proposed to build classifiers and several features were extracted from MIDI music to recognize the music styles (Dannenberg, Thom, and Watson 1997). Chai and Vercoe (2001) modeled the melody feature by hidden Markov models (HMM) to classify folk music from different countries. And another group of researches turn to focus on spectral characteristics and implement them in practical systems for classifying real music. In 1995, multi-layer neural network was used on the average amplitude of Fourier transform coefficients to separate music into classic and pop (Matityaho and Furst 1995). Considering the temporal information, Soltau (1998) used HMM and Explicit Time Modeling with Neural Network (ETM-NN) to extract the temporal structure from the cepstral coefficients to classify music into rock, pop, techno, and classic. Pye (2000) extracted the Mel-Frequency Cepstral Coefficients (MFCC) features and used the Gaussian mixture model (GMM) as classifier to classify six music types which include blues, easy listening, dance, classic, opera, and India rock. Jiang (2002) proposed an octave-based spectral contrast feature which can present the relative spectral characteristics to classify music into baroque, romantic, pop, jazz, and rock.

In this paper, we propose an approximate method to extract three new features to represent the high-level characteristics of unstructured music, such as mp3 and wav. The distribution of instruments is proposed to represent the percentage that each group of instruments is played in the music performance. And the means and standard deviations of notes in each instrument group are extracted to represent the instrument-based melody and the statistical features on some time range. Experiments show that these three features have good discriminating performance and are more representative than MFCC features in discriminating the musical genre such as pop, jazz and rock

The rest of the paper is organized as follows. The representation of these novel features will be given in Section 2. Section 3 describes our approximate methods to

extract these features in detail. In Section 4, experiments were performed to evaluate our features. We conclude our work in Section 5.

2 Feature Representation

The distribution of instruments is a measure of the percentage of each instrument that is played in the music performance. In other words, this feature reflects that to what degree each instrument is important in the music. The means and standard deviations of the notes are extracted separately according to their instruments. These features consider the properties of performance in each instrument and can give us rough contours of the music scores. Sieger (1997) proposed a concept of "dictionary" to detect pitch. In the following, we extend the idea of "dictionary" and propose our method to extract both pitches and instruments' information.

2.1 Dictionary of Instruments

A dictionary of instruments' spectrum is constructed in order to get enough information on different instruments. Fig. 1 shows the spectrums of violin and piano.

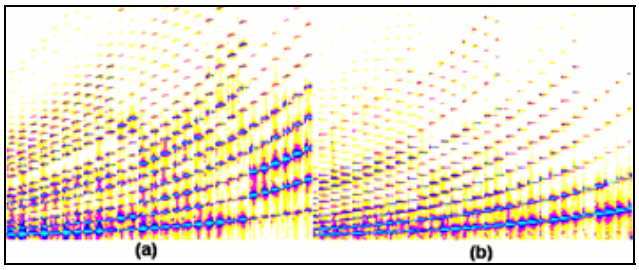

Fig. 1 Spectrum of (a) **Violin** and (b) **Piano**.
[From note B at octave 4 to note F# at octave 8]
(Frequency ranges from 493.8833HZ to 6644.875HZ)

The spectrum can be obtained by performing FFT on the sample audios. The dictionary D is made up with the spectrum of J different representative instruments and each instrument is represented by K notes.

$$D = \begin{bmatrix} D_{1,1}, & \cdots & D_{1,N} \\ \vdots & \ddots & \vdots \\ D_{M,1}, & \cdots & D_{M,N} \end{bmatrix} \quad (1)$$

where M is number of frequency bins in spectrum, and $N = K \times J$.

This dictionary works in the extraction notes and instrument types as follows; if instruments and notes of each instrument are sufficiently collected in the dictionary and representative enough, the spectral vectors of each audio frame can be represented as a linear combination of the vectors in dictionary. Suppose A is the matrix of spectrum of each frame in the target music, we can extract the note and its instrument name by calculating the matrix X in the following formulation:

$$DX = A \quad (2)$$

We will discuss how to solve this formulation in more detail and give an approximate method to obtain X in Section 3.

2.2 Distribution of Instruments

It can be determined from X that at every frame what instrument and what note are most likely to be played. All these J instruments are divided into L groups by a mapping function ϕ , where $\phi(j) = l, (0 < j \le J, 0 < l \le L)$ and the mapping is many-to-one. Thus, the feature for distribution of each instrument group ($P_1 \ldots P_i$) is extracted in two steps:

i) Calculate the importance of each instrument group

$$P_i = \sum_{t=1}^{M} \sum_{s=1}^{n} X_{t,s} \quad (\text{For } \phi(s) = i) \quad (3)$$

ii) Normalize the features to [0.0-1.0]

$$P_i = P_i / \sum_{t=1}^{L} P_t \quad (0 < i \le L) \quad (4)$$

It should be noted that the features are calculated on clips, each of which contains n frames.

2.3 Distribution of Instrument-Based Notes

The distribution of instrument is an L-dimension feature, which reflects the importance of each instrument group. The means and standard deviations of notes in each instrument group compose the rest $2L$-dimension features:

Means:

$$\mu_i = \frac{1}{n} \sum_{t=1}^{M} \sum_{s=1}^{n} X_{t,s} \times (s \bmod K) \quad (5)$$

$$(\text{For } \phi(s) = i \text{ and } 0 < i \le L)$$

Standard Deviations:

$$\sigma_i = \sqrt{\frac{1}{n} \sum_{t=1}^{M} \sum_{s=1}^{n} X_{t,s} [(s \bmod K) - \mu_i]^2} \quad (6)$$

$$(\text{For } \phi(s) = i \text{ and } 0 < i \le L)$$

3 Approximate Extraction Method

There are many possible solutions to matrix X when $N >> M$, which is usually the case in practice. It is reasonable to define the optimal solution as the one which contains greatest number of zeros for each column. This depends on the assumptions that few musicians play a lot of notes at the same time, and the number of instruments involved in a song rarely exceeds 3 or 4. Under such conditions, we still have no efficient method to obtain the optimal solution for Equation (2).

3.1 Decomposition

If we can obtain the correct information of dominant notes and instruments, it is acceptable for some errors and inaccuracies in solving the optimal solution of Equation (2). We then propose a simple approximate solution to X, which was proven to be effective in classification of music.

First, all these notes in the dictionary are scanned to find the nearest one to the target vector. The distance between note D_i and target vector A_j can be measured by the 1-norm distance as below.

$$\text{Distance}_{i,j} = \sum_{k=1}^{M} \left| D_{k,i} - A_{k,j} \right| \qquad (7)$$

where $(D_{1,i},...,D_{M,i})^T$ is the candidate note and $(A_{1,j},...,A_{M,j})^T$ is the target vector.

Then the nearest note is removed from the target vector. When the decomposition is performed, a weight is given to the removed note to represent the probability of the dominance of the note. The weight can be simply calculated as the reciprocal of the distance. When the target vector is output, it is adjusted by removing corresponding component and then normalized again.

This procedure is repeated until the target vector has been decomposed for 4 times or nearly a zero vector.

3.2 Decomposition Result

Fig. 2 illustrates the result of decomposition for three different audio signals by using our method.

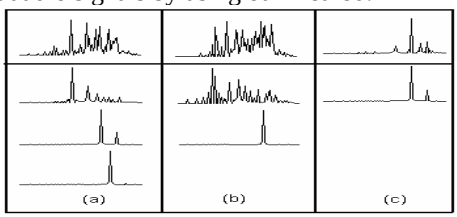

Fig. 2 Three Different signals (in the first row of the figure) are decomposed into (a) 3 (b) 2 (c) 1 note(s) by using our method.

It can be seen from Fig. 2 that this method works pretty well in decomposing audio frames (60ms) into combination of note signals. For example, in Figure.2 (a), a signal is recognized as a combination of 3 notes; in Fig. 2 (b), the signal is decomposed into 2 notes, and in Fig. 2 (c), single note is enough to represent the target signal.

Fig. 3 gives the decomposition result of rock music sample in temporal order. In the figure, the horizontal direction is the time and the vertical direction is 8 instruments. The instruments are represented in different colors and the instruments include piano, guitar, violin, drum, etc. The different position in the vertical direction represents the note information.

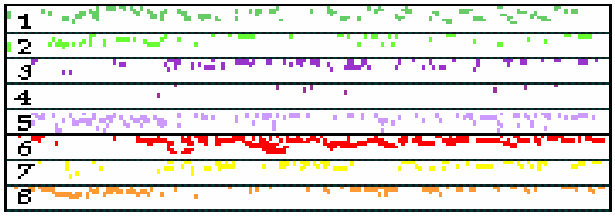

Fig. 3 Decomposition result of a rock music sample.

It can be seen from Fig. 3 that the music contains about 4 main instruments (which represented by the 1st, 3rd, 5th, and 6th rows respectively). And the guitar like instrument plays the central role performance, which is represented in the 6th row. Furthermore, note information can also be retrieved by the trajectory of lines and position of dots.

4 Experiments

In the experiment, we evaluate the classification performance by using our proposed features. The classifier we employed is GMM with 16 components. In classification, the probability of the whole music is calculated by multiplying the probability of each clip. The classification method is same to that used in Jiang's work (2002).

4.1 Dictionary for Experiments

30 MIDI instruments are selected from the total 127. We assume that these 30 instruments can approximately represent the instruments in real music. Table 1 lists all the instruments which are grouped into 8 instrument types, which are illustrated in the table by different background colors.

0	Acoustic Piano	9	Acoustic Guitar	19	Soprano Sax
1	Bright Piano	10	Acoustic Guitar(steel)	20	Alto Sax
2	Electric Grand Piano	11	Electric Guitar(jazz)	21	Tenor Sax
3	Honky Tonk Piano	12	Guitar Harmonics	22	Baritone Sax
4	Church Organ	13	Violin	23	Piccolo
5	Reed Organ	14	Viola	24	Flute
6	Accordion	15	Cello	25	Whistle
7	Harmonica	16	String Ensemble 1	26	Steel Drums
8	Tango Accordion	17	String Ensemble 2	27	Woodblock
		18	Synth Strings 1	28	Taiko Drum
				29	Melodic Tom

Table 1 The 30 instruments we selected in our dictionary. 0-3 are in Piano Group, 4-8 are in Organ Group, 9-12 are in Guitar Group, 13-15 are in Strings Group, 16-18 are in Ensemble Group, 19-22 are in Reed Group, 23-25 are in Pipe Group, and 26-29 are in Percussive Group.

For each instrument, 79 notes are sampled in 11,025Hz (each note with length of 3 seconds). The duration of the sample data is 1 hour and 58 minutes in all. Thus, the dictionary we built is a matrix of 95×2370 dimensions.

4.2 Database for Experiments

In our experiments, we collected 1,699 mp3 songs and music, including 667 for rock type, 280 for classical music, 487 for jazz, and 255 for pop. The classical music database includes literature by Beethoven, Chopin, Mozart, Schubert, Bach, and Liszt. The pop music database consists of 14 male and 12 female singers' albums. In each type of music database, different musical instruments are included.

These 1,699 mp3s are all first converted into 11,025 Hz, 16 bits, mono wave files. Every song is divided into clips with each clip lasting 10 seconds. Then, the music database consists of 31,004 10-second-clips. We randomly select

four-fifths clips for training models and the remaining one-fifth clips for testing.

4.3 Experiments Results

Experiments were first performed to classify the music into pop, jazz, rock and classic by using our proposed features. The distribution of instrument group composes the first 8-dimension features and the means and standard deviations of notes in each group compose the rest 16-dimension features. Since Pye (2000) reported that adding an energy term with MFCC features could have better performance of musical genre classification, we also extracted the mean and standard deviation of MFCC and energy term from 60ms frames for comparison. Fig. 4 illustrates the classification result on our database by using both 24-dimension new feature and 26-dimension MFCC + Energy feature.

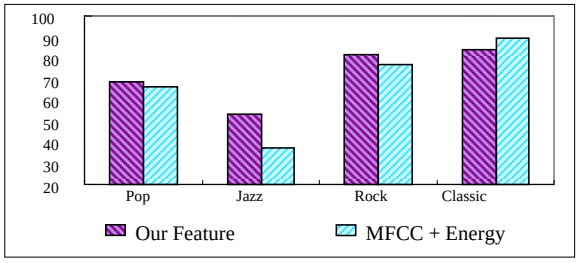

Fig. 4 Comparison results of our instrumental features and MFCC + Energy.

It can be seen from Fig. 4 that our features perform quite satisfactorily in most cases of classifying music, especially in jazz, the accuracy of our feature are higher than the MFCC + Energy method by 19%. Although for classic, the performance of our proposed features is lower than that of MFCC + Energy, the classification performance is already satisfactory, i.e. accuracy of 86%.

We also perform the experiment on testing the performance of adding our feature to MFCC + Energy feature. Fig.5 shows the experimental result by our features, MFCC, MFCC + Energy, and all of them. It can be seen that adding our features to MFCC + Energy feature can significantly improve the performance of classification.

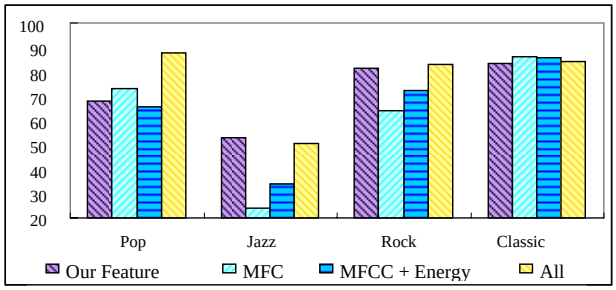

Fig. 5 Comparison result of our feature, MFCC, MFCC + Energy, and all of them.

We also perform experiment on further classifying music types such as classifying pop into male and female singers. The experimental result is shown in Fig 6.

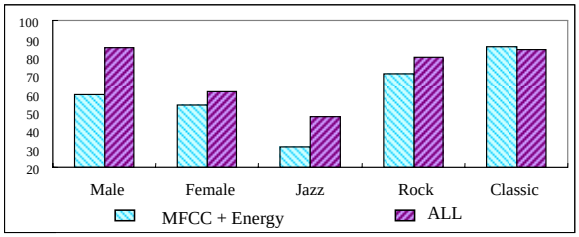

Fig. 6 Result of further classifying pop songs

It can be seen from Fig. 6, for further classification, the average accuracy can be improved by 11.3 % by using all the combined features than MFCC + Energy, and for male singers, the average accuracy can be improved by nearly 27%.

5 Conclusion

This paper proposed new instrumental features that can represent the information of instrument and instrument-based melody characteristics of unstructured music. Features are extracted by the decomposition of musical signal based on "Dictionary". The experimental results showed that our high-level features have a 4.5% higher average classification accuracy than that of MFCC with energy term. Furthermore, the average enhancements of 11.3% and nearly 27% can be achieved by combining our features with commonly used MFCC features with energy for further classification.

6 Acknowledgements

This work was supported in part by NSF of China under contracts 60003017 and 60373020, 863 Plans under contracts 2002AA103065, and Shanghai Municipal R&D Foundation under contracts 03DZ15019 and 03DZ14015, MoE R&D Foundation under contracts 104075.

7 References

Dannenberg, R.B. Thom B. and Watson D (1997). "A Machine Learning Approach to Musical Style Recognition". In *Proceedings of the International Computer Music Conference.*

Chai W. and Vercoe B (2001). "Folk Music Classification Using Hidden Markov Models". In *Proceedings of International Conference on Artificial Intelligence.*

Matityaho B. and Furst. M.(1995). "Neural Network Based Model for Classification of Music Type", In *Proceedings. of Convention of the Electrical and Electronic Engineers*, pp. 4.3.4/1-5, Israel.

Soltau.H., Schultz. T., Martin Westphal, and Alex Waibel (1998). "Recognition of Music Types". *International Conference on Acoustics, Speech, and Signal Processing*, Vol. II, pp. 1137-1140.

Pye. D. (2000). "Content-Based Methods for the Management of Digital Music" *International Conference on Acoustics, Speech, and Signal Processing*, Vol. IV, pp.2437-2440

Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Lian-Hong Cai, Jian-Hua Tao, (2002)."Music Type Classification By Spectral Contrast Features:" *In Proceedings. of IEEE International Conference on Multimedia and Expo*, Lausanne Switzerland,

Sieger N.J., Tewfik A.H, (1997) "Audio coding for conversion to MIDI." *IEEE First Workshop on Multimedia Signal Processing*, pp. 101-106.