AUDIO MOOD CLASSIFICATION USING RHYTHM AND BASS-LINE PATTERN INFORMATION

Emiru Tsunoo Taichi Akase Nobutaka Ono Shigeki Sagayama
Graduate School of Information Science and Technology University of Tokyo, Japan
{tsunoo, t-akase, onono, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper discusses an approach for the feature extraction for audio mood classification which is an important and tough problem in the field of music information retrieval (MIR). In this task the timbral information has been widly used, however many musical moods are characterized not only by timbral information but also by musical scale and temporal features such as rhythm patterns and bass-line patterns. In particular, modern music pieces mostly have certain fixed rhythm and bass-line patterns, and these patterns can characterize the impression of songs. We have proposed the extraction of rhythm and bass-line patterns, and these unit pattern analyses are combined with statistical feature extraction for mood classification. Experimental results show that the automatically calculated unit pattern information can be used to effectively classify musical mood.

1. INTRODUCTION

Due to the increasing size of music collections available on computers and portable music players, researches related to music information retrieval (MIR) including automatic genre classification and mood classification from audio has surged recently. Mood classification systems share the same formulation with genre classification systems which have been provided a established way of evaluating new representations of musical content and generally constructed of two stage processes: feature extraction and classification. These audio classification tasks have been competed in the contest called MIREX [1] for MIR researches for a long time. Mood classification is one of a variation of the audio classification however has even more elusive ground truth [2, 3]. In mood classification, not only instrumental information but also musical scale, rhythm and bass-line informations are thought to be important. In many case especially of modern popular music, some fixed bar-long percussive patterns and bass-line patterns are repeated in whole a song and the unit patterns frequently characterize the music. If such representative bar-long percussive patterns and bass-line patterns in music can be captured automatically as templates, they can potentially be used to characterize different music mood directly from audio signals.

In previous research, timbral features, rhythmic features and pitch features have been used for audio genre classification [4], and this sort of feature extraction is widely used in audio classification including mood classification in the past MIREX contests. In this work, the timbral features were the most dominant and the other statistical features came out not to be so useful for audio classification. This feature extraction method was tested for mood classification [2] and the effectivity was verified. The statistical features are also classified in various ways such as fuzzy classification [5] and active learning of support vector machine (SVM) [6]. In comparison to these statistical feature extraction, we have proposed temporal feature extractions including bar-long percussive pattern information [7] and bar-long bass-line pattern information [8], and these two features are tested to the genre classification task.

In this paper, we discuss an approach for extracting unit rhythm and bass-line patterns from a number of audio tracks and propose a feature vector for the application to mood classification. First, we separate percussive sounds and harmonic sounds of the audio tracks. Then we propose a clustering method specialized to rhythm patterns using a combination of dynamic programming and k-means clustering, and to bass-line patterns based on the k-means clustering algorithm. For the purpose of an application to audio mood classification, the scheme to extract feature vector based on clustered patterns which contain temporal information is suggested. All process is implemented using an open source software *Marsyas*[1] which is open source software with specific emphasis on Music Information Retrieval (MIR) [9].

2. HARMONIC/PERCUSSIVE SOUNDS SEPARATION

Generally, harmonic and percussive sounds are mixed in the observed spectrograms of audio pieces. Therefore in order to perform rhythm and bass-line pattern analysis it is useful to separate these components as a preprocessing, and percussive components are used for rhythm analysis and harmonic components for bass-line analysis. We utilize the harmonic/percussive sound separation (HPSS) technique proposed by Ono [10] that is based on the difference of general timbral features. By looking at the upper left figure in Fig. 1, a typical instance of spectrogram, one can observe that harmonic components tend to be continuous along the temporal axis in particular frequencies. On the other hand, percussive components tend to be continuous along the frequency axis and temporally short. Mask functions for separating the two components (harmonic and percussive) are calculated following a maximum a priori (MAP) estimation approach using the expectation maximization (EM) algorithm. Applying this approach to the shown spectrogram, harmonic and percussive components are separated (harmonic and percussive components are shown in the upper right and the lower left of Fig. 1 respectively).

3. BAR-LONG UNIT PATTERN CLUSTERING

3.1 Rhythm Pattern Clustering

Bar-long percussive patterns are frequently common and characteristic of a particular mood or style. Automatically

[1] http://marsyas.sness.net/

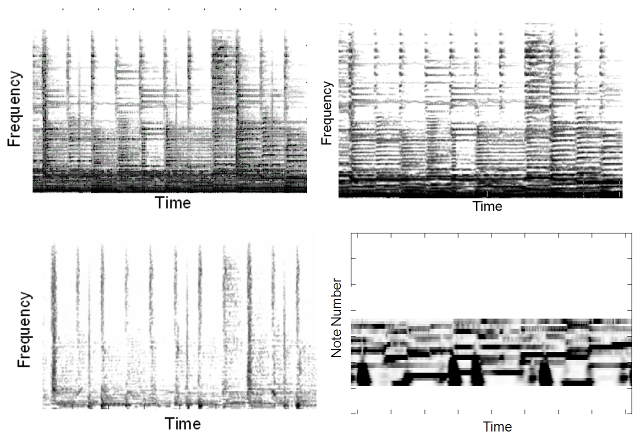

Figure 1. The original spectrogram (upper left), the harmonics-emphasized spectrogram (upper right) and the percussion-emphasized spectrogram (lower left) of a popular music piece (RWC-MDB-G-2001 No.6 [11]). The low-pass filtered logarithmic spectrogram calculated using wavelet transform from harmonics-emphasized spectrogram is shown in lower right figure.

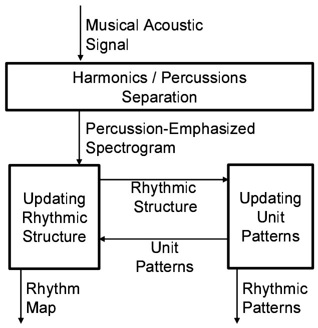

Figure 2. The flow diagram of the extraction of rhythm patterns.

detecting these patterns is a "chicken-and-egg" problem in that sets of bar-long unit rhythm patterns may be determined only after their corresponding unit boundaries in the music pieces are given, and vice versa. This is complicated by tempo fluctuations which might cause the unit pattern to stretch or shrink.

In order to solve these problems, the rhythm map which we have proposed [7, 12] is employed. The rhythm map is an approach to estimate representative bar-long percussive patterns and its segmentation (i.e. bar lines) simultaneously. This algorithm is processed to the percussive-emphasized spectrogram. By iterating dynamic programming (DP) matching and updating the templates used for DP matching based on the k-means-clustering-like update rules, both segmentation and templates themselves are updated. After convergence, the multiple percussive patterns in the input song are learned and the optimal segmentation is obtained. We use this estimated segmentation as bar lines. Fig. 3 illustrates the flow of this algorithm.

3.2 Bass-line Pattern Clustering

Bar-long bass-line patterns are also frequently common and characteristic of a particular mood. Bar lines have been already estimated in the process shown in section 3.1. Therefore, only problem to be solved is that unit bass-line patterns are shifted in pitch according to the chord played. For example, a pattern consists of only root notes in a uniform rhythm, all notes in this pattern need to be pitch-

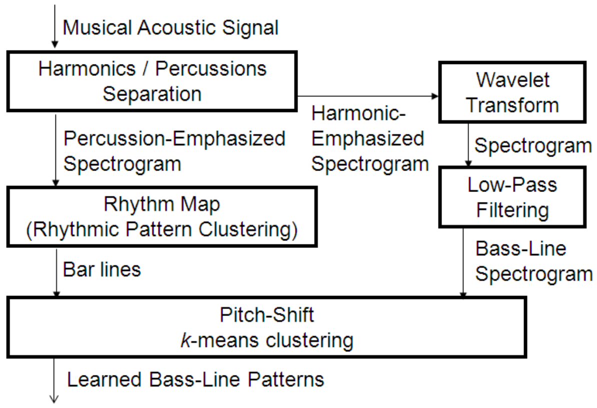

Figure 3. The flow diagram of the bass-line pattern extraction.

shifted by the same amount of notes according to the chord, because the root note changes accompany with the chord changes. Fig. 3 illustrates the flow of the bass-line pattern clustering algorithm.

If there was no pitch shift a simple k-means clustering approach can be used to estimate the representative bar-long bass-line patterns: Distances between each bar-long spectrogram pattern and centroid spectrogram patterns are calculated, and the centroids are updated by averaging the sample patterns. In order to deal with pitch-shift problem, we have proposed an approach where every possible pitch-shift is compared in k-means framework [8]. The lower right of Fig. 1 shows the logarithmic spectrogram which is processed low-pass filtering after Gabor wavelet transform whose frequency resolution is semitone (100 cents). The low-pass filtering (actually a band-pass filtering) was done by setting high and low frequency components to be zero. This kind of spectrogram is used for this purpose.

4. FEATURE EXTRACTION

4.1 Mood Classification via Pattern Clustering

Ideally percussive patterns for a particular mood would be fixed and would be automatically extracted perfectly. If that was the case then automatic mood classification could be performed simply by looking at which particular rhythm and bass-line patterns are used in a music piece. However in practice there is no guarantee that patterns are fixed for a particular mood or that their automatic extraction will be perfect. Therefore in many cases the percussive and bass-line patterns of a particular music piece will belong to more than one mood. To address these problems simultaneously we utilize a pattern occurrence histogram and distance related representation followed statistical machine learning to automatically classify music mood. Supervised learning classifiers such as Support Vector Machines (SVM) [13] can be used for this purpose.

4.2 Percussive Pattern Feature Extraction

One possible way to extract percussive feature vector is count up which percussive pattern templates are contained in a song and calculating the mood pattern occurrence histogram, similarly to Latent Semantic Indexing approach [14].

If K pattern templates are learned from mood m ($m = 1, \ldots, M$), an alignment can be calculated using dynamic programming to get the templates $T_{k,m}$ that exist in the song s. Then, the occurrence number of the patterns from mood m can be simply calculated by summation as follows:

$$c_{s,m} = \sum_{k=1}^{K} c_{s,k,m} \quad (1)$$

where $c_{s,k,m}$ is the number of the template $T_{k,m}$ in the song s, and the eventual M dimensional pattern occurrence histogram features vector $\boldsymbol{x}$ of song s can be written as

$$x_g = \frac{c_{s,m}}{N_s} \quad (2)$$

which is normalized by N_s, the number of measure in the song s.

4.3 Bass-line Pattern Feature Extraction

For bass-line pattern, one way to extract feature vector is calculating distances between every measure of input spectrogram and every template pattern and averaging them through whole an input piece. Even though there is a possibility for a music piece to belong to more than one mood templates, the distances between spectrogram in the input piece and learned templates are still affected.

The mathematical definition of the feature vector is following. After bass-line pattern templates are learned, we have $K \cdot M$ templates when K templates are learned from M moods. Then the distances between input song which have N measures and learned templates are calculated. The averaged distances are obtained as follows:

$$d_l = \frac{1}{N} \sum_{n=1}^{N} D(X_n, B_l) \quad (3)$$

where $1 \leq l \leq KM$ is the template number and $D(\cdot, \cdot)$ is a distance with pitch-shift problem proposed in [8]. The feature vector $\mathbf{x}$ can be written as

$$\mathbf{x} = (\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_{KM}})^{\mathbf{T}}. \quad (4)$$

4.4 Statistical Feature Extraction

4.4.1 Mel-Frequency Cepstral Coefficients

MFCCs are one of the most compact and efficient spectral expression that contain the general frequency characteristics important to human hearing. Originally MFCCs are developed for automatic speech recognition however they have been found to be powerful also for other auditory domains such as MIR [15].

In order to calculate MFCC feature, the log-magnitude spectrogram is wrapped to the Mel frequency scale and the inverse discrete cosine transform is performed. The first 13 coefficients are calculated in each time window of short time Fourier transform and the mean and the standard deviation for each song are calculated on the whole piece.

4.4.2 Musical Scale Feature

Musical scale also characterizes musical mood. For instance, when a note “do” was sounded the note “mi” makes the harmony bright, however “mi flat” makes it dark. In order to extract this kind of information, the correlation of chroma vector is calculated.

Chroma vector is a 12 dimensional vector whose columns represent overlapped energies of 12 semitones over octaves. Let the ith element of chroma vector at the time t be $c(t,i)$, a normalized correlation of the vector can be written as

$$S(t,\tau) = \sum_{i=1}^{12} \frac{\left(\frac{e(t)}{12} - c(t,i)\right)\left(\frac{e(t)}{12} - c(t,i+\tau)\right)}{e(t)^2} \quad (5)$$

where $e(t) = \sum_{i=1}^{12} c(t,i)$. It means just an energy in the case $\tau = 0$ and this correlation is symmetrical. Therefore, the coefficients which contain useful information are only 6 of them, and the mean and the standard deviation for each song are calculated on the whole piece.

5. CONCLUSIONS

We discussed an approach for clustering bar-long common percussive patterns and bass-line patterns for particular mood and extracting feature vectors for mood classification. We used HPSS technique to separate percussive components and harmonic components from audio signals as a preprocessing. Percussive patterns were clustered using a combination of one-pass DP and k-means clustering algorithm. Bass-line patterns were clustered using a new clustering method based on k-means clustering with pitch-shift. For audio mood classification, new feature vectors were defined as pattern occurrence histograms for percussive patterns and as averaged distances from each template for bass-line patterns. They were implemented in combination with statistical features including MFCCs and musical scale feature.

Future work includes using n-gram model approach rather than only looking at the uni-gram histogram for rhythm pattern features. Additionally, other features than pattern distance vector of bass-line pattern can be devised. Combination with other features like chord transition information can be done to improve further audio classification as well.

6. REFERENCES

[1] Mirex 2008. http://www.music-ir.org/mirex/2008/.

[2] T. Li and M. Ogihara. Detecting emotion in music. In *Proc. of ISMIR*, pages 239 – 240, 2003.

[3] X. Hu, M. Bay, and SJ Downie. Creating a simplified music mood classification ground-truth set. In *Proc. of ISMIR*, pages 309–310, 2007.

[4] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transaction on Speech and Audio Processing*, 10(5):293–302, 2002.

[5] Y.-H. Yang, C.-C. Liu, and H.-H. Chen. Music emotion classification: A fuzzy approach. In *Proc. of ACM Multimedia*, pages 81 – 84, 2006.

[6] M. Mandel, G. Poliner, and D. Ellis. Support vector machine active learning for music retrieval. *ACM Multimedia Systems*, 12(1):3 – 13, 2006.

[7] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama. Audio genre classification using percussive pattern clustering combined with timbral features. In *Proc. of ICME*, pages 382 – 385, 2009.

[8] E. Tsunoo, N. Ono, and S. Sagayama. Musical bass-line pattern clustering and its application to audio genre classification. In *Proc. of ISMIR*, 2009.

[9] G. Tzanetakis. *Marsyas-0.2: A Case Study in Implementing Music Information Retrieval System*, chapter 2, pages 31 – 49. Idea Group Reference, 2007. Shen, Shepherd, Cui, Liu (eds).

[10] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive componets in music signals. In *Proc. of the 9th Int. Conf. on Music Information Retrieval*, pages 139–144, September 2008.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Music genre database and musical instrument sound database. In *Proc. of the 4th Int. Conf. on Music Information Retrieval*, pages 229–230, October 2003.

[12] E. Tsunoo, N. Ono, and S. Sagayama. Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals. In *Proc. of ICASSP*, pages 185–188, 2009.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[14] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[15] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of ISMIR*, 2000.