Multimed Tools Appl (2016) 75:3013–3026
DOI 10.1007/s11042-014-2418-z

Importance of audio feature reduction in automatic music genre classification

Babu Kaji Baniya · Joonwhoan Lee

Received: 17 June 2014 / Revised: 18 November 2014 / Accepted: 24 November 2014 /
Published online: 23 December 2014

Abstract Multimedia database retrieval is rapidly growing and its popularity in online retrieval systems is consequently increasing. Large datasets are major challenges for searching, retrieving, and organizing the music content. Therefore, a robust automatic music-genre classification method is needed for organizing this music data into different classes according to specific viable information. Two fundamental components are to be considered for genre classification: audio feature extraction and classifier design. In this paper, we propose diverse audio features to precisely characterize the music content. The feature sets belong to four groups: dynamic, rhythmic, spectral, and harmonic. From the features, five statistical parameters are considered as representatives, including the fourth-order central moments of each feature as well as covariance components. Ultimately, insignificant representative parameters are controlled by minimum redundancy and maximum relevance. This algorithm calculates the score level of all feature attributes and orders them. Only high-score audio features are considered for genre classification. Moreover, we can recognize those audio features and distinguish which of the different statistical parameters derived from them are important for genre classification. Among them, mel frequency cepstral coefficient statistical parameters, such as covariance components and variance, are more frequently selected than the feature attributes of other groups. This approach does not transform the original features as do principal component analysis and linear discriminant analysis. In addition, other feature reduction methodologies, such as locality-preserving projection and non-negative matrix factorization are considered. The performance of the proposed system is measured based on the reduced features from the feature pool using different feature reduction techniques. The results indicate that the overall classification is competitive with existing state-of-the-art frame-based methodologies.

Keywords Music genres · Dimensionality · Locality preserving projection · Non-negative matrix factorization

B. K. Baniya · J. Lee (✉)
Department of Computer Science and Engineering, Chonbuk National University, 561-756 Jeonju, South Korea
e-mail: chlee@jbnu.ac.kr

B. K. Baniya
e-mail: everwith_7@jbnu.ac.kr

1 Introduction

Music genres are categorical labels created by musicians or composers for identifying the content (style) of the music. Music genres area growing research area in the field of information retrieval because they can be applied to practical purposes, such as efficient organization and categorization of online data collections. It is therefore essential to design automatic grouping tools that provide meaningful and efficient ways of describing music genre classifications. Several well-known approaches have been proposed in this area. Efficient and accurate automatic music information processing remains the key issue; therefore, it has been attracting the attention of researchers and musicologists. Consequently, improvement is still needed in automatic genre classification. The main concern is in describing, organizing, and categorizing music content on the Internet. This can be achieved by finding the important characteristics of audio signals. Music audio signals belong to the same genre; that is, they share certain characteristics because they are composed of similar types of instruments that use similar rhythmic patterns and pitch distributions [30]. The extracted music features must be comprehensive, compact, and effective [8].

Audio feature selection is generally based on areas of application. Jialie et al. [24] used four groups of features (timbral, spectral, rhythmic, and melodic) for music tagging. Later, they used five different groups of features for large-scale music searching [23]. Features commonly exploited for music genre classification can be roughly classified into timbral texture, rhythmic, pitch content, or their combinations [27]. Once descriptive features are extracted, different pattern recognition algorithms are employed for their classification into genres. These features can be categorized into two types: low-level and high-level. Low-level features are derived from a short segment, such as a frame. In contrast, high-term features usually characterize the variation of spectral shape or beat information within a long segment. The low and high segments are also referred to as the analysis window and texture window, respectively [6, 29]. In short time estimates, the signal is partitioned into successive frames using small-sized windows. If an appropriate window size is chosen, the signal within each frame can be considered a stationary signal. The windowed signal is then transformed into another representation space to achieve good discrimination.

Music genre classification with reduced feature sets using the support vector machine (SVM) classifier is shown in Fig. 1. It represents an overview of the proposed method of genre classification. Basically, there are two problems to be addressed in music genre classification: audio feature extraction and classifier design. In addition, feature analysis plays a significant role in genre classification. Feature analysis means identifying the most discriminative feature or set of features among all extracted features. In this scheme, the minimum redundancy maximum relevance (MRMR) approach is implemented for feature reduction. It gives the maximum relevance value as a score of each feature statistic in descending order. Based on this requirement, we can select the number of required feature statistics to achieve maximum classification accuracy. Furthermore, the unique aspect of MRMR is to maintain the original features as they are. Other important feature reduction methodologies include PCA [25], LDA [28], LPP [10], and NMF [14], unsupervised feature selection [12] etc. Later, we perform the classification based on reduced features from the feature pool to obtain optimum classification accuracy.

In this study, we implemented four groups of audio features: dynamic, rhythmic, spectral, and harmonic. They belong to either low- or high-level feature categories according to frame size. The frame length for low- and high-level features were 46 ms and 2 s, respectively, with a 50 % overlap for both. From each feature, we extracted up to the fourth order of central moments for music, such as mean, variance, skewness, and kurtosis. It was recently reported

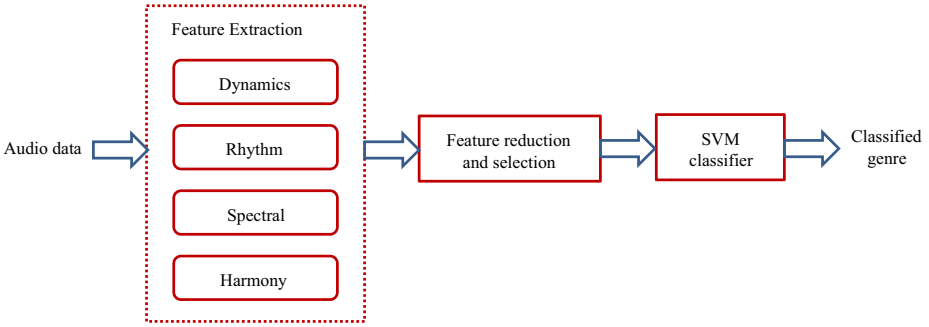

Fig. 1 Overview of proposed method

that the higher-order moments, such as skewness and kurtosis [1], are also useful for music genre classification [2]; therefore, we included them in our genre classification. In addition, covariance components of pairwise features were likewise included within the same group of features. For feature extraction, we utilized the MIRtoolbox [11].

From the extracted features, we calculated five statistical parameters: mean, standard deviation, skewness, kurtosis, and covariance components. Ultimately, feature dimensions sharply increased. However, it was not known whether all the features were equally significant for genre classification in order to be evaluated by MRMR. For this purpose, we employed GTZAN, a well-known dataset collected by Tzanetakis [17]. It includes ten different classes; each class has 100 songs. The statistical parameters, such as covariance, were calculated for each feature group. In the spectral category, the covariance component—such as the mel frequency cepstral coefficient (MFCC) [21], delta mel frequency cepstral coefficient (DMFCC), and delta-delta mel frequency cepstral coefficient (DDMFCC) were separately calculated.

The experimental results demonstrated that the high-order moment helps to increase classification accuracy when it is combined with other low-level spectral features; it generally provides the supplementary statistical information for the audio signal. Skewness is a measure of the asymmetry of the data distribution regarding the sample mean, which represents the relative disposition of the tonal and non-tonal components of the audio signal. Kurtosis is the measure of the degree of peakedness or flatness of a distribution [9]. Therefore, we considered $4n$ components for the n texture features.

The remainder of this paper is organized as follows. Feature extraction, which is the critical portion of genre classification, is described in Section II. In Section III, we discuss audio feature reduction. In Section IV, we explain the experimental setup and data preparation, and the results and analysis are provided in Section V. In Section VI, we present our conclusions of music genre classification.

2 Feature extraction

Feature extraction involves the analysis and extraction of meaningful information from audio files to obtain a compact and concise description that is machine-readable. Features are typically selected in the context of a specific task and domain. The features used in our research are divided into four categories: dynamic, rhythmic, spectral, and harmonic.

The different frame-based features listed in Table 1 are extracted. We can integrate each of them using the different statistical values, such as mean, standard deviation, skewness,

Table 1 Extracted audio feature sets (*M* mean, S_{td} standard deviation, S_k skewness, K_t kurtosis, *Cov* pairwise covariance, *HCDF* Harmonic Change Detection Function)

No.	Category	Feature	Acronyms
1	Dynamic	RMS energy	M, S_{td}, S_k K_t, Cov
2		Slope	”
3		Attack	”
4	Rhythm	Tempo	M, S_{td}, S_k K_t
5		Spectral centroid	
6		Brightness	”
7		Spread	”
8		Rolloff85	”
9		Rolloff95	”
10		Spectral entropy	”
11		Flatness	”
12		Irregularity	”
13		Roughness	”
14		Zerocrossing	”
15		Spectral flux	”
16		MFCC (1 ~ 13)	”
17		DMFCC (1 ~ 13)	”
18	Harmony	DDMFCC (1 ~ 13)	”
19		Chromagram peak	”
20		Chromagram centroid	”
21		Key clarity	”
22		Key node	”
23		HCDF	”

kurtosis, and covariance in a music piece. The mean (μ) and standard deviation (σ) for frame-wise feature values (X_n) in an*N*-frame music piece are given by

$$Mean\,(\mu) = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{1}$$

$$Std(\sigma) = \frac{1}{N}\sum_{n=1}^{N} (X_n - \mu)^2 \tag{2}$$

Skewness is a measure of the asymmetry of the distribution, which represents the relative disposition of the tonal and non-tonal components of each band. If the tonal components frequently occur in a band, the distribution of its spectrum will be left-skewed; otherwise, it will be right-skewed. Mathematically, the skewness in a music piece can be defined as

$$Skewness = \frac{\sum_{n=1}^{N} (X_n - \mu)^3}{(N-1)\sigma^3} \tag{3}$$

Kurtosis is the measure of whether the data are peaked or flat in relation to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean.

The kurtosis measure can sketch the effective dynamic range of the audio spectrum. It can be mathematically defined as

$$Kurtosis = \frac{\sum_{n=1}^{N} (X_n - \mu)^4}{(N-1)\sigma^4} - 3 \tag{4}$$

Covariance is measured between two random variables or features. Covariance is typically considered to determine if there is any relationship between the random variables. It is useful to measure the polarity and degree of the correlation between two features. The covariance of two features X_n and Y_n, in a music piece is given as

$$Cov(X_n, Y_n) = \frac{1}{N} \sum_{n=1}^{N} (X_n - \mu_X)(Y_n - \mu_Y) \tag{5}$$

where μ_X and μ_Y are corresponding means of X_n and Y_n, respectively. For n timbral texture features, we acquire $n(n-1)/2$ mutual covariance values.

3 Feature reduction

3.1 Minimum redundancy maximum relevance (MRMR)

The MRMR criterion was proposed in [19], [18] in combination with a forward selection search strategy. Given set X_s of selected variables, the criterion consists of updating X_s with the variable $X_i \in X_t$ that maximizes u_i-z_i, where u_i is a relevance term and z_i is a redundancy term. Moreover, I is the mutual information of two variables, u_i is the relevance of X_i to the output Y alone, and z_i is the average redundancy of X_i to the selected variables $X_j \in X_s$.

$$u_i = I(X_i; Y) \tag{6}$$

$$z_i = \frac{1}{d} \sum_{X_j \in X_s} I(X_i; Y) \tag{7}$$

$$X_i^{MRMR} = \arg \max_{X_i \in X_t} \{u_i - z_i\} \tag{8}$$

At each step, this method selects the variable with the best trade-off between relevance and redundancy. The selection criterion is fast and efficient. At step d of the forward search, the search algorithm computes n-d evaluations; the evaluation requires the estimation of $(d+1)$ bivariate densities (one for each of the already selected variables plus one with the output). It was shown in [18] that the MRMR criterion is an optimal first-order approximation of the mix-dependency if a feature is selected at the given time. Furthermore, MRMR avoids the estimation of multivariate densities by using multiple bivariate densities. Although the method addresses the bivariate redundancy issue through the term z_i, it cannot account for the complementarities between variables.

3.2 Principal component analysis (PCA)

PCA is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of variables called principal components. The

number of principal components is less than or equal to the number of original variables. This transformation is defined such that the principal component has the largest possible variance; moreover, each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

PCA is referred to as a discrete version of the Karhunen-Loeve transform. For a set of M-dimensional $X=[x_1,x_2,\ldots,x_M]^T$, and $[w_1,w_2,\ldots,w_R,\ldots,w_M]$ be the eigenvalues in descending order and the corresponding orthonormal eigenvectors of $E[XX^T]$. To reduce the M-dimensional data to R-dimensional space, the reduced number of eigenvectors $W=[w_1,w_2,\ldots,w_R]$ is applied as given below

$$Y = W^T X \tag{9}$$

This transformation is defined such that the principal component has the largest possible variance; additionally, each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. We considered the transform feature up to a variance range of 0.98; it covered up to 200 transform features.

3.3 Locality preserving projection (LPP)

In this paper, we also consider LPP, a linear dimensionality reduction algorithm [3]. It builds a graph that incorporates neighborhood information of the data set. Based on the notation of the Laplacian of the graph, we then compute a transformation matrix that maps the data points to a subspace. This linear transformation optimally preserves local neighborhood information in a certain sense. The representation map generated by the algorithm may be viewed as a linear discrete approximation to a continuous map that naturally arises from the geometry of the manifold [20].

The generic problem of linear dimensionality reduction is the following. Given set x_1, $x_2,\ldots,x_m$ in R^n, find transformation matrix A that maps these m points to a set of points y_1, y_2, $\ldots,y_m$ in R^l ($l \ll m$), such that y_i "represents" x_i, where $y_i=Ax_i$. This method is of particular applicability in the special case where $x_1, x_2,\ldots,x_m\in$ and is a manifold embedded in R^n.

Locality preserving projection is a linear approximation of the nonlinear Laplacian eigenmap [7]. The algorithm procedure is outlined below.

1. Constructing the adjacency graph: Let G denote a graph with m nodes. An edge exists between nodes i and j if x_i and x_j are "close". Two variations include:

   a. ϵ-neighborhoods [parameter $\epsilon \in$ R]: Nodes i and j are connected by an edge if $\|x_i\text{-}x_j\|^2 < \epsilon$, where the norm is the usual Euclidean norm in R^n.
   b. k-nearest neighbors [parameter $k\in N$]:Nodes i and j are connected by an edge if i is among k-nearest neighbors of j or if j is among k-nearest neighbors of i.
2. Two variations exist for weighting the edges. W is a symmetric $m\times$m matrix with W_{ij} having the weight of the edge-joining vertices i and j, or 0 if there is no such edge.

   a. Heat kernel [parameter t$\in$ R]:If nodes i and j are connected, put

$$w_{ij} = e^{\frac{\|xi-xj\|^2}{t}} \tag{10}$$

   The justification for this choice of weights can be traced to [8].
   b. Simple-minded: $W_{ij}=1$ if and only if vertices i and j are connected by an edge.

3. Eigen maps: Compute the eigenvectors and eigen values for the generalized eigenvector problem:

$$XLX^T a = \lambda XDX^T a \tag{11}$$

where D is a diagonal matrix whose entries are column sums of W, $D_{ii} = \sum_j w_{ji}$. L=D-W is the Laplacian matrix. The i^{th} column of matrix X is x_i. Let the column vectors $a_0,\ldots, a_{l-1}$ be the solutions of Eq. (11), which are ordered according to their eigenvalues, $\lambda_0<\ldots.<\lambda_{l-1}$. Thus, the embedding is as follows:

$$x_i \rightarrow y_i = A^T x_i, A = (a_0, a_1, \ldots, a_{l-1}), \text{ where } y_i \text{ is an } l\text{−dimensional vector, and } A \text{ is a } n \times l \text{ matrix.}$$

3.4 Linear discriminant analysis (LDA)

LDA or Fisher's linear discriminant (FLD) approach is a widely used method for feature extraction in audio and image files. It is a dimensional reduction technique used for classification problems. This approach is used to find the projection direction in which audio belonging to different classes is maximally separated. Mathematically, it attempts to find the projection matrix so that the ratio of the between-class and within-class scatter matrices of projected features is maximized.

The optimal discrimination projection matrix W_{opt} is given as

$$W_{opt} = \underset{w}{\operatorname{argmax}} \frac{\left|W^T S_B W\right|}{\left|W^T S_w W\right|} \tag{12}$$

The basic procedures of LDA are as follows. Calculate the scatter matrix of within-class S_w as

$$S_w = \sum_{i=1} \left(x_i - \mu_{k_i}\right)\left(x_i - \mu_{k_i}\right)^T \tag{13}$$

Calculate the scatter matrix of between-class S_B as

$$S_B = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^T \tag{14}$$

Calculate the eigenvectors of the projection matrix

$$w = eig\left(S_T^{-1} S_B\right) \tag{15}$$

S_B is the between-class scatter matrix, S_W is the within-class scatter matrix, and S_T=S_B+S_W is the total scatter matrix. Notation c is the total number of samples in the whole audio data set, x_i is the feature vector of a sample, and μ_{k_i} , is the vector of the genre class to which x_i belongs. In addition, μ_i is the mean feature vector of class i, and n_i is the number of samples in genre class i.

3.5 Non-negative matrix factorization (NMF)

NMF is an algorithm that is used to decompose any non-negative matrix V into non-negative-based vectors matrix B and non-negative coefficient matrix W. The real value V of size K-by-T produces the two matrices, B and W:

$$V \approx BW \tag{16}$$

B is size K-by-I and W is size I-by-T, with I being a user-defined parameter. The approximation can be performed by minimization of the Euclidean distance and Kullback–Leibler divergence. The Euclidean distance between X and BW is given as follows:

$$\min_{B,W} \left(\|V - BW\|_2^2 \right) \tag{17}$$

where

$$\|V - BW\|_2^2 = \sum_{k,t} \left(V_{k,t} - (BW)_{k,t} \right)^2$$

The second cost function is the divergence of V from BW, which yields the following optimization problem:

$$\min_{B,W} D\left(V \middle|\middle| BW \right) \tag{18}$$

where

$$D(V\|BW) = \sum_{k,t} \left(V_{k,t} \log \frac{V_{k,t}}{(BW)_{k,t}} - V_{k,t} + (BW)_{k,t} \right)$$

After matrices B and W are initialized with the absolute values of Gaussian noise, the multiplicative update rules for Eq. (17) are as follows:

$$B \leftarrow B. \times \frac{VW^T}{BWW^T} \tag{19}$$

$$W \leftarrow W. \times \frac{B^T V}{BB^T W} \tag{20}$$

Equation (18) is minimized by:

$$B \leftarrow B. \times \frac{\frac{V}{BW} W^T}{1W^T} \tag{21}$$

$$W \leftarrow W. \times \frac{B^T \frac{V}{BW}}{B^T 1} \tag{22}$$

where 1 corresponds to a K-by-T matrix (the same size as V) containing only one.

4 Experimental setup and data preparation

We employed for the performance evaluation the renowned GTZAN dataset, which is widely used for music genre classification. The dataset consists of 1000 music pieces divided into ten genres: classical, blues, hiphop, pop, rock, jazz, reggae, metal, disco, and country. Each class consists of 100 music pieces; each piece is 30s in duration. Each music piece is stored in the database as a 22,050Hz, 16bits, and mono-audio file. A ten-fold cross validation scheme was used to evaluate the performance of the proposed system using the GTZAN dataset.

5 Result and analysis

The proposed method improved music classification accuracy with minimum feature sets and maximum discriminative capability. We analyzed the audio features (statistics) in two stages: low-order statistics (mean and standard deviation) and high-order statistics. The experiment included118 features in total when considering the low-order statistics. The objective was to reduce feature statistics by using PCA, LDA, LPP, NMF, and MRMR without degrading the classification accuracy and their comparison. A ten-fold cross validation scheme was used to evaluate the performance of the proposed system in the GTZAN dataset using SVM [7]. The classification accuracy of 80.75 % was achieved using two different statistics of each feature from the reduced feature sets. The feature reduction was performed using the MRMR algorithm. Similarly, 75.0, 73.50, 77.25, and 79.80 % classification accuracies were obtained using, respectively, PCA, NMF, LDA, and LPP-based feature reduction methodologies. Of the five feature reduction techniques, MRMR, LPP, and LDA were comparatively better than the others.

We later included all statistics (lower-order moments, higher-order moments, and covariance components) extracted from the music genre dataset with a total feature dimension of 538. The results showed that the feature dimension sharply increased over the previous one (only for low-order statistics). It was unknown whether all the feature statistics were equally significant for music genre classification. We aimed to find the most discriminative feature statistics from the feature pool. Therefore, several feature reduction approaches we reused to reduce the insignificant feature statistics from the feature pool without degrading the overall performances, which are outlined in Table 2.

The number of reduced feature statistics obtained in multiples of 25 showed no significant differences, nor did their corresponding classification accuracies, which were measured from the previous to current accuracies. Among them, the MRMR-based feature reduction method

Table 2 Feature reduction methodologies percentage of reduced features and their corresponding classification accuracies

S.No.	Feature reduction method	% of remaining reduced feature	% of classification accuracy
1	PCA	32.53 %	78.25 %
2	NMF	27.88 %	75.20 %
3	LDA	23.23 %	84.50 %
4	MRMR	37.17 %	87.90 %
5	LPP	23.23 %	85.20 %

demonstrated better classification accuracy (87.09 %) than that of the others when only 37.17 % of feature statistics were considered. Similarly, LPP and LDA showed competitive classification accuracies when fewer dimensions (23.23 %) than those of MRMR were considered. PCA showed an average classification accuracy compared to LPP and LDA. Finally, it was evident that 37.17 % of feature statistics were significant and sufficient for achieving the optimum classification accuracy. This means that 62.83 % of features (statistics) were reduced; i.e., they were insignificant to genre classification. Ultimately, the computational complexity of the classifier was also reduced. Furthermore, we obtained the visible impact of feature reduction and their consequences in genre classification by using MRMR, LPP, LDA, PCA, and NMF.

Table 3 provides an overview of the feature statistics involved in genre classification. Among the 538 features in our study, only 37.17 % of important feature statistics were selected using the MRMR algorithm. From the experiment, it was found that the spectral features had a maximum contribution compared to the other feature groups. MFCC feature statistics, such as variance and covariance components, provided an especially significant contribution. They contributed 57 % out of 200 selected feature statistics. This is because MFCC feature statistics capture audio features in different frequency ranges that differ from those of other features. The spectral feature without MFCC remained the second-highest contributor for genre classification. The following feature attributes were selected: covariance component, variance, and mean; their corresponding contribution rate was 27 % in total. Harmonic features also played a significant role for genre classification; its statistical parameters were almost equally selected and their overall contribution rate was 8.5 %. Similarly, dynamic and rhythmic features remained in the least-selected feature groups. The order of selected feature statistics was: covariance component, variance, mean, skewness, and kurtosis; their corresponding proportions were 54.0, 36.5, 10, 6.0, and 2.5 %, respectively.

Table 4 shows the results of the comparison of our proposed method with other approaches in term of average classification accuracy while using the GTZAN dataset. Shin-Chelo et al. [11] extracted timbral features, such as MFCC coefficients, and used the spectro-temporal features to capture the temporal evolution and variation of the spectral characteristics of the music signal. They calculated the four different types of statistics from the extracted features; i.e., mean, variance, minimum, and maximum. Later, the insignificant statistical parameters were controlled using the SVM ranker. Consequently, the kernel SVM was used as the classifier for the music genre classification. Our method included differed from that of Shin-

Table 3 Proportions of selected feature statistics from each feature group

Type of feature	Feature from each group	Percentage of each feature statistics				
		Mean	Variance	Skewness	Kurtosis	Covariance
Dynamic	6	1.5	1.0	0.5	0.0	0.0
Rhythmic	1	0.0	0.5	0.0	0.0	0.0
Spectral without MFCC	58	4.55	4.5	3.0	2.0	15.0
MFCC	118	2.0	20.0	0.0	0.0	37.0
Harmonic	17	2.0	1.5	2.5	0.5	2.0
Total	200	10.0	26.5	6.0	2.5	54.0

Table 4 Comparison of classification accuracies with other methods on the GTZAN dataset

Reference	CA
Our approach	87.90 %
Shin-Chelon Lim et al.[16]	87.40 %
Jin S. Seo [22]	84.09 %
Bergstra et al. [5]	82.50 %
Li et al. [13]	78.50 %
Lidy et al. [15]	76.80 %
Benetos et al. [4]	75.00 %
Tzanetakis [26]	61.00 %

Chelo et al. by considering distinct statistics, except mean and variance, and large numbers of feature reduction methods.

To further elucidate the classification accuracy of an individual music genre, the confusion matrix is given in Table 5 for the GTZAN dataset. The confusion matrix is an n x n matrix in which each column represents the instances in a predicted class, while each row represents the instances in an actual class. The diagonal entries of the confusion matrix are the rates of music genre classification that are correctly classified, while the off-diagonal entries correspond to misclassification rates. The genres are arranged in the order of blues (Bl), classical (Cl), country (Co), disco (Di), hip-hop (Hi), jazz (Ja), metal (Me), pop (Po), reggae (Re), and rock (Ro), respectively.

From the confusion matrix, it is evident that some music genres are classified with significant accuracies, such as blues, classical, disco, hip-hop, jazz, and metal. Except for rock, all other music genres showed competitive classification performances. Rock music had a lower classification accuracy rate than the other music genres and was confused with blues, pop, and country. In our conjecture, rock music is characteristically diverse compared to other genres; therefore, it may share characteristics with other genres. Similarly, reggae was primarily confused with hip-hop and pop; it was somewhat confused with disco.

Table 5 Confusion matrix of GTZAN dataset classification accuracy using SVM

	Bl	Cl	Co	Di	Hi	Ja	Me	Po	Re	Ro
Bl	90.0	3.0	2.5	0.0	0.0	2.5	2.0	0.0	0.0	0.0
Cl	0.0	98.5	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
Co	3.0	0.0	85.0	0.0	0.0	0.0	0.0	5.0	0.0	7.0
Di	0.0	0.0	0.0	89.5	0.0	0.0	0.0	6.5	4.0	0.0
Hi	0.0	0.0	0.0	2.0	97.0	0.0	0.0	1.0	0.0	0.0
Ja	1.0	0.0	0.0	0.0	0.0	92.0	0.0	0.0	3.0	4.0
Me	0.0	0.0	0.0	0.0	3.5	0.0	88.5	0.0	0.0	8.0
Po	0.0	0.0	0.0	8.5	0.0	0.0	0.0	86.5	3.0	2.0
Re	0.0	0.0	3.0	4.0	6.0	0.0	0.0	5.0	82.0	0.0
Ro	10.0	0.0	5.0	4.0	0.0	2.0	3.0	5.0	0.0	71.0

6 Conclusions

In this paper, diverse audio features (dynamic, rhythmic, spectral, and harmonic) were selected for music genre classification. These features were then integrated using lower- (mean and standard deviation) and higher-order moments (skewness and kurtosis). Similarly, covariance components were also calculated to improve the classification. Based on obtained feature statistics, the experiments were performed in two stages. The goal was to determine the extent of classification accuracy improvement after adding the other parameters (higher-order statistics and covariance components) to the lower-order statistics using reduced features.

In the first stage, lower-order statistics were considered for classification. Later, all statistics (lower, higher, and covariance components) were included to determine overall genre classification performances. The direct consequence of considering all statistics was rapidly incrementing feature dimensions. Therefore, they were controlled by different feature reduction methodologies; i.e., PCA, LDA, LPP, NMF, and MRMR. The classification accuracies of the lower-order statistics were recorded as 75.0, 73.50, 77.25, 79.80, and 80.75 % using PCA, LDA, LPP, NMF, and MRMR reduction feature sets, respectively. Similarly, 63 % of features were insignificant when considering all feature statistics. From the remaining feature statistics, the overall classification accuracies were recorded upto 87.9, 85.20, 84.50, 78.2, and 75.20 %, respectively, using MRMR, LPP, LDA, PCA, and MNF feature reduction methodologies.

The MRMR-based feature reduction method using SVM showed a performance competitive with other contemporary methodologies, as shown in Table 4. Moreover, the MRMR algorithm determines other maximum relevance features among the all features, as shown in Table 3. Therefore, it is clearly evident which features and their corresponding statistics had a significant impact on genre classification in MRMR.

Acknowledgments This paper was supported by research funds of Chonbuk National University in 2013 and also partially supported by the National Research Foundation of Korea grant funded by the Korean government (2011-0022152).

References

1. Baniya BK, Ghimire D, Lee J (2013) Evaluation of different audio features for musical genre classification. In Proc. IEEE workshop on Signal Processing Systems, Taipei, Taiwan
2. Baniya BK, Ghimire D, Lee J (2014) A novel approach of automatic music genre classification based on timbral texture and rhythmic content features. Int. conference on Advance Communications Technology (ICACT), pp.96–102
3. Belkin M, Niyogi P (2002) "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering", Advances in Neural Information Processing Systems 14. Vancouver, British Columbia
4. Benetos E, Kotropoulos C (2008) A tensor-based approach for automatic music genre classification. Proceedings of the European Signal Processing Conference, Lausanne
5. Bergstra J, Casagrande N, Erhan D, Eck D, Kegl B (2006) Aggregate features and AdaBoost for music classification". Mach Learn 65(2–3):473–484
6. Casey M, Veltkamp RC, Goto M, Leman M, Rhodes C, Slaney M (2008) Content-based music information retrieval: current directions and future challenges. Proc IEEE 96(4):668–696
7. Cortes C, Vapnik V (1995) Support vector networks. J Mach Learn
8. Dowling WJ, Harwood DL (1986) Music cognition. Academic
9. Groeneveld RA, Meeden G (1984) Measuring Skewness and Kurtosis. J R Stat Soc Ser D (Stat) 33:391–399
10. He X, Niyogi P (2003) Locality Preserving Projections. Proceedings of the 17thAnnual Conference on Neural Information Processing Systems. The MIT Press, Cambridge, pp 153–160
11. Lartillot O, Toiviainen P (2007) MIR in Matlab (II): A toolbox for musical feature extraction from audio. In Proc. Int. Conf. Music Inf. Retrieval, pp. 127–130 [Online]. Available: http://users.jyu.fi/lartillo/mirtoolbox/

12. Li Z, Liu J, Yang Y, Zhou X, Lu H. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans. Knowl. Data Eng. [Online]. Available: http://www.computer.org/csdl/trans/tk/preprint/06509368-abs.html
13. Li T, Ogihara M, Li Q (2003) A comparative study on content-based music genre classification. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 282–289, Toronto
14. Li Z, Yang Y, Liu J, Zhou X, Lu H (2012) Unsupervised feature selection using nonnegative spectral analysis. In: AAAI
15. Lidy T, Rauber A, Pertusa A, Inesta J (2007) Combining audio and symbolic descriptors for music classification from audio. Music Information Retrieval Information Exchange (MIREX)
16. Lim S-C, Lee J-S, Jang S-J, Lee S-P, Kim MY (2012) Music-genre classification system based on spectro-temporal features and feature selection. IEEE Trans Consum Electron 58–4:1262–1268
17. Marasys, Data sets http://marsysas.info/download/data
18. Peng H, Long F (2004) An efficient max-dependency algorithm for gene selection. In 36th Symp. Interface: Computational Biology and Bioinformatics
19. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
20. S Roweis, LK Saul (2000) Nonlinear dimensionality reduction by locally linear embedding. Science, 290
21. Scheirer E, Slaney M (1997) Construction and evaluation of a robust multi-feature speech/music discriminator, in Proc. Int. Conf. Acoustics, Speech, Signal Processing, Munich
22. Seo SS, Lee S (2011) Higher-order moments for musical genre classification. Signal Process 91(8):2154–2157
23. Shen J, Pang H, Wang M, Yan S(2012) Modeling concept dynamics for large scale music search. In Proc ACM SIGIR, pp. 455–464
24. Shen J, Meng W, Yan S, Pang H, Hua X (2010) Effective music tagging through advanced statistical modeling. In Proc. of ACMSIGIR
25. L. Smith (2002) A tutorial on principal components analysis, www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
26. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(3):293–302
27. Tzanetakis G, Essl G, Cook P (2001)Automatic musical genre classification of audio signals, in Proc. Int. Conf. Music Information Retrieval, Bloomington, pp. 205–210
28. Welling M (2000) Fisher linear discriminant analysis. Technical report, University of Toronto, Kings College Road, Toronto, M5S 3G5, Canada
29. Wold E, Blum T, Keislar D, Wheaten J (1996) Content based classification, search, and retrieval of audio. IEEE Trans Multimedia 3(3):27–36
30. Xu C, Maddage NC, Shao X (2005) Automatic music classification and summarization. IEEE Trans Speech and Audio Process 6(5):441–450

Babu Kaji Baniya received the B.E degree in Computer Engineering, Nepal in 2005 and M.E. degree in Electronic Engineering from Chonbuk National University, Rep. of Korea in 2010. Currently he is pursuing his Ph.D. degree in Computer Science and Engineering at Chonbuk National University, Rep. of Korea from 2011. His main research interest includes audio signal processing, music information retrieval, source separation, pattern recognition, machine learning etc.

Joonwhoan Lee received his BS degree in Electronic Engineering from the University of Hanyang, Rep. of Korea in 1980. He received his MS degree in Electrical and Electronics Engineering from KAIST University, Rep. of Korea in 1982 and the Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA, in 1990. He is currently a Professor in Department of Computer Engineering, Chonbuk National University, Rep. of Korea. His research interests include image processing, computer vision, emotion engineering etc.