MUSICAL BASS-LINE PATTERN CLUSTERING AND ITS APPLICATION TO AUDIO GENRE CLASSIFICATION

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

ABSTRACT

This paper discusses a new approach for clustering musical bass-line patterns representing particular genres and its application to audio genre classification. Many musical genres are characterized not only by timbral information but also by distinct representative bass-line patterns. So far this kind of temporal features have not so effectively been utilized. In particular, modern music songs mostly have certain fixed bar-long bass-line patterns per genre. For instance, while frequently bass-lines in rock music have constant pitch and a uniform rhythm, in jazz music there are many characteristic movements such as walking bass. We propose a representative bass-line pattern template extraction method based on k-means clustering handling a pitch-shift problem. After extracting the fundamental bass-line pattern templates for each genre, distances from each template are calculated and used as a feature vector for supervised learning. Experimental result shows that the automatically calculated bass-line pattern information can be used for genre classification effectively and improve upon current approaches based on timbral features.

1. INTRODUCTION

Due to the increasing size of music collections available on computers and portable music players, the need for music information retrieval (MIR) has surged recently. In particular, automatic genre classification from audio is a traditional topic of MIR and provides a structured way of evaluating new representations of musical content. In this task, not only instrumental information but also a bass-line information is thought to be important. For instance, bass parts in most rock songs consist of root notes of the chords in a uniform rhythm. In comparison to this, bass parts in jazz songs have a lot of characteristic movements which are called walking bass. If such representative bar-long bass-line patterns in music can be captured automatically as templates, they can potentially be used to characterize different music genres directly from audio signals.

In previous research, timbral features, rhythmic features and pitch features have been used for audio genre classification [1]. In this work, the timbral features were the most dominant and the pitch features used were not limited to the bass register. Studies more related to bass part extraction have been presented [2] where musical melodic and bass notes were modeled with Gaussian mixture model (GMM) and estimated. Using this pitch estimation method, Marolt [3] worked on the clustering of melodic lines using GMM. Researches using bass-line information for genre classification include McKay [4] and Tsuchihashi [5]. However the bass-line features discussed were based on overall statistics and did not represent directly temporal information.

In this paper, we discuss an approach for clustering unit bass-line patterns from a number of audio tracks and propose a feature vector based on the distances from templates for the applicationto genre classification. First, we emphasize only harmonic sounds of the audio tracks and estimate measure segments which divide tracks into measures. Then we propose a clustering method specialized to bass-line patterns based on the k-means clustering algorithm. For the purpose of an application to audio genre classification, the scheme to extract feature vector based on the bass-line patterns which contain temporal information is suggested. Finally, the effectiveness of the proposed bass-line pattern information for genre classification is verified experimentally.

2. BASS-LINE PATTERN CLUSTERING

2.1 Challenges in Bass-line Pattern Clustering

Bar-long bass-line patterns are frequently common and characteristic of a particular genre. In order to extract those representative patterns from audio signals, there are several challenges to be cleared. Especially in modern popular music, pieces comprise of both harmonic and percussive sounds and percussive components might disturb the bass-line analysis. Additionally, the bar lines of the music pieces need to be estimated. Another problem 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-shifted by the same amount of notes according to the chord, because the root note changes accompany with the chord changes.

Therefore the problems in extraction of representative bar-long patterns can be summarized as the following three problems:

I. audio signals may contain not only harmonic sounds but also percussive sounds,

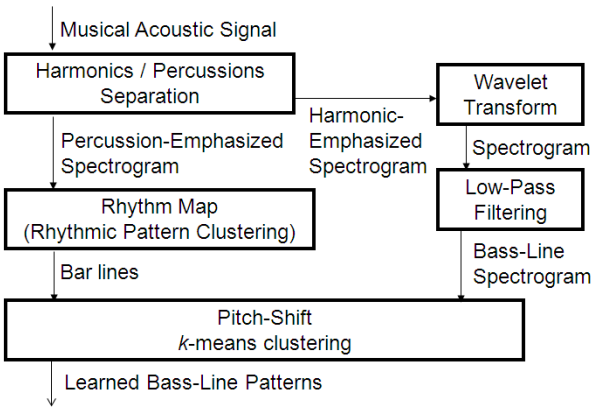

Figure 1. The flow diagram of the system.

II. measure segmentation is to be estimated, and
III. bass-line patterns are pitch-shifted according to chords.

In the next subsections, we describe our approach to solving these challenges. Fig. 1 illustrates the flow of the algorithm.

2.2 Emphasizing Harmonic Components

Generally, harmonic and percussive sounds are mixed in the observed spectrograms of audio pieces (the problem I). Therefore in order to perform bass-line pattern analysis it is useful to separate these components as a preprocessing step. We utilize the harmonic/percussive sound separation (HPSS) technique which we have proposed [6] that is based on the difference of general timbral features. By looking at the upper left figure in Fig. 2, 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 and harmonic ones are emphasized (harmonic and percussive components are shown in the upper right and the lower left of Fig. 2 respectively). In order to capture only the bass part we apply low pass filtering to the harmonic-only spectrogram.

2.3 Bar Line estimation Using Percussive Clustering Method

There are many possible ways to solve problem II which is the estimation of bar lines. One way is beat tracking [8] in which onset, chord changes and drum patterns are used as cues. Instead, in the case where it is unknown how many beats are in one measure or songs do not always start with the head of the measure, the pattern matching approach is rather useful to estimate bar lines. Therefore the rhythm map which we have proposed [9] 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 also requires the

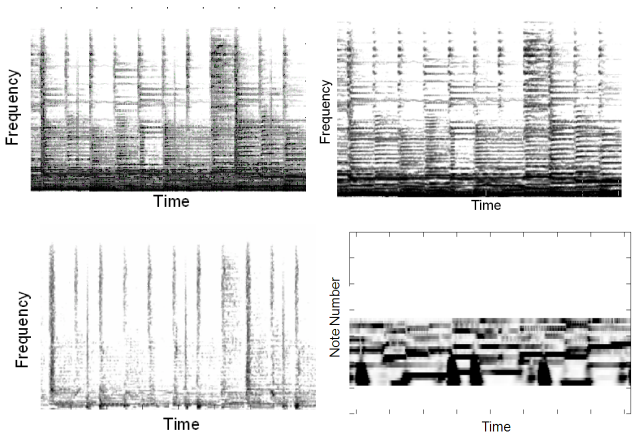

Figure 2. The original spectrogram (upper left), the harmonics-emphasized spectrogram (upper right) and the percussion-emphasized spectrogram (lower left) of a popular music piece (RWC-MDB-G-2001 No.6 [7]). The low-pass filtered logarithmic spectrogram calculated using wavelet transform from harmonics-emphasized spectrogram is shown in lower right figure.

source separation method shown in previous subsection as a preprocessing and deals with 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.

2.4 Iterative Update of Bass-line Pattern Cluster

If there was no pitch shift (problem III) 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 propose an new approach where every possible pitch-shift is compared in k-means framework.

Here we should note that both centroid template spectrogram and input spectrogram need to be logarithmic along frequency axis in order to consider pitch-shift. Because the musical notes are fashioned logarithmic in linear frequency domain. This kind of logarithmic spectrogram can be obtained by using wavelet transform. The lower right of Fig. 2 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. In this figure, the energy of bass drum is still dominant even after the source separation. That is because the bass drum sounds sometimes have long duration and the duration components also have the same feature with harmonic sounds which are continuous along the temporal axis. However these energies are thought not to be so harmful because when averaging spectrogram patterns to update templates, those parts become relatively small.

Mathematically, we modeled a bass-line template as a matrix and find out the most matched template for each bar-long spectrogram by calculating distances over all possible pitch-shift. The centroid bass-line pattern is defined as a matrix whose rows represent semitone numbers and whose columns are time instants. This matrix contains energy distribution of bar-long harmonic spectrogram. When the input piece has M measures, mth measure processed low-pass filtering can be written as X_m, the $I \times N$ matrix, where N is the number of semitones to capture in low frequency band and I is the resolution of time that divides a bar-long. On the other hand, kth bass-line pattern template can be written as the $I \times 2N$ matrix B_k. Since there are N notes to pitch-shift potentially, at least N rows of margin have to be provided.

As a distance measure of two bass-line pattern matrices, we use the following distance introduced by Frobenius norm:

$$d(X,Y) = \sum_{i=1}^{I}\sum_{n=1}^{N}\left(x_{i,n} - y_{i,n}\right)^2, \tag{1}$$

where $x_{i,n}$ and $y_{i,n}$ are the (i,j)th entries of matrices X and Y. Considering pitch-shift, we can define the distance between input spectrogram X_m and template bass-line pattern B_k as

$$D(X_m, B_k) = \min_{1\le n\le N} d(X_m, B_{k,n}) \tag{2}$$

where $B_{k,n}$ is a $I \times N$ submatrix of B_k which is from nth row to $(n+N-1)$th row.

Like k-means algorithm, for every input bass-line pattern X_m, the distances are calculated according to Eq. (2) and classes and pitch-shift intervals are determined as

$$\hat{k}_m = \operatorname*{argmin}_{1\le k\le K} D(X_m, B_k) \tag{3}$$

and

$$\hat{n}_m = \operatorname*{argmin}_{1\le n\le N} d(X_m, B_{\hat{k},n}). \tag{4}$$

Then the update rule of a template pattern can be written as following, just by averaging patterns in a particular class,

$$B'_k = \frac{\sum_{m\in\{m|k=\hat{k}_m\}} S^{\hat{n}_m} E X_m}{\sum_{m\in\{m|k=\hat{k}_m\}} 1}, \tag{5}$$

where E is an $N \times 2N$ extending matrix and S is a $2N \times 2N$ pitch-shift matrix respectively defined as

$$E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \tag{6}$$

$$S = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}. \tag{7}$$

This update rule satisfies following equation as far as Euclidean distance metric is used,

$$\begin{aligned} \sum_{m\in\{m|k=\hat{k}_m\}} D(X_m, B'_k) &\le \sum_{m\in\{m|k=\hat{k}_m\}} d(X_m, B'_{k,\hat{n}}) \\ &\le \sum_{m\in\{m|k=\hat{k}_m\}} D(X_m, B_{k,\hat{n}}), \end{aligned} \tag{8}$$

and convergence of iterative update is guaranteed. After the convergence, K centroid bass-line patterns B_k $(k = 1,\dots,K)$ are obtained.

3. BASS-LINE PATTERN FEATURE EXTRACTION

As the proposed clustering method applied to audio genre classification, there is a problem that the learned templates cannot be used directly. Ideally bass-line patterns for a particular genre would be fixed and would be automatically extracted perfectly. If that was the case then automatic genre classification could be performed simply by looking at which particular bass-line patterns are used in a music pieces by calculating the distance defined in Eq. (2). However in practice there is no guarantee that patterns are fixed for a particular genre or that their automatic extraction will be perfect. Therefore in many cases the bass-line patterns of a particular music piece will belong to more than one genre. To address these problems simultaneously we utilize a feature vector based on the distances from each template followed by statistical machine learning to automatically classify music genre. Supervised learning classifiers such as support vector machines (SVM) [10] can be used for this purpose.

One possible way to extract feature vector is calculating distances between every measure of input spectrogram and every template pattern following Eq. (2) and averaging them through whole an input piece. Even though there is a possibility for a music piece to belong to more than one genre templates, the distances between spectrogram in the input piece and learned templates are still affected, e.g., one measure spectrogram in blues song is close enough to the templates learned from blues collection even if its distance is not the smallest.

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

$$d_l = \frac{1}{M}\sum_{m=1}^{M} D(X_m, B_l) \tag{9}$$

where $1 \le l \le KG$ is the template number. The feature vector $\boldsymbol{x}$ can be written as

$$\boldsymbol{x} = (d_1, d_2, \dots, d_{KG})^T. \tag{10}$$

We use this feature vector for a supervised learning classification to classify music genre.

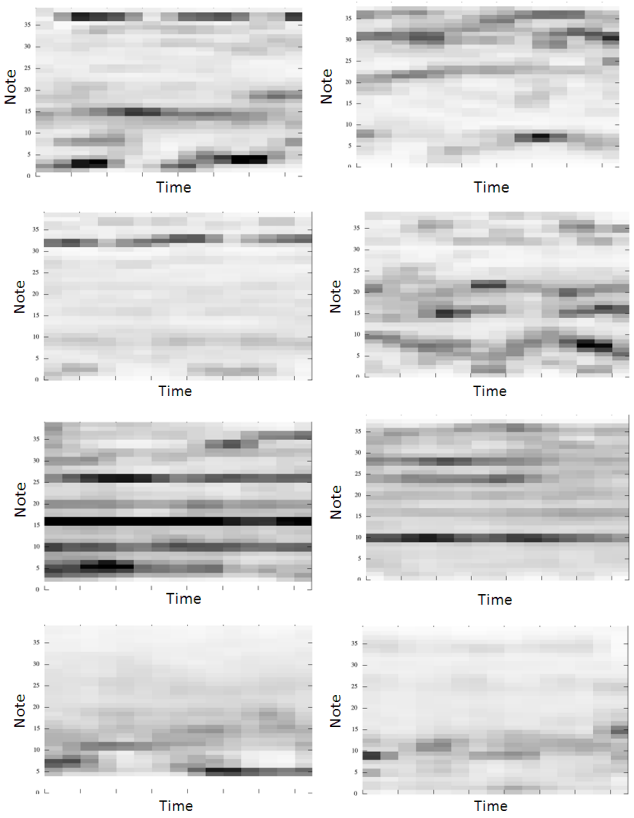

Figure 3. Two examples of learned bass-line templates from jazz, blues, rock, and hiphop, in descending order. While a lot of movements of bass notes are shown in jazz and blues, bass notes in rock are unchanged, and hiphop bass-line templates are quite sparse.

4. PROCEDURAL SUMMARY OF THE ALGORITHM

The overall algorithm can be summarized as follows:

1. Preprocessing
   (a) Emphasis of harmonic components using HPSS
   (b) Apply low-pass filtering
   (c) Estimate bar lines using rhythm map
2. Bass-line Pattern Clustering
   (a) Provide initial (random value) templates
   (b) Match the templates patterns with pitch-shifting by Eq. (2)
   (c) Update the template patterns following Eq. (5)
   (d) Iterate steps b and c until the convergence
3. Genre Classification
   (a) Calculate the distances between patterns (Eq. (10)) and use it as a feature vector characterizing a music pieces
   (b) Perform classification into genres using a machine learning technique

5. EXPERIMENTAL RESULTS

5.1 Dataset

Experiments with the proposed algorithms were conducted on the GTZAN dataset [1]. The dataset had 10 genres:

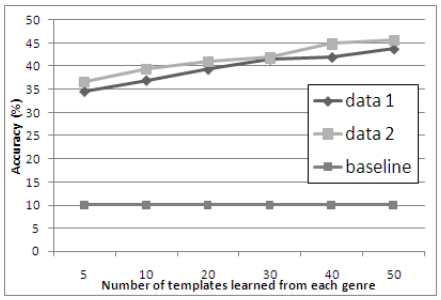

Figure 4. Classification accuracy using only bass-line features for each number of templates learned from one genre. The baseline accuracy of random classifier was 10.0%

blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. The dataset had 100 songs per genre all of which were single-channel and sampled at 22.05kHz. All songs were processed harmonic-percussive sound source separation and wavelet transform. Then they were processed low-pass filtering and obtained the spectrogram only from 82.4Hz (E1) to 330.0Hz (E3). The reason we didn't use the spectrogram under 82.4Hz was the dominance of the energy of bass drums in that area.

5.2 Template Learning and Feature Extraction

First, common bass-line pattern templates were learned using the proposed algorithm for each genre. The proposed algorithm was implemented using audio processing framework *Marsyas*[1] which is open source software with specific emphasis on Music Information Retrieval (MIR) [11]. To generalize the learning templates part, we divided the dataset 50-50 into two parts randomly and obtained two sets of templates for each genre. In this experiment, the number of iteration was fixed to 15 times because it was enough to converge, and the number of the time resolution was fixed to 16.

The examples of learned templates of jazz, blues, rock and hiphop are shown in Fig. 3. While jazz bass patterns had some movements, rock patterns showed the horizontally straight lines. In blues bass-line templates, a typical swing rhythm is shown, and in hiphop templates, there are sparse bass notes because hiphop songs mostly have strong drum sounds but bass sounds are not so melodic.

After learning templates two sets of templates were obtained, and next, we extracted feature vector (Eq. (10)) using those template sets. The templates learned from data 1 were used to extract feature vectors of data 2, and vice versa.

5.3 Classification results

In order to train a classifier in the feature space, the "Weka" machine learning toolkit [12] was employed. All the results shown were based on 10-fold cross validation using a linear SVM as a classifier. The labeled data was split into 10 folds and each fold was used once for testing with the remaining 9 folds used for training the classifier to generalize the classification, for both of divided data sets.

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

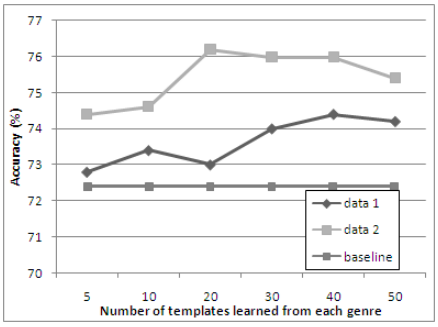

Figure 5. Classification accuracy using both bass-line and timbral features for each number of templates learned from one genre. The baseline accuracy of existing features was 72.4%.

Table 1. Genre classification accuracy using only bass-line pattern features and merged with timbral features.

Features	data 1	data 2
Baseline (random classifier)	10.0%	10.0%
Only bass-line (400 dim.)	42.0%	44.8%
Existing (Timbre, 68 dim.)	72.4%	72.4%
Merged (468 dim.)	74.4%	76.0%

The number of templates learned for each particular genre were decided experimentally. Fig. 4 shows the results using only the bass-line pattern features whose dimension was $10K$ when the number of templates learned from one genre was K. As can be seen the proposed features had enough information for genre classification and the accuracy improved as the number of templates was increased.

An existing state-of-the-art genre classification system which uses 68 dimensional timbral features like MFCC and spectrum centroids proposed by Tzanetakis was used for comparison. The system performed well on several audio classification tasks in MIREX 2008 [13]. Merging this timbral features and bass-line features, the classification accuracies shown in Fig. 5 were obtained. When the number of templates K was 40 the performance was the best, and when we increased the number K more than that the performance got worse. That was because the dimension of feature space became larger and curse of dimensionality was thought to occur. In particular case the number of templates K was fixed to 40 for each particular genre, the precise result is shown in Table 1 and the confusion matrices are shown in Table 2 and Table 3. These results were higher than existing genre classification systems that rely on timbral information and verify the effectiveness of the proposed bass-line patterns. One can see that classical was the most distinguished by the system and some reggae songs were mistaken for disco.

6. CONCLUSIONS

We discussed an approach for clustering common bar-long bass-line patterns for particular genres, and proposed a feature vector which represented relations to learned bass-line templates. We used HPSS technique to extract harmonic components from audio signals and rhythm map to estimate measure segmentation. After processing low-pass filtering, bass-line patterns were clustered using a new clustering method based on k-means clustering with pitch-shift. For audio genre classification, a new feature vector was defined as averaged distances from each template. Experiments over music pieces from various genres confirmed that the proposed algorithm can improve the accuracy of classification systems based on timbral information.

Future work includes more experiments with the parameters of the algorithms such as the resolution of time in templates. Additionally, other features than pattern distance vector can be devised. Combination with other features like percussive pattern can be done to improve further genre classification as well.

ACKNOWLEDGMENTS

The authors thank George Tzanetakis for sharing his dataset and some useful experimental discussions with us.

7. REFERENCES

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

[2] M. Goto. A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[3] M. Marolt. Gaussian mixture models for extraction of melodic lines from audio recordings. In *Proc. of ISMIR*, pages 80–83, 2004.

[4] C. McKay and I. Fujinaga. Automatic genre classification using large high level musical feature sets. In *Proc. of ISMIR*, pages 525–530, 2004.

[5] Y. Tsuchihashi and H. Katayose. Music genre classification from bass-part using som. In *IPSJ SIG Technical Reports*, volume 90, pages 31–36, 2006.

[6] 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.

[7] 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.

[8] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171, June 2001.

[9] 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.

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

[11] 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).

[12] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

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

Table 2. The confusion matrix of data 1 in the case 40 templates learned from each genre

classified as	bl	cl	co	di	hi	ja	me	po	re	ro
bl: blues	38	0	5	0	0	2	1	0	1	3
cl: classical	0	48	0	0	0	0	0	0	0	2
co: country	3	0	30	1	1	2	1	3	0	9
di: disco	0	0	1	36	1	1	1	2	2	6
hi: hiphop	0	0	1	6	34	0	1	2	4	2
ja: jazz	4	2	1	1	0	41	0	0	0	1
me: metal	0	0	2	1	0	0	43	0	0	4
po: pop	1	0	3	3	3	2	0	34	0	4
re: reggae	0	0	4	4	2	0	0	0	37	3
ro: rock	3	0	4	3	2	1	3	0	3	31

Table 3. The confusion matrix of data 2 in the case 40 templates learned from each genre

classified as	bl	cl	co	di	hi	ja	me	po	re	ro
bl: blues	38	0	5	2	0	2	2	0	1	0
cl: classical	0	47	0	0	0	3	0	0	0	0
co: country	4	0	40	0	0	0	0	0	0	6
di: disco	3	0	0	34	1	0	1	2	3	6
hi: hiphop	1	0	0	3	41	0	0	0	5	0
ja: jazz	3	1	0	0	0	43	2	0	0	1
me: metal	1	0	0	0	0	1	44	0	1	3
po: pop	4	0	1	2	1	0	0	35	3	4
re: reggae	4	0	0	9	4	2	0	3	27	1
ro: rock	6	0	3	5	0	0	3	2	0	31