Automatic Music Genre Classification Using Timbral Texture and Rhythmic Content Features

Babu Kaji Baniya*, Deepak Ghimire, Joonwhoan Lee

Division of Computer Science and Engineering
Chonbuk National University, Jeonju 761-756, South Korea
everwith_7,deep,chlee@jbnu.ac.kr

***Abstract*— Music genre classification is a vital component for the music information retrieval system. There are two important components to be considered for better genre classification, which are audio feature extraction and classifier. This paper incorporates two different kinds of features for genre classification, timbral texture and rhythmic content features. Timbral texture contains the Mel-frequency Cepstral Coefficient (MFCC) with other several spectral features. Before choosing a timbral feature we explore which feature contributes a less significant role on genre discrimination. This facilitates the reduction of feature dimension. For the timbral features up to the 4-th order central moments and the covariance components of mutual features are considered to improve the overall classification result. For the rhythmic content the features extracted from beat histogram are selected. In the paper Extreme Learning Machine (ELM) with bagging is used as the classifier for classifying the genres. Based on the proposed feature sets and classifier, experiments are performed with two well-known datasets: GTZAN and the ISMIR2004 databases with ten and six different music genres, respectively. The proposed method acquires better and competitive classification accuracy compared to the existing approaches for both data sets.**

***Keyword*— Classification, music genres, ELM (Extreme Learning Machine) with bagging, covariance matrix, timbral texture, rhythmic contents**

I. INTRODUCTION

TUTOMATIC music genre classification is an important for the information retrieval task since it can be applied for practical purposes such as efficient organization of data collections in the digital music industry. There have been several well-known distinct approaches put forward on this. Still, efficient and accurate automatic music information processing remains as the key issue, and it has been consistently attracting the attention of a growing number of researchers, musicians, and composers. A current challenging topic in automatic music information retrieval is the problem of organizing, describing, and categorizing music contents on the internet [1]. Although music genre classification is done manually, sometimes it is difficult to precisely define the genre of music content. The reason for such difficulties is due to fact that music is a state of art that evolves, where composers and musicians have been influenced by the music of other genres. Despite these difficulties, there are still some possibilities that remain for genre classification. The audio signals of music belonging to the same genre mean they share the certain common characteristics, because they are composed of similar types of instruments, having similar rhythmic patterns, and similar pitch distributions [2]. The extracted features must be comprehensive (representing music very well), compact, and effective.

The overview of our music genre classification is shown in Fig.1. It depicts the backbone of genre categorization. There are two associated problems that need to be addressed in genre classification, i.e., feature extraction and classification. The first stage is to extract the meaningful and relevant features from audio that could sufficiently discriminate the music genre. The next stage is to classify the genre based on the extracted features. In our method the extreme learning machine (ELM) combined with bagging is used as a classifier. Several bags of the dataset are constructed and each bag is trained using individual ELMs. The final decision is made based on the majority voting score. ELM is an unstable classifier, therefore ELM combined with bagging increases the stability, as well as generalization performance of the classifier.

For constructing a robust music genre classifier, extracting features that allows direct access to the relevant genre-specific information is crucial. Most musical genre classification systems utilize the low-level spectral features of the short time audio signal in the range of 10ms to 100ms, such as pitch extraction, mel-frequency cepstral coefficients (MFCCs), and other timbral texture features [3]. Then the short-time low-level spectral features are integrated on long duration. The most widely used integrating method is mean and standard deviation of the short time feature [4, 5].

In this paper, we attempt to implement timbral texture features which represent short-time spectral information, and rhythmic content features like beat histogram which represent

Manuscript received May 9, 2014. This work was partially supported by a National Research Foundation of Korea grant funded by the Korean government (2011-0022152) and PK21PLUS project.

Babu Kaji Baniya is with the Computer Engineering Department of Chonbuk National University, Jeonju, South Korea (e-mail: everwith_7@jbnu.ac.kr).

Deepak Ghimire is with the Computer Science and Engineering Department of Chonbuk National University, Jeonju, South Korea (e-mail: deep@jbnu.ac.kr).

Corresponding author: Joonwhoan Lee is with the Computer Engineering Department of Chonbuk National University, Jeonju, South Korea (phone: +82-10-9855-2406; Fax: +82-63-270-2406; e-mail: chlee@jbnu.ac.kr).

the long-term properties. Timbral texture features include spectral centroid, flux, rolloff, flatness, energy, zero crossing, and MFCCs, respectively. We divide the timbral texture features into two groups for convenience; the first group (FG1) does not include MFCCs and the second group (SG2) includes only MFCCs. After the frame-wise extraction of each timbral texture feature among FG1 from all genres of music, the next stage is to calculate the standard deviation for all genres of music. The aim of calculating the standard deviation for each feature in whole genres is to find out which feature is insignificant for genre discrimination. The feature which has a small value of standard deviation contributes the insignificant impact on genre discrimination. Based on the standard deviation value, we considered a limited number of timbral features.

A Similar procedure has been preceded for the SG2 of MFCC features as well. Out of thirteen, twelve coefficients give meaningful standard deviation values. This shows that twelve MFCC coefficients are meaningful for genre classification. For our experiment, we consider both the first seven and twelve coefficients separately for genre classification.

Timbral texture features are based on short time low-level spectral components that are integrated on long duration. The integration method is mean and standard deviation. Beside this, high order moments such as skewness and kurtosis are also implemented for integrating the frame-wise features. The aim of considering the high order moment is that even if there are the same values of mean and standard deviation, the position of location (shape of skewness and kurtosis position) could be different because each feature cannot be modelled by the Gaussian distribution.

Ultimately, the high order moment increases the 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 for the degree of peakedness or flatness of a distribution [6]. Therefore we have considered 4*n* components for the *n* texture features.

In addition we propose to use the covariance components of mutual timbral texture features. Each of them gives the statistical property of mutual random variables associated features. For each song the covariance values of selected features from FG1 and SG2 are calculated, respectively. Therefore, additional $n(n-1)/2$ components are included for *n* timbral texture features.

Note that we can have $4n+ n(n-1)/2$ for *n* features, which increases rapidly as the number of features increases. This is the reason why we remove the relatively less important features by checking the corresponding variances.

Rhythm is a property of an audio signal that represents a changing pattern of timbral and energy over time. Rhythmic features characterize the movement of music signals over time and contain such information as the regularity of the rhythm, beat, tempo, and time signature. The feature set for representing the rhythmic structure is based on detecting the most salient periodicities of the signal and it is usually extracted from the beat histogram. Rhythmic content features contain relative amplitude of the first and second histogram peaks, period of first and second peaks, ratio of the amplitude of the second peak divided by the amplitude of first peak, and overall sum of the histogram.

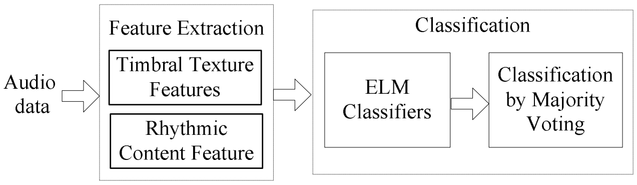

Fig.1. Overview of music genre classification

There are different types of classifiers which have been proposed for genre classification. We prefer the distinct classifier than the previously applied one. Extreme Learning Machine (ELM) is a recently proposed classifier which has high generalizing capability and takes minimum time for training. The reason for selecting ELM is that it does not require a tuning parameter, has the smallest training error, and is free from the local maxima problem. However, ELM is unstable because the weights connected with hidden units are randomly determined. Therefore, we combine ELM with bagging in order to increase the stability. Bagging is almost always more accurate than a single classifier. Other classifiers like K-Nearest Neighbour (K-NN), Neural Networks (NN) have some drawbacks. In case of neural network, when learning rate is too small, the algorithm converges very slowly. It also requires a tuning parameter and probably faces the local maxima problem. K-Nearest Neighbour is a simple nonparametric classifier. It is proven that the error of K-NN's is twice large than Bayesian error rate.

This paper is organized as follows. A review of related work is provided in section II. Feature extraction is the critical portion of genre classification; and is describes in section III. Section IV deals with the classifier, similarly section V explains the experimental setup and data preparation, and section VI explains the result and analysis. Finally, section VII describes the conclusion of the proposed method and future work of the genre.

II. Related Work

Many different features have been introduced for music genre classification. The primary aim of feature extraction is to acquire a meaningful representative part of music in the reduced form. The acoustic features include tonality, pitch, beat, and symbolic features extracted from the scores, and text-based features can be obtained from the song lyrics. In this paper, we only focus on timbral texture and rhythmic content which are sub-groups of content-based features.

The content-based acoustic features are divided into timbral texture features, rhythmic content features, and pitch content features [7]. Timbral features are often calculated for every short-time frame of sound based on the Short Time Fourier Transform (STFT) [8]. Timbral texture features contain MFCCs, spectral centroid, spectral flatness, spectral flux, spectral rolloff, zero crossing, energy, and Linear Prediction Coefficients (LPCs) [7, 8]. These features are widely used in different applications based on the requirement of applications. MFCCs have been extensively used in speech recognition [8]. Later, MFCC features are used for discriminating the music and speech as well. Rhythmic content features possess information about continuity of rhythm, beat and tempo. Tempo and beat tracking are excessively used in music search and retrieval systems. The tempo value is a number which represents the speed of music or music measured by beats per minute (bmp) [9, 10]. The pitch content feature deals with frequency information of music.

Bergstra et al. [11] extracts the several timbral texture features like MFCCs, spectral centroid, spectral flux, spectral rolloff, zero crossing, energy, and Linear Prediction Coefficients. These features are almost similar with the features used in [3, 5]. AdaBoost is used as a classifier.

C.-H. Lee et al. [12] considers the Octave-Based Spectral Contrast (OSC) and MFCC for feature extraction. There is a range of nine different frequencies in octave-based spectral contrast. Music genres are classified by using Linear Discriminant Analysis (LDA). Recently, Seo et al. [13] also implemented the Octave-Based Spectral Contrast (OSC) for feature extraction. Beside this, he consider the high order moment for improving the performance of classification accuracy. The genre classification is performed by using Support Vector Machines (SVM).

Li et al. [1] mention several audio feature extraction methodologies. Later, he proposed a new approach for feature extraction, i.e. Daubechies Wavelet Coefficients Histograms (DWCHs). The effectiveness of this new feature is compared using various machine learning algorithms, SVMs, Gaussian Mixture Models (GMMs), K-NNs, and LDAs.

The spectral similarity of the timbral texture feature is described by Pampalk et al. [14]. The audio signal is chopped into thousands of very short frames and their order in time is ignored. Each frame is described by MFCCs. The large set of frames is summarized by a model obtained by clustering the frames. The distance between two pieces is computed by comparing their cluster models. Later, GMM is considered for genre classification.

Tzanetakis and Cook [7] proposed a comprehensive set of features for direct modelling of music signals and explore the different applications of those features for musical genre classification using K-Nearest Neighbor and GMM. Other researchers like Lambrou et al. [15] use statistical features in the temporal domain as well as three different wavelet transform domains to classify music into rock, piano, and jazz.Soltau et al. [16] propose an approach of representing temporal structures of input signals. He shows that this new set of abstract features can be learned via artificial neural networks and can be used for music genre identification. Deshpande et al. [17] use Gaussian Mixtures, SVM, and K-Nearest Neighbor to classify the music into rock, piano, and jazz based on timbral texture features.

III. FEATURE EXTRACTION

Feature extraction encompasses the analysis and extraction of meaningful information from audio in order to obtain a compact and concise description that could be machine readable. Features are usually selected in the context of a specific task and domain. The features that are used in our research are divided into two categories, the timbral texture feature and rhythmic content feature.

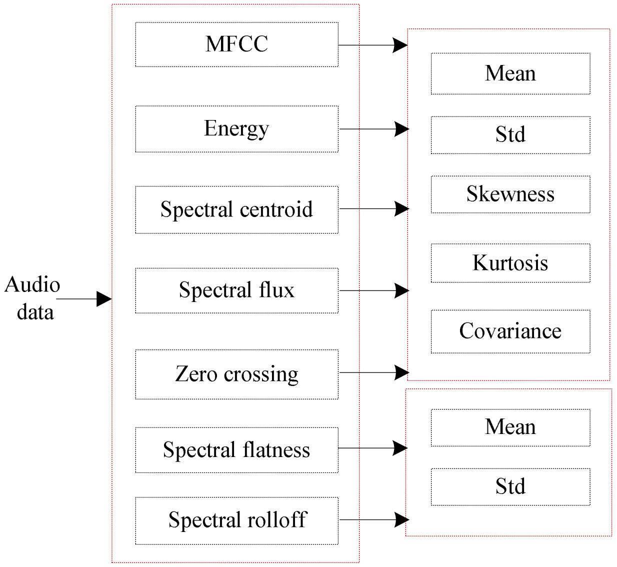

Fig. 2. Overview of Timbral texture features extraction of audio.

A. Timbral Texture features

These features are used to differentiate mixture of sounds that possibly have similar pitch and rhythm [8]. The features used to represent timbral texture are based on standard features proposed for music-speech discrimination [18]. To extract the timbral features, audio signals are first divided into frames by applying a windowing function at fixed intervals. The window function of this research is hamming window which helps to remove the edge effects. Timbral texture features in Fig.2 have been computed and later we calculated different statistical values like mean, standard deviation, skewness, kurtosis, and covariance matrix from feature values. The mean (μ) and standard deviation (σ) for frame-wise feature values (X_n) in a N -frame song are given by

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

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

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

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

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean. It is difficult to specify the exact contribution of kurtosis in music genre classification [13]. However, the kurtosis measure can sketch the effective dynamic range of the audio spectrum. Mathematically it can be 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. The aim of considering the covariance is usually to see if there is any relationship between the random variables. It is useful to measure the polarity and the degree of the correlation between two features. The covariance of two features X_n and Y_n in a song 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 acquired $n(n-1)/2$ mutual covariance values.

We consider two groups of timbral texture features FG1 and SG2 described as

1) *FG1 features*

Spectral flux: It is defined as the variation value of the spectrum between the adjacent two frames in a short-time analyze window. It measures how quickly the power spectrum changes and is used to determine the timbral of an audio signal.

$$F_t = \sum_{n=1}^{N}(N_t[n]-N_{t-1}[n])^2 \tag{6}$$

where $N_t[n]$ and $N_{t\text{-}1}[n\text{-}1]$ are normalized magnitudes of the Fourier transform at the present frame t, and previous frame t-1 respectively.

Spectral centroid: The spectral centroid is described as the gravity centre of the spectral energy. It determines the point in the spectrum where most of the energy is concentrated and is correlated with the dominant frequency of the signal. It is closely related to the brightness of a single tone.

$$C_t = \frac{\sum_{n=1}^{N} M_t[n]*n}{\sum_{n=1}^{N} M_t[n]} \tag{7}$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n.

Short Time Energy: The short time energy measurement of an audio signal can be used to determine voiced and unvoiced speech. It can also be used to detect the transition from unvoiced to voice and vice versa [19]. The energy of voiced speech is much greater than the energy of unvoiced speech. Short-time energy can be defined as

$$E_n = \sum_{m=1}^{N}[x(m)w(n-m)]^2 \tag{8}$$

where, $x(m)$ is discrete time audio signal, n is time index of short-time energy, and $w(m)$ is window of length N.

Zero Crossing: It is a process of measuring the number of times in a given time interval that the amplitude of speech signals crosses through a value of zero. It is random in nature. Moreover, the zero crossing rate for unvoiced speech is greater than that of voice speech. Moreover, it is often used as a crucial parameter for voiced/unvoiced classification and end point detection.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N}|\operatorname{sgn}(x[n]-\operatorname{sgn}(x[n-1])| \tag{9}$$

where sgn is a short notation of sign function. The sgn function is 1 for positive arguments and 0 for negative arguments and x[n] is the time domain for signal for frame t.

Spectral Rolloff: It is a measure of the bandwidth of the audio signal. It is the fraction of bins in the power spectrum in which 85% of the power is at lower frequencies.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85\sum_{n-1}^{N} M_t[n] \tag{10}$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n.

Spectral flatness: It is used to characterize an audio spectrum. Spectral flatness is typically measured in decibels, and provides a way to quantify how tone like a sound is, as opposed to being noise-like.

$$F = \frac{\exp\left(\frac{1}{N}\sum_{m=0}^{N-1}\ln x(m)\right)}{\frac{1}{N}\sum_{m=0}^{N-1}x(m)} \tag{11}$$

where $x(m)$ represents the magnitude of bin number m.

From the above mentioned features in FG1, the normalized standard deviation of all the data has been calculated. Since the standard deviation generally depends on the mean value in general, the standard deviation is divided by corresponding mean to find out the less important features. Note that a smaller value of the standard deviation means a smaller change in the values of the frame-wise timbral texture feature. This means any derived central moments from the feature and the covariance with the feature is not significant for the discrimination of music genres. Therefore we removed such features to reduce the feature dimension.

Spectral centroid, flux, short time energy, and zero crossing possess large normalized standard deviations compared to the rolloff and flatness as shown in Table IV. We only consider four features (Spectral centroid, flux, short time energy, and zero crossing) and their mean, std, skewness, kurtosis and $n(n\text{-}1)/2$ covariance components, respectively. The feature dimension is given in Table I.

2) *SG2 Features: Mel-Frequency Cepstral Coefficients*

TABLE I
FEATURE DIMENSION OF FOUR DIFFERENT TIMBRAL TEXTURE FEATURES

Mean	Std. dev.	Skew.	Kurt.	Cov.	Total features
4	4	4	4	6	22

Earlier MFCCs widely used in automatic speech recognition later on evolved into one of the prominent techniques in every domain of audio retrieval. They represent most distinctive information of signal. MFCCs have been successfully implemented to timbral measurements by H. Terasawa [20].

We took the MFCC feature based on the paper that mentioned the mel frequency cepstral coefficients for music modelling [21]. Fig. 3 shows the process of creating MFCC features. The first step is to divide the audio signal into frames, by applying a window function at fixed intervals. The aim is to model small (having 10ms) sections of the signal that are statistically stationary. The window function is hamming window. We generate the cepstral feature vector for each frame. The next step is to take the Discrete Fourier Transform (DFT). The phase information has been discarded because perceptual studies have shown that the amplitude of the spectrum is much more important than the phase. The logarithm of the amplitude spectrum has been taken because the perceived loudness of a signal has been estimated to be approximately logarithmic. The next stage is to smooth the spectrum and emphasize perceptually meaningful frequencies. This is achieved by collecting the spectral components into frequency bins. As we know, lower frequencies are perceptually more important than the higher frequencies. Therefore, the bin spacing follows the so-called 'Mel' frequency scale [22]. The components of the Mel-spectral vectors calculated for each frame are highly correlated. In order to reduce the number of parameters in the MFCC, we need to apply a transform to the Mel-spectral vectors which decorrelates their components. The cepstral features of each frame are obtained by using DCT.

by applying a window function at fixed intervals. The aim is to model small (having 10ms) sections of the signal that are statistically stationary. The window function is hamming window. We generate the cepstral feature vector for each frame. The next step is to take the Discrete Fourier Transform (DFT). The phase information has been discarded because perceptual studies have shown that the amplitude of the spectrum is much more important than the phase. The logarithm of the amplitude spectrum has been taken because the perceived loudness of a signal has been estimated to be approximately logarithmic. The next stage is to smooth the spectrum and emphasize perceptually meaningful frequencies. This is achieved by collecting the spectral components into frequency bins. As we know, lower frequencies are perceptually more important than the higher frequencies. Therefore, the bin spacing follows the so-called 'Mel' frequency scale [22]. The components of the Mel-spectral vectors calculated for each frame are highly correlated. In order to reduce the number of parameters in the MFCC, we need to apply a transform to the Mel-spectral vectors which decorrelates their components. The cepstral features of each frame are obtained by using DCT.

There are thirteen coefficients in the mel-frequency cepstral coefficient. After analysis of the normalized variance we selected 12 out of 13 coefficients. The last coefficient has a very small value of the variance as shown in Table V, so that it could be removed. We try to implement the early seven and twelve coefficients separately. The first choice is just for reducing the dimension. The different feature dimension of MFCC while considering seven and twelve coefficients are given in table II and III.

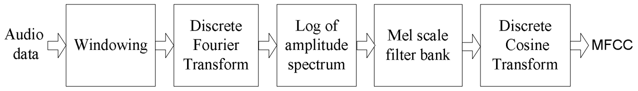
Audio data
Windowing
Discrete Fourier Transform
Log of amplitude spectrum
Mel scale filter bank
Discrete Cosine Transform
MFCC

Fig. 3. Mel frequency cepstral coefficients feature extraction of audio.

TABLE II

FEATURE DIMENSION OF MFCC CONSIDERING FIRST SEVEN COEFFICIENTS

Mean	Std. dev.	Skew.	Kurt.	Cov.	Total features
7	7	7	7	21	49

TABLE III

FEATURE DIMENSION OF MFCC CONSIDERING TWELVE COEFFICIENTS

Mean	Std. dev.	Skew.	Kurt.	Cov.	Total features
12	12	12	12	66	114

B. *Rhythmic Content Features*

Rhythmic content features characterize the movement of music signals over time and contain such information as the regularity of the rhythm, beat, and tempo. For the rhythmic feature, beat histogram has been taken. It is a compact global representation of the rhythmic content of audio music. The beat histogram [5] can be obtained by the wavelet decomposition of a signal and can be interpreted as successive high-pass and low-pass filtering of the time domain signal. The decomposition is defined by

$$y_{high}[k] = \sum_{n} x[n]g[2k-n] \qquad (12)$$

$$y_{low}[k] = \sum_{n} x[n]h[2k-n] \qquad (13)$$

where $y_{high}[k]$ and $y_{low}[k]$ are the output of high-pass and low-pass filters respectively, and $g[n]$ and $h[n]$ are the filter coefficients for the high-pass and low-pass filters associated to the wavelet function for fourth order Daubechies wavelets (DW) [22]. Wavelet Transform deals with the similarity of the decomposed signal to the octave filter band. Once the signal is decomposed, the additional signal processing operation is required. The building blocks as shown in Fig. 4 are used for the beat analysis feature extraction.

1) *Full Wave Rectification:*

$$y[n] = abs(x[n]) \qquad (14)$$

where x[n] is the output of the wavelet decomposition at that specific scales.

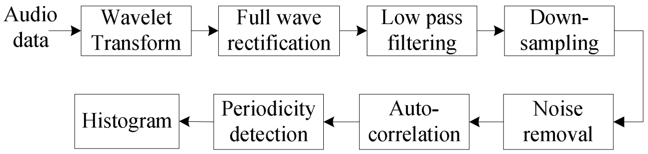

Fig.4. The block diagram of beat histogram for feature extraction

2) *Low-Pass Filtering:*

$$a[n] = (1-\alpha)y[n] + \alpha a[n-1] \qquad (15)$$

For one-pole filter with an alpha value of 0.99 which is used to smooth the envelope.

3) *Downsampling:*

$$b[n] = a[kn] \qquad (16)$$

Downsampling the signal reduces computation for the autocorrelation calculation without affecting the performance of the algorithm. The value of *k* is 16.

4) *Normalization (mean removal)*

Mean removal is applied in order to make the signal centered to zero for the autocorrelation stage.

$$c[n] = b[n] - E[b[n]] \tag{17}$$

5) *Autocorrelation*

$$d[k] = \frac{1}{N} \sum_{n} c[n]c[n-k] \tag{18}$$

where c[n] is periodic signal with period N.

6) *Periodicity detection and beat histogram calculation:*

There are six different features extracted from the beat histogram. They are relative amplitude of the first and second histogram peak, period of the first and second histogram peak measure in beat per minute (bpm), ratio of the amplitude of the second peak divided by the amplitude of the first peak, and overall sum of the histogram.

IV. CLASSIFIER

Traditionally, all the parameters of the feed-forward networks need to be tuned and thus there exists the dependency between different layers of parameters (weights and biases). In particular the gradient descent-based methods have been used in various learning algorithms of feed-forward neural networks [23]. However, the weakness of this kind of learning method is that it is generally very slow due to diverse learning steps and may easily converge to local minima. They also require many iterative learning steps in order to obtain better learning performance.

ELM [24] resolves the problem associated with the gradient-based algorithm by analytically calculating the optimal weights of single-hidden layer feed-forward neural networks (SLFNs). Where the weights between input layers and the hidden layer biases are arbitrarily selected and then the optimal values for the weights between the hidden layer and output layer are determined by calculating the linear matrix equations.

For N distinct samples and $\tilde{N}$ hidden nodes, the activation function g(x) of the SLFN neural network is defined as

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i . x_j + b_i) = o_j, \; j = 1, \ldots\ldots, N \tag{19}$$

where $w_i = [w_{i1}, w_{i2}, .., w_{in}]^T$ is the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, .., \beta_{im}]^T$ is the weight vector connecting the i-th hidden nodes and output nodes, and b_i is the threshold of the i-th hidden node. $w_i.x_j$ denotes the inner product of w_i and x_i.

The standard SLFNs with $\tilde{N}$ hidden nodes with the activation function $g(x)$ can approximate these N samples with zero error means that $\sum_{j=1}^{\tilde{N}} \| o_j - t_j \| = 0$ i.e., there exist β_i, w_i, and b_i such that

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i . x_j + b_i) = t_j, \; j = 1, \ldots\ldots, N \tag{20}$$

where t_j is the target vector of the j-th input data. Equation (19) can be written as a matrix equation to form a new equation by using the output matrix of the hidden layer H.

$$H\beta = T \tag{21}$$

where

$$H = \begin{bmatrix} g(w_1 . x_1 + b_1) & \cdots & g(w_{\tilde{N}} . x_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(w_1 . x_N + b_1) & \cdots & g(w_{\tilde{N}} . x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} \tag{22}$$

$$\beta = \begin{bmatrix} \beta_1{}^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad and \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{23}$$

From the above equation (21), the target vector T and the output matrix of the hidden layer H can comprise a linear system. Thus, the learning procedure of the network helps to find the optimal weight matrix β between the output layer and the hidden layer β can be determined by using the Moore-Penrose generalized inverse of H.

$$\hat{\beta} = H^{\dagger} T \tag{24}$$

From the above equation (24) we can draw the following important properties. The first one is that we can take minimum training error, because the solution $\hat{\beta} = H^{\dagger} T$ is one of the least-square solutions of the general linear system $H\beta = T$. In addition, the optimal $\tilde{\beta}$ is also minimum norm among these solutions. Thus, ELM has the best generalization performance compared to the typical back propagation network. In summary the ELM algorithm can be summarized as follows.

Algorithm ELM: For the given training set $\aleph = \{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1, ..., N\}$, activation function $g(x)$, and hidden neuron number $\tilde{N}$,

1) Assign random input weight w_i and bias b_i, i=1,…, $\tilde{N}$.
2) Calculate the hidden layer output matrix H.
3) Calculate the output weight β:

$$\hat{\beta} = H^{\dagger} T$$

Where $H^{\dagger}$ is the Moore-Penrose generalized inverse of hidden the layer output matrix H.

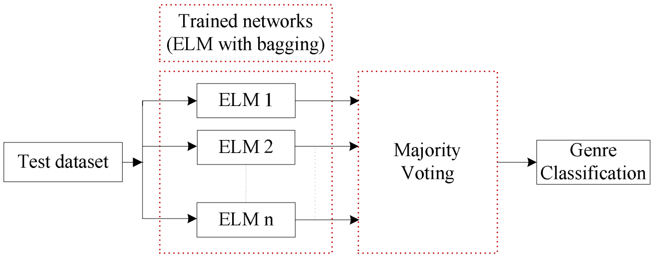

Fig. 5: Block diagram of music genre classification

A. *Bagging Algorithm*

Bagging [25] is a well-known ensemble learning algorithm that has been shown to be very effective in improving generalization performance compared to the individual base models. Breiman indicated that bagging is a smoothing operation which turns out to be advantageous when aiming to improve the predicative performance of regression or classification. It is a “bootstrap” ensemble method that creates bags for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement; many of the original samples may be repeated in the resulting training set while others may be left out. Each bag classifier in the ensemble is generated with a different random sample of the training set. The algorithm then applies a base classifier to classify each bag. Bagging is almost always more accurate than a single classifier. Finally the decision is taken by majority voting of all the base classifier results. Fig. 5 shows the overview of genre classification. Our base classifier is ELM.

```
The Bagging Algorithm
Inputs: Training set $S$, based classifier $L$, integer $T$ (number of bootstrap sample)
    for $i$ = 1 to $T$ {
        $S_i$=bootstrap sample from $S$ (i.i.d. sample with replacement)
        $E_i$=$L(S_i)$ }
  $E^*(x) = \arg\max_{y \in Y} \sum_{i:E_i(x)=y} 1$   (the most often predicated label $y$)
  Output: Compound classifier $E^*$
```

V. Experimental setup and Data Preparation

Different datasets widely used for music genre classification are employed for performance comparison. The first dataset (GTZAN) consists of 1000 songs over ten different genres: Classical, Blues, Hiphop, Pop, Rock, Gazz, Reggae, Metal, Disco, and Country. Each class consists of 100 songs having duration of 30s. The dataset was collected by Geroge Tzanetakis [26]. Each song in the database was stored as a 22050Hz, 16bits, and mono audio file. The second dataset is ISMIR2004 [27] which were used in the Music Genre Classification Contest 2004. This dataset has an unequal number of distributions of music tracks in each class. It consists of six different classes: Classical, Pop and Rock, Metal and Punk, Electronic, World, and Jazz and Blues respectively. This dataset consists of 1458 music tracks in which 729 music tracks are used for training and the other 729 tracks for testing. The audio files are stored in MP3 format having a sampling frequency of 44.1 kHz, 128-kbps, 16 bit, and stereo files. For our research, each stereo MP3 file was first converted into a 44.1 kHz, 16 bit, mono audio file before feature extraction. In summary, the music tracks used for training/testing include 320/320 tracks of Classical, 115/114 tracks of Electronic, 26/26 tracks of Jazz/Blues, 45/45 tracks of Metal/Punk, 101/102 tracks of Rock/Pop, and 122/122 tracks of World music genre.

A five-fold cross validation scheme is used to evaluate the performance of the proposed system in the GTZEN dataset whereas in order to compare our proposed method with the results from the ISMIR2004 Music Genre Classification Content, our experiment on the ISMIR2004 genre dataset used the same training and testing set as in the contest. In the contest, the classification performance is evaluated based on 50:50 training and testing set instead of five-fold cross validation.

VI. Result and Analysis

At first, in order to reduce the dimensionality of the extracted feature set, the normalized standard deviation of each timbral texture feature is calculated in both the GTZAN and ISMIR2004 dataset. As the number of timbral texture feature increases the dimensionality of the extracted feature set increases rapidly, therefore we removed the relatively less important features by checking the corresponding normalized standard deviations. Table IV and V contain the average normalized standard deviation of the GTZAN and ISMIR2004 datasets. The data seen in table IV indicates that flatness and rolloff are less significant for genre classification than the other four in FG1. A similar approach is also applied for SG2 of MFCC coefficients. There is only one MFCC coefficient which has a relatively small value of normalized standard deviation. Aside from that coefficient, the remaining twelve coefficients are useful for genre classification.

TABLE IV
NORMALIZED STANDARD DEVIATION OF TIMBRAL TEXTURE FEATURES (EXCLUDING MFCC)

	Energy	Centroid	Flux	Zerocrossing	Flatness	Rolloff
Nor.Std	1.085	0.690	0.886	0.671	0.239	0.274

TABLE V
NORMALIZED STANDARD DEVIATION OF MFCC

MFCC Coefficients	1	2	3	4	5	6	7
Normalized Standard Deviation	2.03	2.31	2.36	2.43	2.49	2.38	2.43
MFCC Coefficients (contd.)	8	9	10	11	12	13	
Normalized Standard Deviation	2.24	2.28	2.48	2.29	1.60	0.02	

The genre classification result of the GTZAN and ISMIR2004 datasets is shown in table VI and VII respectively. Several experiments have been conducted among the different feature sets. The first experiment was conducted within timbral texture features in FG1 like spectral centroid, flux, energy, and zero crossing for genre classification (excluding MFCC). The second and third experiments were only conducted for SG2 with seven and twelve mel-frequency cepstral coefficients (feature dimension shown in table II and III respectively). The fourth and fifth experiments only considered mean, standard deviation, skewness, and kurtosis of timbral texture feature including seven and twelve MFCC coefficients separately with the rhythmic content feature. The final experiment was conducted taking the covariance matrix and all other features. The experiment was performed into two different steps to find the classification accuracy in regard to minimum (seven MFCC coefficients) and maximum (twelve MFCC coefficients) feature dimensions considering mean, standard deviation, skewness, and kurtosis.

The combination of different feature sets gives different classification accuracy. The feature extracted from FG1 of energy, centroid, flux, and zero crossing gives 68.33 % of accuracy. Similarly, SG2 (MFCC feature sets) with seven and twelve coefficients give 64.62% and 66.26% accuracy respectively. This experimental result shows that seven or eleven coefficients of MFCC do not make a big difference in genre classification. We also tried to find out the overall impact of classification accuracy with covariance components. The classification accuracy of GTZAN dataset without covariance components comes around 78.26% (7-MFCCs) and 80.21% (12-MFCCs) respectively. Among them, maximum accuracy is obtained while combining the covariance components with other feature sets. The classification accuracy of the GTZAN dataset increases from 80.21% to 85.58% while including covariance components. Similarly, ISMIR2004 classification accuracy also increases from 81.53% to 86.46%. Hence, covariance components had significant impact in improving the genre classification.

TABLE VI
CLASSIFICATION ACCURACY (CA) OF GTZAN DATASET IN DIFFERENT FEATURE SETS

Feature set	ELM(CA)
Energy+Centroid+Flux+Zerocrossing	68.33%
MFCC (7 coff)	64.62%
MFCC (12 coff)	66.26%
[ECFZ+MFCC, 7 coff.] [without covariance]+beat histogram	78.26%
[ECFZ+MFCC, 12 coff.] [without covariance]+beat histogram	80.21%
[ECFZ+MFCC , 7 coff.] [with covariance]+beat histogram	84.52%
[ECFZ+MFCC, 12 coff.][with covariance]+beat histogram	85.15%

TABLE VII
CLASSIFICATION ACCURACY (CA) OF ISMIR2004 DATASET IN DIFFERENT FEATURE SETS

Feature set	ELM(CA)
Energy+Centroid+Flux+Zerocrossing (ECFZ)	73.62%
MFCC (7 coff)	66.58%
MFCC (12 coff)	68.78%
(ECFZ+MFCC (7 coff.)) (without covariance)+beat histogram	79.65%
(ECFZ+MFCC (12 coff.)) (without covariance)+beat histogram	81.53%
(ECFZ+MFCC (7 coff.)) (with covariance)+beat histogram	85.15%
(ECFZ+MFCC (12 coff.)) (with covariance)+beat histogram	86.46%

In our approach, the extreme learning machine combined with bagging algorithm is used for the classification of the music genre. Before we applied the ELM with bagging as a classifier, we attempted to find out how many bags were needed to obtain maximum classification accuracy. There were twenty three bags combined in case of the GTZAN dataset to get maximum classification as shown in Fig.6. In the ISMIR2004 dataset maximum classification accuracy was achieved when twenty five bags were used as shown in Fig.7.

Table VIII compares our proposed method with other approaches in terms of average classification accuracy in the GTZAN dataset. It is clear that our proposed method achieves the classification accuracy of 85.15% which is better than other approaches. Similarly, Table IX shows the comparison results with previous different approaches as well as the ISMIR2004 Music Genre Classification Contest. The classification accuracy is 86.46%. It is also comparatively competitive with Chang-Hsing Lee's method [26] and better than all other approaches including the ISMIR Music Genre Classification Contest (classification accuracy 84.07%) shown in table IX.

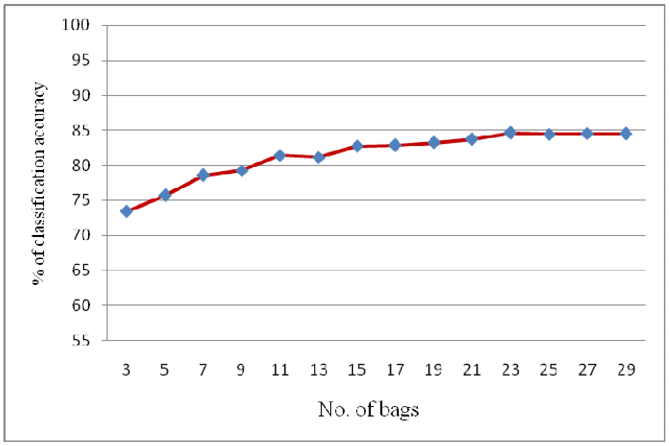

Fig. 6: Number of bags Vs classification accuracy of ZIGEN dataset.

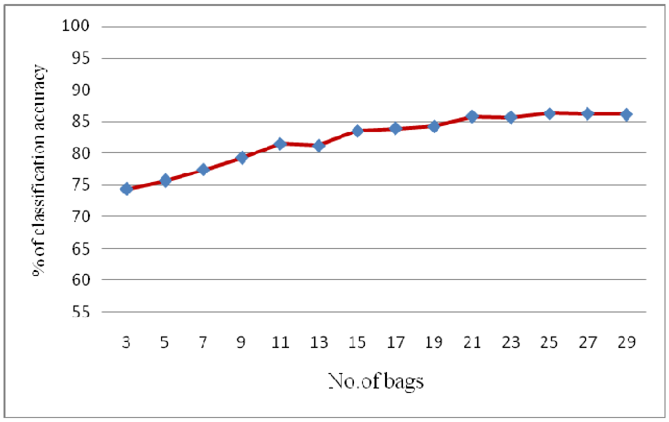

Fig.7: Number of bags Vs classification accuracy of ISMIR2004 dataset.

TABLE VIII
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER APPROACH OF GTZAN DATASETS (OUR APPROACH BASED ON FIVE-FOLD CROSS VALIDATION

Reference	CA
Our approach	85.15%
Jin S. Seo [11]	84.09%
Bergstra et al [9]	82.50%
Li et al.[1]	78.50%
Tzanetakis [5]	61.00%

TABLE IX
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER APPROACH OF ISMIR2004 DATASETS

Reference	CA
Our approach (ECFZ+MFCC+beat histogram)	86.46%
Chang-Hsing Lee[10]	86.83%
Jin S. Seo [11]	84.90%
Pampalk et al. [12]	84.07%
Bergstra et al [9]	82.34%
Our approach (ECFZ+MFCC+beat histogram)	86.46%

To get a better picture of the classification accuracy of an individual music genre, the confusion matrices are given. The confusion matrix is *n* x *n* matrix, at which each column of the matrix 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.

Table VIII shows the confusion matrix of the GTZAN dataset. The genres are arranged in the order of Classical (Cl), Blues (Bl), Hiphop (Hi), Pop (Po), Rock (Ro), Jazz (Ja), Reggae (Re), Metal (Me), Disco (Di), and Country (Co) respectively. Similarly, Table IX shows the confusion matrix of genre classification of the ISMIR2004 dataset. The genres are arranged in the order of Classical (Cl), Pop and Rock (P&R), Metal and Punk (M&P), Electronic (Ele), World (Wo), and Jazz and Blues (J&B) respectively.

Form the confusion matrix of the GTZAN dataset; we can see that some music genres are classified with significant accuracy like Classical, Pop, Metal, and Reggae. Except for Rock, other music genre classification rates are also competitive. Rock music has a minimum classification rate. It is confused with Metal. Beside this, it is diverse in nature as compared to other genre and also overlaps its characteristic with other genres. Music genre Disco and Country are also confused with Reggae and Metal respectively.

Similarly, Table XI shows the confusion matrix of the ISMIR2004. The classification rate of Classical, Pop and Rock, and Electronic are significant. Genres like Jazz and Blues, and World are also relatively better than Metal and Punk. The World music is diverse in nature, so it is confused with Classical and Pop and Rock. Genre like Jazz and Blues are also confused with Classical. Among the six genres, Electronic has minimum classification rate as compared to others.

TABLE X

CONFUSION MATRIX OF GTZAN DATASETS CLASSIFICATION ACCURACY WITH TIMBRAL TEXTURE AND RHYTHMIC CONTENT FEATURES

	Cl	Bl	Hi	Po	Ro	Ja	Re	Me	Di	Co
Cl	**95.0**	3.67	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.33
Bl	0.0	**85.25**	5.25	0.0	0.0	0.0	2.48	4.52	2.50	0.0
Hi	0.0	3.24	**81.26**	2.06	7.39	0.0	0.0	0.0	8.03	0.0
Po	0.0	0.0	0.0	**96.82**	1.14	0.0	0.0	0.0	2.04	0.0
Ro	2.18	1.15	0.0	1.39	**74.92**	0.0	3.55	7.09	5.03	2.69
Ja	5.21	4.0	0.0	0.0	2.52	**83.12**	0.0	0.0	0.0	5.15
Re	0.0	0.0	0.0	8.64	2.23	3.68	**85.45**	0.0	0.0	0.0
Me	0.0	0.0	0.0	3.89	0.0	5.13	0.0	**89.57**	2.28	1.15
Di	0.0	0.0	2.35	0.0	4.95	0.0	6.25	2.35	**83.90**	0.0
Co	0.0	6.12	0.0	4.19	0.0	0.0	0.0	9.18	0.0	**80.51**

TABLE XI

CONFUSION MATRIX OF ISMIR2004 DATASET CLASSIFICATION ACCURACY OF TIMBRAL TEXTURE AND RHYTHMIC CONTENT FEATURES

	Cl	P&R	M&P	Ele	Wo	J&B
Cl	**95.75**	0.0	0.0	3.14	1.13	0.0
P&R	1.45	**90.62**	0.0	4.58	3.35	0.0
M&P	7.18	10.19	**78.63**	0.0	2.52	1.48
Ele	0.0	3.28	4.46	**87.02**	3.85	1.39
Wo	6.47	7.82	0.0	2.86	**80.85**	0.0
J&B	8.14	2.62	0.0	3.32	0.0	**85.92**

VII. CONCLUSIONS

In this paper, first we analysed the validity of timbral texture features. The validity criterion is determined by the normalized standard deviation of each feature. In the second stage, the frame-wise features have been integrated by using central moments including mean, standard deviation, skewness and kurtosis. Also, we propose the covariance components between timbral texture frame-wise features to be included for improving the classification performance. By considering these feature values, several experiments have been performed separately to analyse classification accuracy among the different feature sets. The ELMs combined with bagging is used to build the classifier. The ELM is an unstable classifier; therefore ELMs with bagging improved the classification accuracy as well as the generalization performance.

The classification accuracy of both datasets (GTZAN and ISMIR2004) is shown in table IV and V, respectively. According to our proposed method, the classification accuracy of 85.15% is achieved in GTZAN datasets. The experimental results on the ISMIR2004 genre datasets have also shown that our proposed approach achieves higher classification accuracy (86.46%) than the ISMIR Music Genre Classification Contest with classification accuracy (84.07%) competitive with Chang-Hsing Lee (86.83%).

Experimental results show that there are no significant differences seen while considering seven and twelve MFCC coefficients for genre classification. It concludes that the minimum feature dimension is also sufficient for music genre discrimination. In addition, our experiment shows that covariance components have a significant impact in improving the genre classification. By adding the components we could improve approximately 5% of the overall accuracy. We expect that a more accurate classifier can be constructed with more features added such as segment-based ones after partitioning audio data into pieces, even though it increases the complexity of the classifier.

Acknowledgement

This work was partially supported by the National Research Foundation of Korea grant funded by the Korean government 2011-0022152 and BK21PLUS project.

REFERENCES

[1] Tao Li, Mitsunori Ogihara, and Qi Li, "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, Canada, 2003

[2] Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Trans.Speech Aud. Processing*, vol. 13,no. 3, pp. 441–450, May 2005.

[3] A. Meng, P. Ahrendt, J. Larsen, L. Hansen, Temporal feature integration for music genre classification, *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2007).

[4] Babu Kaji Baniya, Deepak Ghimire, and Joonwhoan Lee, "Evaluation of different audio features for musical genre classification" *in proc. IEEE workshop on Signal Processing Systems*, Taipei, Taiwan, 2013

[5] Babu Kaji Baniya, Deepak Ghimire, and Joonwhoan Lee, "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, 2014

[6] R. Groeneveld, G. Meeden, Measuring skewness and kurtosis, The Statistician 33 (1984) 391-399

[7] Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans Speech Audio Process*., vol. 10, no.3, pp. 293-302, Jul. 2002

[8] L. Rabiner and B. Juang. Fundamentals of Speech Recognition. Prentice-Hall, NJ, 1993.

[9] B. Logan. Mel frequency cepstral coefficients for music modeling. *In Proc. Int. Symposium on Music Information Retrieval ISMIR*, 2000.

[10] Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle , P. Cano, An Experimental Comparison of Audio Tempo Induction Algorithms, *IEEE Transactions on Speech and Audio Processing*, vol.14, page 1832-44, 2006.

[11] Bergstra, J., Casagrande, N., Erhan, D., Eck, D. and Kegl B. “Aggregate features and AdaBoost for music classification”, *Machine Learning*, Vol. 65, No. 2-3, pp. 473-484, 2006.

[12] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin, “Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features”, *IEEE Trans. of Multimedia*, Vol.11, no. 4, pp. 670-82, June 2009.

[13] Jin S. Seo, Seungjae Lee, Higher-order moments for musical genre classification, *Signal Processing*, vol. 91, Issue 8, pp. 2154-57, 2011

[14] Pampalk, E., Flexer, A. and Widmer, G. “Improvements of audio-based music similarity and genre classification”, *Proceedings of the Sixth International Symposium on Music Information Retrieval*, London, UK, 2005.

[15] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney. Classification of audio signals using statistical features on time and wavelet transform domains. *In Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-98)*, volume 6, pages 3621–3624, 1998.

[16] H. Soltau, T. Schultz, and M. Westphal. Recognition of music types. *In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.

[17] H. Deshpande, R. Singh, and U. Nam. Classification of music signals in the visual domain. In Proceedings of the COST-G6 *Conference on Digital Audio Effects*, 2001.

[18] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1997, pp.1331-1334.*

[19] T.Zhang and C.J.Kuo, “Audio Content Analysis for Online Audiovisual Data Segmentation and Classification,” *IEEE Trans. Speech and Audio Processing*, Vo1.9, No.4, pp. 441-457, May 2000.

[20] H. Terasawa, M. Slaney, and J. Berger. “Perceptual distance in timbrals space”. *In Proceedings of Eleventh Meeting of the International Conference on Auditory Display*, pages 61-68 Limerick, Ireland, July 2005.

[21] Beth Logan, Mel Frequency Cepstral Coefficients for Music Modeling, Int. *Symposium on Music Information Retrieval* (2000)

[22] I. Daubechies, “Orthonormal bases of compactly supported wavelets,” Commun. Pure Appl. Math, vol. 41, pp. 909-996, 1988

[23] K. Hornik, “Approximation capabilities of multilayer feedforward networks,*” Neural Networks*, vol. 4, pp. 251—257, 1991.

[24] Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: Theory and Applications. *Neurocomputing* 2006, 70, 489-501.

[25] Beriman, L. Bagging Predictors. *Machine Learning* 1996, 24, 123-140.

[26] Marasys, “Data sets” http://marsysas.info/download/data

[27] [Online]. Available: http://ismir2004.ismir.net/ISMIR_Contest.html

BIOGRAPHIES

Babu Kaji Baniya received the B.E. degree in Computer Engineering from Pokhara University, 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 etc.

Deepak Ghimire received the B.E. degree in Computer Engineering from Pokhara University, Nepal in 2007 and M.S. degree in Computer Science and Engineering from Chonbuk National University, Rep. of Korea in 2011. 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 image processing, computer vision, pattern recognition, facial emotion analysis 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.