Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches

Article *in* GSTF Journal on Computing (JoC) · August 2013

DOI: 10.7603/s40601-013-0014-0

2 authors:

Devindu Chathuranga
Sri Lanka Institute of Information Technology
2 PUBLICATIONS **25** CITATIONS

SEE PROFILE

Lakshman Jayaratne
University of Colombo
52 PUBLICATIONS **128** CITATIONS

SEE PROFILE

DOI 10.7603/s40601-013-0014-0

Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches

Y.M.D. Chathuranga
University of Colombo School of Computing
Sri Lanka
dhanithc@gmail.com

K.L. Jayaratne,
University of Colombo School of Computing
Sri Lanka
klj@ucsc.cmb.ac.lk

***Abstract*—Musical genre classification is put into context by explaining about the structures in music and how it is analyzed and perceived by humans. The increase of the music databases on the personal collection and the Internet has brought a great demand for music information retrieval, and especially automatic musical genre classification. In this research we focused on combining information from the audio signal than different sources. This paper presents a comprehensive machine learning approach to the problem of automatic musical genre classification using the audio signal. The proposed approach uses two feature vectors, Support vector machine classifier with polynomial kernel function and machine learning algorithms. More specifically, two feature sets for representing frequency domain, temporal domain, cepstral domain and modulation frequency domain audio features are proposed. Using our proposed features SVM act as strong base learner in AdaBoost, so its performance of the SVM classifier cannot improve using boosting method. The final genre classification is obtained from the set of individual results according to a weighting combination late fusion method and it outperformed the trained fusion method. Music genre classification accuracy of 78% and 81% is reported on the GTZAN dataset over the ten musical genres and the ISMIR2004 genre dataset over the six musical genres, respectively. We observed higher classification accuracies with the ensembles, than with the individual classifiers and improvements of the performances on the GTZAN and ISMIR2004 genre datasets are three percent on average. This ensemble approach show that it is possible to improve the classification accuracy by using different types of domain based audio features.**

Keywords- Music genre classification; Features extraction; Machine learning; Ensemble classification; Feature selection

I. INTRODUCTION

Music can be divided into many categories mainly based on rhythm, styles and cultural background. The styles are what we call the music genres. Musical genres are categorical labels created by human experts and it used for categorizing, describing and even comparing songs, albums, or authors in the vast universe of music [1]. There is a number of top-level or song-level perceptive descriptions, such as instrumentation, genre, mood and artist. Musical genre is the one of the main top-level descriptors and it encapsulates semantic information of the given music piece. Different genres differ from one another in their pitch content, instrumentation, rhythmic structure and timbre features of the music [2].

Nowadays, a personal music collection may contain thousands of songs, while professional collections typically contain hundreds of thousands. Most of the current music databases are indexed based on the artist's name or the title of the song. When songs are indexed improperly in the database, it can cause unexpected search results. Browsing and searching such large collections of songs are very difficult while associating a genre to a musical piece would be more user friendly, in finding what they are looking for. Most music listeners may only be interested in certain types of music. Therefore, music genre classification system would enable them to search for the music they are interested in.

Traditionally, the music genres are labeled by human musical experts. As the number of manually generated rules increases, it may produce unexpected interactions and side effects. Expert classification process is mostly implemented without following a universal taxonomy and this labeling process to audio indexing is prone to error. Human perception of music is dependent on a variety of personal, cultural and emotional aspects. Therefore its genre classification results may avoid clear definition and the boundaries among genres are fuzzy [3,4]. However, there are no any complete agreement exists in their definition and strict distinguishing boundaries among genres. Automatic music genre classification is the classification of music into genres by a machine and as a research topic, it mostly consists of the selection of best features and development of algorithms to perform this classification. A key problem in music classification is how to efficiently and effectively extract low level audio features for high level classification.

In this paper we present a novel approach for automatically classifying audio signal into a hierarchy of musical genres using ensemble of classifiers and comparative study using different machine learning algorithms. Aim of this project is increase accuracy of genre classification result in more robust way. We proposed frequency domain, temporal domain, cepstral domain and

Published online: 16 August 2013

modulation frequency domain audio features and two types of feature vectors are designed for individual classification according to short term and long term based audio features. For feature selection purposes, wrapper method is used for short term feature vector and filtering method is used for long term feature vector. Feature selection is significantly influence the final classification accuracy and the best features are always producing the most accurate results for a task with the least computational expense. The support vector machine classifier with polynomial kernel function is employed as the base classifiers for each of the feature vectors to infer the genre. Then novel late fusion weighted combination ensemble method is employed for produce the final class label and it outperform the trained fusion functions for proposed feature vectors.

The paper is structured as fallows.

A brief overview on related work is provided in section II. Feature extraction and the four domains of features are described in section III. Section IV deals with feature selection, the automatic musical genre classification using different machine learning algorithms and evaluation of the proposed ensemble methods and finally, Section V provides conclusions and suggest potential directions for future improvements.

II. Related Work

Automatic musical genre recognition does not have a long history but there has been a lot of interest in the recent ten years. It is quite interdisciplinary and draws especially from areas such as digital signal processing, machine learning, and music theory.

A normal machine learning approach is to use support vector machines (SVM) with some kernels. Anan et al. proposed an alternative approach to music genre classification [4]. They employed a (dis)similarity-based learning framework. They convert MIDI format into string data of three types such as Pitch string, Rhythm string and Note string for dissimilarity measure. Computational experiments show that their method combined with string kernels such as the n-gram and the mismatch kernels outperform SVM with all kernels. But it is impractical because of the music genre classifiers always need all MIDI recording of the corresponding audio files and accurate polyphonic transcription system is a much harder problem than genre classification.

One of the most significant proposals specifically to deal with studies on automatic musical genre classification was proposed by Tzanetakis and Cook in 2002 [2]. In that paper, the researchers use timbral related features, texture features, pitch related features based on the multi-pitch detection algorithm and the rhythmic content features based on Beat Histogram. For classification and evaluation, the authors propose the Gaussian Mixture Model (GMM) classifiers and k Nearest Neghbor (KNN) classifiers. The overall genre classification accuracy of the system reaches a 61% of correct classifications over the 10 musical genres.

Sound analysis process was used for different sound representation techniques such as waveform, spectrum and spectrogram for the different purpose. Costa et al. proposed an alternative approach for musical genre classification which is based on texture images [5]. They convert the audio signal into spectrograms and then extract features from this visual representative image, which are divide into zones so that features can be extracted locally. However, larger musical structures other than the instantaneous surface features are difficult to identify by only viewing a spectrogram.

Previously most of the music classification researchers worked on fusion of feature subspaces [6]. Lately some approaches were built on classifier ensemble techniques, which fusion of the genre labels assigned separately by each single classifier [7]–[10]. Music is an inherently multi-modal type of data and Mayer et al. proposed an approach for multi-modal classification of music using classifier ensemble techniques [8]. Mayer et al. presented on how the lyrics domain of music combined with the acoustic domain. Using late fusion rules for combining classifiers outcomes, they have created a Cartesian classifier which is two dimensional ensemble systems and it combines different feature subspaces from different domains and different classification algorithms. Using the ensemble approach, they achieved better results than using single algorithm on a single feature set alone.

Silla et al. presented ensemble method, which combines the multiple feature vectors extracted from the beginning, middle and end parts of 30 second music segments [9]. In this research we tried to combine short term and long term based features of a music piece, using our novel ensemble approach.

A novel feature selection and extraction method was proposed by many researchers for the musical genre classification [2], [11], [12]. Matsui et al. investigated a novel method of musical feature extraction [12]. They used novel gradient-based musical features which extracted using a SIFT algorithm. This feature can effectively capture the local dynamic information in the logarithmic frequency domain. They randomly select samples and represent them as 2D spectrogram images. Then they extract SIFT key points for each image. Using SVM with a linear kernel as their classifier, they examined the temporal and frequency independence of their method through comparison with a simple method based on the GMM with MFCC feature vectors. The experimental results confirm that the SVM method together with their novel feature is robust to variations in tempo, pitch and also have the frequency independent and the time independent characteristics.

Texture window was firstly introduced for musical genre classification by Tzanetakis and Cook [2]. They used variances and means to capture the long term features of

sound texture. A novel approach to musical genre classification using temporal information was proposed by Tao et al [13]. In that paper introduced seven different temporal evolution descriptors are used in the texture window such as minimum, maximum, mean, standard deviation, temporal kurtosis and temporal centroid. The experimental results show that using only mean and standard deviation achieves the best accuracy and when adding more temporal evolution descriptors, it does not cause any improvement of the overall classification accuracy. They showed that standard deviation and mean are simple but powerful for discriminating different music genres.

Viglensoni et al. tried to improve musical genre classification performance using cultural feature [14]. They extracted Information of cultural features from both web searches and mined listener tags using Yahoo! and Last.fm web services. Using the SAC (Symbolic, Audio and Cultural) dataset they achieved some improvements but in the 10-genre classification experiments using social tags which performed the worst amongst all configurations and also observed a 17.7% decrease in performance.

McKay et al. tried to improve musical genre classification performance using lyrical feature [15]. They investigated the genre classification utility of combining features extracted from symbolic, audio, lyrical and cultural sources of musical information. The experimental results show that features extracted from lyrics were less effective than the other feature types.

III. Feature Extraction

Audio features can mainly divide into two levels as top-level and low-level according to perspective of music understanding [16]. The top level labels provide information on how listeners interpret and understand music using different genres, moods, instruments, etc. Low-level audio features can also be categorized into short-term and long-term features on the basis of their time scale [17]. Figure 1 characterizes audio features from different levels and perspectives. Most of the features that have been proposed in the literatures are short-time timbre features, which only consider the immediate frequencies and extract the characteristics of the audio signal in a 10-30ms duration small sized window. Long-term features such as rhythm and beat features contain the structural information and normally extracted from the local windows on the large time-scale full song or a sound clip.

Whatever the format music is stored its data can be decoded and transformed into a succession of digital samples to represent the waveform. But this data cannot be used directly by automatic systems because pattern matching algorithms cannot deal with such an amount of information and formats. So it is necessary to extract some features that describe the audio wave using a compact representation. The set of musical features representing short time features of music was extracted from the audio using the Marsyas (Music Analysis, Retrieval and Synthesis for

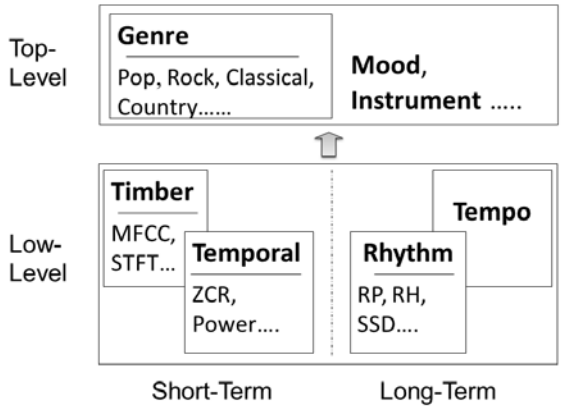

Figure 1. Characterization of Audio Features Based on Raw Digital Audio Signals according to perspective of music understanding

Audio Signals) framework. Marsyas is an open source software framework for audio processing with specific emphasis on MIR applications. We proposed four domains of features for representing short term and long term features.

A. Temporal Domain features

The temporal domain is the native domain for represents the signal changes over the waveform. All temporal features are extracted directly from the raw audio signal without any preceding transformation using Marsyas framework [2].

1) Time Domain Zero Crossings (ZCR): The ZCR measures the noisiness of the sound by computing the number of times the audio waveform crosses the zero axis per time unit. A zero crossing occurs when adjacent audio samples have different signs. The following Equation 1 shows the calculation of Time Domain Zero Crossings.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} \mid sign(x[n]) - sign(x[n-1]) \mid \qquad (1)$$

Where the sign function is 0 for negative arguments and 1 for positive arguments and x[n] denotes the time domain signal for frame t.

2) Amplitude based features: Amplitude based features directly computed from the amplitude or pressure variation of a signal and represents the temporal envelope of the audio signal over time. The MPEG-7 audio waveform descriptor gives a compact description of the shape of a waveform by computing the minimum and maximum amplitude values within successive non overlapping frames.

3) Power based features: The energy of a signal is the square of the amplitude represented by the waveform. The

power of a sound is the energy transmitted per unit time. Short term energy features are measured to discriminate voiced, unvoiced and silence of the song. The following Equation 2 shows the calculation of Short time energy.

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

Where w denotes the window, n is the sample that the analysis window is centered on and N is the window size.

B. *Frequency Domain features*

The frequency domain features are the largest group of audio features used in our short-term feature set. The most popular methods to obtain frequency domain features are the Fourier transform and autocorrelation.

1) Spectral Flux (SF): The SF is defined as the squared change in normalized amplitude of successive spectral distributions between two consecutive time frames. It is often used as an indication of the degree of change of the spectrum between two adjacent frames. The following Equation 3 shows the calculation of Spectral Flux.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \qquad (3)$$

Where N_t [n] and N_{t-1} [n] stand for the magnitude of spectrogram at frequency bin n for current frame t and previous frame display style *t-1* respectively. Signal with nearly constant or slowly varying spectral properties have low SF values and signals with abrupt spectral changes have high SF values.

2) Brightness: Characterizes the spectral distribution of frequencies and describes whether a signal is dominated by high or low frequencies, respectively. A sound becomes brighter as the low-frequency content becomes less dominant and the high-frequency content becomes more dominant.

a) Spectral Centroid: Spectral Centroid (SC) is commonly associated with the measure of the shape or brightness of a sound by calculating the weighted average frequency of every time frame. The spectral centroid is defined as the "center of gravity" of a Short Time Fourier Transform (STFT) using the Fourier transform's frequency and magnitude information. The following Equation 4 shows the calculation of Spectral Centroid.

$$C_t = \frac{\sum_{n=1}^{n=N} M_t[n] \times n}{\sum_{n=1}^{n=N} M_t[n]} \qquad (4)$$

Where M_t[n] represents the magnitude of STFT spectrogram at frame t and frequency bin n. The spectral centroid is a measurement of the spectrogram shape and the centroid is usually a larger value than one might intuitively expect, because there is so much more energy in the high frequency bands.

3) Tonality: Tonality is a property of music in which specific hierarchical pitch relationships are based on a key or tonic. Tonality is related to the pitch strength and music with distinct components tends to produce larger pitch strength than music with continuous spectra.

a) Spectral Rolloff: Spectral rolloff point is defined as the boundary frequency where 85% of the energy distribution in the spectrum is below this point. The following Equation 5 shows the calculation of Spectral Rolloff.

$$\sum_{n=1}^{n=R_t} M_t[n] = 0.85 \times \sum_{n=1}^{N} M_t[n] \qquad (5)$$

Where the spectral Rolloffs frequency point R_t determines where 85% of the window's energy is achieved and M_t[n] is the magnitude of the Fourier transform at frame t and frequency bin n.

4) Pitch: Pitch is a major auditory attribute of musical tones, together with loudness, timbre and duration. Pitch is closely related to frequency but both are not equivalent. Frequency is an objective and scientific concept, although pitch is subjective. Pitch depends to a lesser degree on the sound pressure level such as loudness, volume of the tone.

The pitch histogram describes the pitch content features of a signal based on multiple pitch detection techniques in a compact way. Pitch content features were originally introduced by tazanetakis et al. for musical genre classification [2].

5) Chroma: Human perception of pitch is periodic in the sense that two pitches are perceived as similar in color but differ by an octave. Based on octave a pitch can be separated into two components which are referred to as chroma and tone height. Chroma features show a high degree of robustness to correlate closely in the musical aspect of harmony and variations in timbre.

The chromagram is a spectrogram that represents the spectral energy of each of the twelve pitch classes by maps all frequencies into one octave. Chromas set consists of the twelve pitch spelling attributes: A, A#, B, C, C#, D, D#, E, F, F#, G, G# as used in Western music notation. Chroma is a pitch based feature that projects the frequency spectrum into 12 bins, with one bin for each of the 12 distinct pitches of the chromatic musical scale. The conversion of an audio music into a chromagram representation can be performed

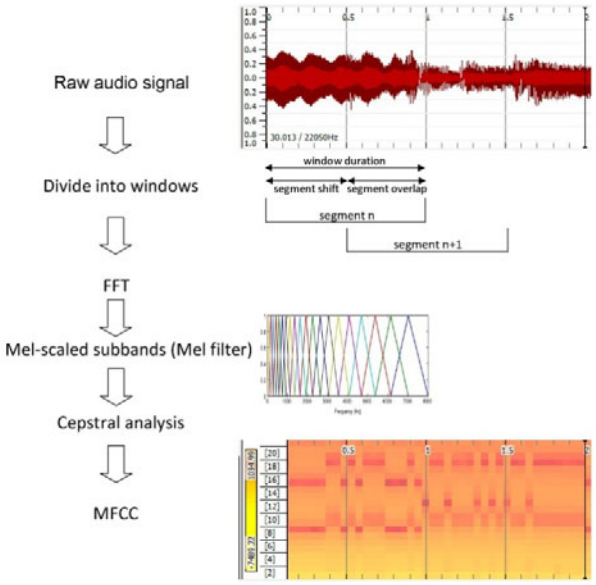

Figure 2. The MFCC Feature Extraction Procedure

by using Short Time Fourier Transform (STFT) in combination with binning strategies [18].

C. Cepstral Features

Cepstral features are frequency smoothed representations of the logarithm of the estimated spectrum of a signal and capture pitch and timbral characteristics.

1) Mel-frequency Cepstral Coefficients: Mel Frequency Cepstral Coefficients (MFCCs) are compact, short time descriptors of the spectral envelope audio feature set and typically computed for audio segments of 10-100ms. MFCC are one of the most popular set of features used in pattern recognition. MFCC was originally developed for automatic speech recognition systems, lately have been used with success in various musical information retrieval tasks [19]. Although this feature set is based on human perception analysis but after calculated features it may not be understood as human perception of rhythm, pitch, etc. Figure 2 illustrates the different steps in the calculation from raw audio signal to the final MFCC features. We selected first 13 MFCC features for our timber feature vector. Normally violin's sounds has much higher values in the third and fifth MFCC than the flute and the fork so mel frequency information may be better suited to discriminate between the different sound sources or different instruments.

D. Modulation Frequency Domain features

Modulation frequency features capture the modulation information of low frequency in the music audio signals. Modulation frequency information is a long term signal variation of frequency that is usually captured by a temporal analysis of the spectrogram. Tempo and rhythm are aspects of musical features which are strongly related to long term

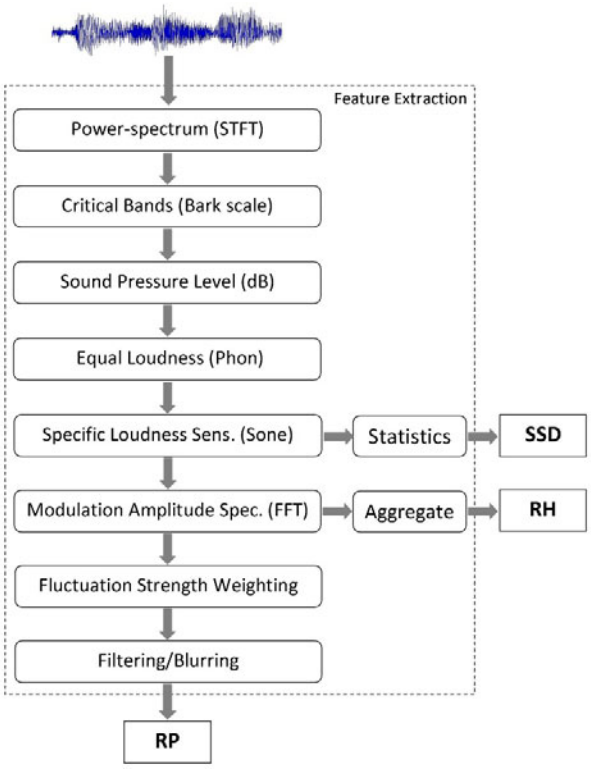

Figure 3. Block Diagram of Rhythm Feature Extraction Method

modulations. We created separate feature vector for modulation frequency domain features.

1) Rhythm features: Rhythm is the timing pattern of musical sounds and silences. These musical sound and silences are put together to form a pattern of regular or irregular pulses caused in music by the occurrence of weak and strong melodic and harmonic beats to create a rhythm. Figure 3 block diagram illustrates the different steps in the calculation from raw audio signal to the final Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram features. These three types of audio features are extracted using the Java audio feature extraction packages.

a) Rhythm Patterns (RP): Pampalk et al. initially proposed RP for music similarity retrieval. The RP contains information about how fast and strong beats are played within the respective frequency bands [20]. RP describe modulation amplitudes for a range of modulation frequencies on a number of frequency bands of the human auditory range. The RP feature extraction process mainly contains two stages. In the first stage, the specific loudness sensation in different frequency bands is computed by using a STFT. In the second stage apply Fast Fourier Transform (FFT) to the sone representation for transform the spectrum into a time invariant representation of the 24 critical bands based on the modulation frequency [21]. In our feature vector 1440 features of RP calculated by using 24 critical bands × 60 modulation frequencies.

b) Statistical Spectrum Descriptor (SSD): During feature extraction process of RP, SSD for the 24 critical bands can be extracted. According to the occurrence of beats or other rhythmic variation of energy on a specific critical band, statistical measures such as median, mean, variance, kurtosis, skewness, minimum and maximum values are used to describe the audio rhythm features. In our feature vector, 168 features of SSD were calculated by using 24 critical bands × 7 statistical moments.

c) Rhythm Histogram (RH): In RH features we can use a descriptor for general rhythmic in an audio music file [21]. The RH features are calculated by taking the median of the histograms of every 6 second segment processed. In our feature vector, 60 features of RH calculated by modulation frequencies are grouped into 60 bins.

IV. Evaluation

A. Feature Selection

In machine learning Feature Selection (FS) is the technique of selecting a subset of relevant features for building robust learning models by removing most redundant and irrelevant features from the feature vector. Relevant features have an influence on the output result and their role cannot be assumed by the rest. The main goal of FS is to determine minimal feature subsets, without affecting the high accuracy in representing the original features. FS method is extremely useful in reducing the dimensionality of the feature vector to be processed by the SVM classifier, improving predictive accuracy by removing irrelevant features or noise data, and speeding up the running time of the learning algorithms.

Feature selection algorithms in general can be broadly split into wrapper methods and filter methods. Filter method tries to rank the list of features by exploiting the intrinsic characteristics of the training data that are relatively independent of the learning algorithm. Wrapper method requires one predetermined learning algorithm and use the classifier as a black box to score the subsets of features. Wrapper methods generally result in better performance than filter methods for the reason that it features selection process is optimized and finds features better suited to the predetermined learning algorithm. In this research we used wrapper method based on SVM for feature selection of short-term feature vector because it contained a few number of features. However, they are generally more computationally expensive if the number of features is large. Therefore in this study wrapper method was not considered for modulation frequency domain feature vector.

We used Classifier Subset Eval along with Best First Search method for short-term feature selection and Info Gain Attribute Eval filtering method for long term feature selection. Classifier Subset Eval is one of the wrapper methods which uses a SVM classifier and evaluates sets of attributes on the training data. The Best First search method is a heuristic algorithm that Searches the space of feature subsets by greedy hill climbing with a stepwise regression facility. It makes at each stage the local optimum choice with the hope of finding the global optimum. The Info Gain Attribute Eval filtering method evaluates the worth of an attribute by measuring the information gained in respect of the genre class.

B. Classification

1) Base Classifiers

In this research we used SVM with polynomial kernel function as the base classifier for genre classification using Weka (Waikato Environment for Knowledge Analysis) library. The kernel function plays the role of the dot product in the feature space and it transforms the data into a higher dimensional space to make it possible to perform the separation so the kernel mapping is a very powerful concept which allows SVM models to perform separations easily even with very complex decision boundaries. In the current research we used polynomial kernel as kernel function for SVM which gives better results than any other kernels for our feature vectors. The polynomial kernel is directional and output depends on the direction of the two vectors in low dimensional space So all vectors with the same direction will have a high output from the polynomial kernel. We used one against one technique and individual classification was done by a max wins voting strategy.

2) Boosting Classifiers

To further improve the performance of classification rules, a number of combining techniques such as Adaboost can be used. Most popular techniques for constructing ensembles are boosting and bagging algorithms. Normally, boosting algorithms often gives a better prediction performance than bagging algorithms so we tried to use the approach of boosting for improve the existing performance of our method. We used AdaBoost.M1 for the multi class classification. AdaBoost algorithm can efficiently convert a weak learning algorithm into a strong learning algorithm and the actual performance dependent on the base learner and the data. It is possible to improve the performance by using SVM as a weak learner for AdaBoost [22].

3) Ensemble Classification

Generally a single classifier is used to determine which genre class belongs to a given style. The ensemble classification approach is the process of using an ensemble of classifiers in order to provide a unified and single classification output from a song into a single genre decision. There are two major ways for ensemble classification. In the first approach the information is fused in the early or later during the classification phase. In the second approach the classification of one classifier's genre label is selected according to the some criterion. Information Fusion integrates data and knowledge from

multiple sources which are potentially more accurate and more efficient than using means of a single source. Early fusion and late fusion are two subsets of fusion techniques that differ in the way they integrate the results from feature extraction on the various modalities [23]. In this research we applied late fusion for combining classifier outcomes

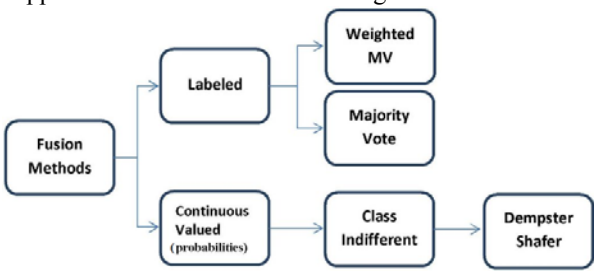

Figure 4. Classifiers Fusion Methods

rather than features. Late fusion method generally apply after classification of each genre classes, using individual modalities and then it merges the scores of individual classifiers for detecting final genre labels.

We can roughly characterize classifier combination methods into two groups based on the forms of classifier outputs as shown in Figure 4. The first group combination of decisions is performed on single genre class labels. Amongst the methods of first group the weighted majority vote is by far the most popular and simple approach. The second method is concerned with the utilization of probabilities corresponding to genre class labels and class indifferent method use as much as possible information obtained from the sets of classes in calculating the support for each genre class using rank based method of DST.

a) Dempster-*Shafer Theory (DST):* Once an ensemble of classifiers has been created, Dempster-Shafer Theory of evidence is an effective way to improve the performance of the fusion functions of label outputs. Dempster-Shafer Theory of evidence is one of the more complex approaches for classifier combination. DST also known as the theory of belief functions is a generalization of the Bayesian theory of subjective probability and unlike the Bayesian theory, the DST allows each source to contribute information in different levels of detail. The DST is based on the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster-Shafer rules for combining such degrees of belief when they are based on independent items of evidence [24]. We tried to combine classifier outputs using Dempster-Shafer rules for computing belief functions for evidence combination, which can be trained using available training samples.

b) Classifier Weighing: The weighing system makes multiple classifiers more robust to the choice of the number of individual classifiers. Dynamic weighting and static weighting are two approaches to weighting of classifiers [25]. The dynamic weights are assigning to the individual classifiers which can change for each test pattern. We used static weighting system for the ensemble classification which weights are computed for each classifier in the training phase and then maintained constant during the classification of the test patterns. The weight of each genre classifier can be set proportional to its accuracy performance on a genre training set. Moreno-Seco et al. proposed re-scaled weighted vote, best-worst weighted vote and quadratic best-worst weighted

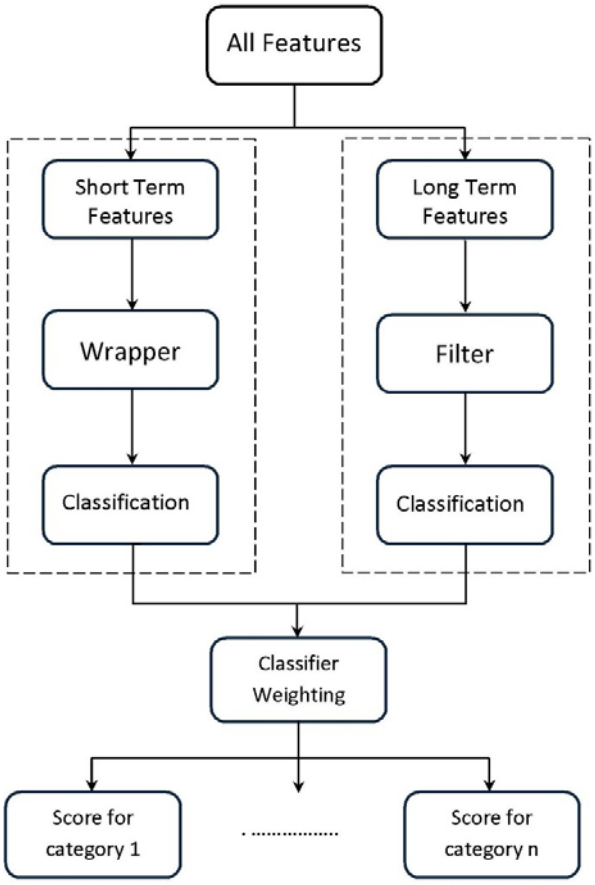

Figure 5. Proposed Classifier Ensemble System using Classifier Weighting

voting approaches for performance weighting [26]. We select the weights for the classifiers are formalized through the following theorem Equation 6.

$$w_i = log\frac{a_i}{1-a_i} \quad (6)$$

It is consider an ensemble of two independent classifiers with individual accuracies a_1 and a_2. The outputs are combined by the weighted majority vote and the accuracy of the ensemble is maximized by assigning weights. The overall scheme of our proposed ensemble classification system is shown in Figure 5.

C. Results

In our experiments we used the GTZAN dataset and ISMIR2004 genre dataset. GTZAN and ISMIR datasets are commonly and widely used as a standard reference collection in genre classification studies and human evaluations were made of both. GTZAN dataset has the following ten classes: blues (bl), classical (cl), country (co), disco (di), hiphop (hi), jazz (ja), metal (me), pop (po), reggae (re) and rock (ro). Each genre contains 100 audio recordings of 30 seconds long. The number of songs for each genre is similar and it is a well-balanced dataset.

Table I
Short-Term feature set Confusion Matrix: Before Feature Selection

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**78**	0	5	2	0	1	6	0	3	3
cl	2	**95**	1	0	0	1	0	0	0	1
co	5	2	**69**	2	0	1	3	1	1	16
di	3	0	5	**70**	3	0	0	1	7	11
hi	4	0	0	5	**73**	0	2	4	11	1
ja	0	6	2	0	2	**85**	0	1	0	4
me	3	0	2	1	0	2	**85**	0	0	7
po	3	0	7	5	3	1	1	**71**	3	6
re	4	0	6	6	12	1	0	6	**60**	5
ro	5	1	10	9	3	1	7	2	5	**59**

ISMIR2004 genre dataset contains 1458 full audio songs using the following six popular musical genre classes: Classical, Electronic (el), Jazz/Blues, Metal/Punk (pu), Rock/Pop and World (wo). All experiment reported in this research were performed using 10 fold cross validation method, in which the dataset was split into 10 parts of equal size. 90% of the songs were used to train the classifier and the rest 10% of the songs were used to test it. This was done 10 times, once for each fold and songs of each of these groups determined randomly on a genre label so classification accuracy will not be biased.

In this research we created two different feature vectors. The first feature vector for describing Short term features consists of the proposed frequency domain, the temporal domain and the cepstral domain audio features resulting in a 124-dimensional feature vector. The second feature vector for describing long term features consists of the proposed modulation frequency domain features resulting in a 1602-dimensional feature vector. Short-term feature vector used wrapper method for feature selection task and long term feature vector used filtering approach for feature selection task. All feature selection tasks were done using WEKA (Waikato Environment for Knowledge Analysis) machine learning toolkit. WEKA contains a collection of machine learning algorithms for data mining tasks.

A confusion matrix is a specific table layout that contains information about actual and predicted classifications done by a genre classification system. In the confusion matrix, each column represents the instances in a predicted genre class, while each row represents the instances in an actual genre class. Table I shows the confusion matrix of Short-term feature vector before apply wrapper feature selection method.

In Short-term feature vector, Classifier Subset Eval alone with Best First Search method was employed to select an optimal subset of 105 dimensional from the initial 124 dimensional feature vector. It was reduced 15.3% number of features by removing most redundant and irrelevant features from the feature vector.

Table II
Short-Term feature set Confusion Matrix: After Feature Selection

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**83**	0	4	1	0	1	5	0	3	3
cl	0	**99**	1	0	0	0	0	0	0	0
co	8	1	**75**	1	0	1	1	2	1	10
di	4	0	4	**71**	1	1	0	1	6	12
hi	2	0	0	5	**68**	0	3	8	13	1
ja	3	5	2	2	0	**85**	0	0	0	3
me	3	0	0	2	1	0	**87**	0	0	7
po	2	0	7	5	1	0	0	**74**	4	7
re	4	0	5	7	13	1	0	2	**64**	4
ro	6	0	9	7	2	0	7	3	7	**59**

Table III
Long Term feature set Confusion Matrix

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**86**	0	4	2	1	2	1	0	1	3
cl	0	**94**	1	0	0	2	0	0	1	2
co	8	0	**68**	3	0	2	0	2	3	14
di	3	1	1	**63**	5	1	4	7	4	11
hi	3	0	2	6	**77**	1	1	5	5	1
ja	2	7	5	0	2	**81**	1	1	0	1
me	1	0	1	1	0	0	**88**	1	0	8
po	0	0	7	7	2	1	0	**73**	4	6
re	5	0	3	7	11	4	1	3	**64**	2
ro	12	1	15	8	2	3	6	6	0	**47**

Table II shows the confusion matrix of Short-term feature vector after applying wrapper feature selection

method. The performance and accuracy of SVM classifier were improved after applying wrapper feature selections methods to the dataset. But Disco and Rock genre class's accuracies did not change significantly. As seen in the Table II, Classical music has the highest classification accuracy of 99% and Rock music has the worst classification accuracy. Normally Rock music has broad nature so it can easily confuse with other genres [2].

Our second feature vector initially had 1602 different features. CfsSubsetEval feature selection method took about 7 days to terminate and generate the optimal subset of our short term feature vector. So normally the number of features becomes very large, and the filter method is usually chosen due to its computational efficiency. We used Info Gain Attribute Eval filtering method for feature selection in the long term feature vector. It removed 103 features of most redundant and irrelevant from the feature vector. Table III shows the confusion matrix of long term feature vector after applying filtering feature selection method.

Table IV
Performance of Individual Long Term Feature Sets

Feature Set	GTZAN Dataset
RH	46.4%
SSD	71.4%
RP	63.8%

As it can be seen from Table III, sixth row Hip Hop music has the highest classification accuracy when compared with Table II Short-term features. Typically, hip hop music consists of intensely rhythmic beats and rapping parts. Our long term feature vector mainly consists of rhythm, tempo and beat features, so using our second feature vector; Hip hop genre can be detected accurately than other features. As seen in the last column of Table III, accuracy of the rock music is below 50%. Rock is a kind of music with a very strong beat and simple tunes that are usually played loudly. So identification of such features using rhythmic features may be a difficult task. Rock music is mostly misclassified as country and blue. This is due to the facts that rock music and country music have similar roots and rock music came from a combination of country music and rhythm and blues. If both rock and country music have similar instrumentations and rhythm, then it can be difficult to distinguish a country song from a rock song.

Table IV shows the individual performance of the different rhythm and tempo based feature set for the task of the musical genre classification. As can be seen, RH and RP perform worse than the SSD features. SSD feature set consist 168 different features and statistical measures such as median, mean, variance, kurtosis, skewness, minimum and maximum values are used to describe the audio rhythm features so it is performed better than other feature sets.

1) Boosting Classifiers

Boosting is an ensemble technique that allows us to improve the classification performance of weak classifiers. As it can be seen from the Table V confusion matrix, Compared with previously received both short term and long term feature vectors, SVM classifier as the base classifier of Boosting techniques obtained a less performance results for the GTZAN dataset. Boosting algorithms are performing better when the classifiers are weak and the data do not have much noise. Using classifiers like SVM with polynomial kernel function for proposed features, that is already strong as the base learner in AdaBoost does not seem to provide any advantages so its performance of the classifier can be keep unchangeable or decreases as the number of rounds increases.

Table V
AdaboostSVM: Final Genre Confusion Matrix

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**88**	0	3	2	2	1	1	0	1	2
cl	0	**96**	1	0	0	2	0	0	0	1
co	7	1	**69**	4	1	1	0	2	1	14
di	1	1	3	**69**	2	0	2	6	3	13
hi	2	0	0	9	**73**	0	2	5	8	1
ja	2	5	2	0	0	**89**	1	0	0	1
me	0	0	1	1	0	0	**89**	1	0	8
po	0	0	7	4	3	0	0	**76**	3	7
re	7	0	3	7	8	1	1	3	**67**	3
ro	11	1	12	9	1	1	6	6	1	**52**

Table VI
One-Vs.-All Classification: Final Genre Confusion Matrix

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**96**	0	0	0	0	1	2	0	1	0
cl	4	**95**	0	0	0	1	0	0	0	0
co	51	0	**48**	0	0	0	0	1	0	0
di	86	0	0	**11**	0	0	0	0	3	0
hi	41	0	0	0	**52**	0	1	3	3	0
ja	19	5	2	0	0	**74**	0	0	0	0
me	23	0	0	0	0	0	**77**	0	0	0
po	34	0	1	0	7	0	0	**57**	1	0
re	60	0	0	2	7	1	0	1	**29**	0
ro	89	0	3	0	0	1	7	0	1	**0**

2) Dempster-Shafer Theory

We designed the Dempster-Shafer Theory based on multi-class musical genres classification system using SVM with polynomial kernel function and one-against-all strategy. We noted that there are more errors occur between the Blue and Rock music using this approach. This can be happen because of the genre overlapping problem. Dempster's rule plays a central role in DST if it agrees with the assumption of independence, or distinctness, of the items of information. Generally, musical genre classifications are arrived at in a variety of uncoordinated ways and also most of genres are never accept the mutually exclusive terms. DST functions might perform worse than the simple fusion functions such as majority weighted voting. It has, in fact, been suggested that, given insufficient training samples and pieces of evidence are not independent, simple fusion functions may outperform some trained fusion functions.

Table VII
Music Genre Classification Accuracy: Using Different Late Fusion Strategies

Late fusion strategies	Classification accuracy
Majority vote rule	77.5%
Product rule Simple	77.2%
weighted vote Weighted	77.7%
majority vote	78.0%

As it can be seen from the Table VI confusion matrix, the classification accuracy of the last row Rock music has the lowest accuracy of 0% and majority of 89% of Rock music misclassified as Blue music. This is due to the fact that Rock music and Blue music have similar roots and Rock music came from a fusion of Rhythm and Blues sub-genres. Rock music and Blues music is related to one another and they use very similar instruments but typical Blues music employs more instruments than Rock and pure Rock is said to contain 3 chords only whereas Blues use 12-bar in a 4/4 time signature blues chord progressions. Most musicians associate the Blues with depressing lyrics about loss or loneliness. The Blues can be played on any instrument or with any combination of instruments and it possessed of other characteristics such as specific lyrics and bass lines so as a narrow category other musical genres can be misclassified as Blues. One of the most unique music genres is the Classical music. The confusion matrix shows that classification accuracy of the second row Classical music has the maximum accuracy of 95%, which is very similar result with the long term feature vector but when compared with our proposed ensemble approach for all genres except Blues music has very high classification accuracy.

3) Ensemble Classification

We used a late fusion technique to combine classifier outcomes with a weighted majority voting strategy in order to obtain a consensus output. We investigated the impact of using weighted and unweighed combination rules that make use of the output probabilities provided by SVM with polynomial kernel classifier. Table VII shows the results of GTZAN dataset using the majority vote, product rule, simple weighted vote and weighted majority vote rule to combine the individual classifiers. We can observe that weighted majority vote combination rule have highest performance of the ensemble approach relative to the other combination rules.

Mayer et al. proposed ensemble approach reaches 77.5% of correct classifications using both audio and symbolic domains after FS [7]. But in our research, a 78% average accuracy of combined feature sets was reported in a 10-fold cross validation on the GTZAN data set only using audio domain. The accuracy resulted in for the classifier ensemble method on the GTZAN data set is shown in Table VIII.

Table VIII
Final GTZAN Genre Confusion Matrix

	bl	**cl**	**co**	**di**	**hi**	**ja**	**me**	**po**	**re**	**ro**
bl	**86**	0	2	2	0	1	3	0	1	5
cl	0	**98**	1	0	0	0	0	0	0	1
co	2	0	**73**	2	1	3	1	2	1	15
di	1	0	2	**76**	1	1	0	2	7	10
hi	2	0	1	5	**75**	0	3	3	11	0
ja	3	5	2	0	0	**84**	0	0	0	6
me	1	0	0	2	0	0	**86**	0	0	11
po	0	0	8	4	3	1	0	**74**	4	6
re	3	0	5	7	11	1	0	3	**65**	5
ro	11	0	7	4	1	0	6	4	4	**63**

Table IX
Final ISMIR Dataset Confusion Matrix

	cl	el	ja bl	me pu	ro po	wo
classical	623	0	0	0	0	17
electronic	2	171	0	1	25	30
jazz blues	9	1	27	0	2	13
metal punk	0	1	0	54	33	2
rock pop	4	14	0	6	153	26
world	61	25	1	0	9	148

As it can be seen from the Table VIII confusion matrix, the ensemble classification accuracy of the last row rock music has the highest accuracy when compared with individual classifiers, especially with long term rock accuracy, of which the accuracy of the rock music is below 50%. We also were able to keep maximum accuracy of the classical music without a significant change while using ensemble method. On our GTZAN datasets, we can observe

higher classification accuracies with the ensembles, than with the individual classifiers. The improvements of the performances are three percent on average.

We used ISMIR2004 dataset for another experiment for the purpose of checking the performance of ensemble classifier with another ground truth dataset. The accuracy resulted in for the classifier ensemble method on the ISMIR2004 dataset is shown in Table IX. The confusion matrix shows that Classical music has the maximum accuracy of 94.6%. In this dataset Classical music has approximately 44% of largest number of songs. As can be seen from Table IX, fourth row Jazz & Blues music has lowest classification accuracy of 51.9%. In this dataset Jazz & Blues music has approximately 3.6% of lowest number of songs which use only 5 songs for testing two types of music genres so accuracy results may be poor if the sample size is not sufficiently large. In general, a large test sample size is particularly essential to accurately evaluate a classifier performance with a low error rate. Using ensemble approach, an 80.66% average accuracy was reported in a 10-fold cross validation on the ISMIR2004 dataset. It is compares favorably to other research's performance with the 10 fold cross validation on the ISMIR2004 dataset.

V. Conclusions

In this paper we addressed the problem of musical genre classification from audio signals. Most researchers in this area are very concerned about the classification accuracy. In this research we verify that it is possible to improve the classification accuracy by using machine learning algorithms and different types of domain based audio features together. We have presented an alternative approach for music genre classification based on classifier ensemble techniques and we evaluated the method by musical genre classification on GTZAN dataset and ISMIR2004 dataset. Results showed that use of late fusion methods can improve the classification results in a more robust way than using early fusion approaches. AdaBoost boosting algorithm is perform well if the classifiers are weak but using our proposed features SVM with polynomial kernel function is act as strong base learner in AdaBoost, so its performance of the SVM classifier cannot improve using boosting method. Musical genres are not mutually exclusive and most genres evolved out of other genres so it difficult to categorize within non-fusion genres. The weighted majority voting rule is the simplest method but it is the best method to accomplish the classifiers decision fusion than using trained fusion functions such as Dempster-Shafer Theory of evidence. We have also used a filtering and wrapping algorithms for feature selection in order to create a reduced feature vector. Filtering approach provides the same accuracy as the feature vector containing all features but with a compact representation and wrapper approach provides both high accuracy and compact representation. We used SVM with a polynomial kernel as an individual classifier and using ensemble classifier, music genre classification accuracy has been obtained 78% and 81% on GTZAN dataset and ISMIR2004 respectively. We achieved a better performance than when using any of the individual types of feature sets alone. It also compares favorably with the performance of other authors with the same experimental set up on the same datasets only using audio waveform.

As future work we intend to experiment bigger amounts of data (Million Song Dataset) and develop new features which are able to extract the musically-meaningful information from the audio signals and use more feature sets such as melodic characteristics. All genres of music share commonalities but each has unique characteristics so we intend to identify unique audio waveforms features especially for Rock and Reggae music.

References

[1] D. Jang, M. Jin, and C. Yoo, “Music genre classification using novel features and a weighted voting method,” in 2008 IEEE Int. Conf. on Multimedia and Expo (ICME), (Hannover, Germany), pp. 1377–1380, Jun. 2008.

[2] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Trans. on speech and audio process., vol. 10, no. 5, pp. 293–302, Jul. 2002.

[3] N. Scaringella, G. Zoia, and D. Mlynek, “Automatic genre classification of music content: a survey,” IEEE Signal Pro- cess. Mag., vol. 23, no. 2, pp. 133–141, Mar. 2006.

[4] Y. Anan, K. Hatano, H. Bannai, and M. Takeda, “Music genre classification using similarity functions,” in 12th Int. Conf. on Music Information Retrieval (ISMIR), (Miami, Florida, USA), pp. 693–698, Oct. 2011.

[5] Y. Costa, L. Oliveira, A. Koericb, and F. Gouyon, “Music genre recognition using spectrograms,” in 18th Int. Conf. on Systems, Signals and Image Process. (IWSSIP2011), (Sara- jevo, Bosnia and Herzegovina), pp. 151–154, Jun. 2011.

[6] M. Wu and J. Ren, “Combining visual and acoustic features for music genre classification,” in Proc. 10th Int. Conf. on Machine Learning and Applications (ICMLA), (Honolulu, Hawaii, USA), pp. 124–129, Dec. 2010.

[7] R. Mayer, A. Rauber, P. Ponce de León, C. Pérez-Sancho, and J. Iñ esta, “Feature selection in a cartesian ensemble of feature subspace classifiers for music categorisation,” in Proc. of 3rd int. workshop on Music and Machine learning (MML), (Firenze, Italy), pp. 53–56, Oct. 2010.

[8] R. Mayer and A. Rauber, “Music genre classification by ensembles of audio and lyrics features,” in 12th Int. Conf. on Music Information Retrieval (ISMIR), (Miami, Florida, USA), pp. 675–680, Oct. 2011.

[9] C. Silla, C. Kaestner, and A. Koerich, “Automatic music genre classification using ensemble of classifiers,” in Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics (SMC 2007), (Montreal, Canada), pp. 1687–1692, Oct. 2007.

[10] C. Silla Jr, A. Koerich, and C. Kaestner, "A machine learning approach to automatic music genre classification," J. of the Brazilian Comp. Soc., vol. 14, pp. 7–18, Sep. 2008.

[11] D. Jang, M. Jin, and C. Yoo, "Music genre classification using novel features and a weighted voting method," in 2008 IEEE Int. Conf. on Multimedia and Expo (ICME), (Hannover, Germany), pp. 1377–1380, Jun. 2008.

[12] T. Matsui, M. Goto, J. Vert, and Y. Uchiyama, "Gradient- based musical feature extraction based on scale-invariant feature transform," in Proc. of the 19th European Signal Process. Conf. (EUSIPCO), (Barcelona, Spain), pp. 724–728, Aug. 29-Sep. 2 2011.

[13] R. Tao, Z. Li, Y. Ji, and E. Bakker, "Music genre classification using temporal information and support vector machine," in ASCI Conf., (Veldhoven, The Netherlands), Nov. 2010.

[14] G. Vigliensoni, C. McKay, and I. Fujinaga, "Using jweb- miner 2.0 to improve music classification performance by combining different types of features mined from the web," in 11th Int. Society for Music Information Retrieval (ISMIR), (Utrecht, Netherlands), pp. 607–612, Aug. 2010.

[15] C. McKay, J. Burgoyne, J. Hockman, J. Smith, G. Viglien- soni, and I. Fujinaga, "Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features," in 11th Int. Society for Music Information Retrieval (ISMIR), (Utrecht, Netherlands), pp. 213–218, Aug. 2010.

[16] Z. Fu, G. Lu, K. Ting, and D. Zhang, "A survey of audio- based music classification and annotation," IEEE Transactions on Multimedia, vol. 13, pp. 303–319, Apr. 2011.

[17] D. Chathuranga and L. Jayaratne, "Musical genre classifica- tion using ensemble of classifiers," in IEEE Fourth Int. Conf. on Computational Intelligence, Modelling and Simulation (CIMSim2012), (Kuantan, Pahang, Malaysia), p. 237242, Sep. 25-27 2012.

18] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in Proc. 2003 IEEE Workshop Appl. Signal Process. Audio Acoust.(WASPAA), (New York, USA), pp. 185–188, Oct. 2003.

[19] B. Logan, "Mel frequency cepstral coefficients for music modeling," in Proc. of the 1st Int. Symp. on Music Infor- mation Retrieval(ISMIR), (Plymouth, Massachusetts, USA), Oct. 2000.

[20] E. Pampalk, A. Rauber, and D. Merkl, "Content-based orga- nization and visualization of music archives," in Proc. of the 10th ACM Int. Conf. on Multimedia, (Juan-les-Pins, France), pp. 570–579, Dec. 2002.

[21] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classi- fication," in 6th Int. Conf. on Music Information Retrieval (ISMIR), (London, U.K.), pp. 34–41, Sep. 2005.

[22] X. Li, L. Wang, and E. Sung, "A study of adaboost with svm based weak learners," in Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN'05), (Montreal, Canada), pp. 196–201, Jul. 31-Aug. 4, 2005.

[23] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in Proc. of the 13th Annu. ACM Int. Conf. on Multimedia, (Singapore), pp. 399–402, Nov. 2005.

[24] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Inc., 2004.

[25] J. S. Valdovinos, R. M. Sanchez and R. Barandela, "Dynamic and static weighting in classifier fusion," in Proc. of the 2nd Iberian Conf. on Pattern Recognition and Image Analysis (IbPRIA'05), (Estoril, Portugal), pp. 59–66, Jun. 2005.

[26] F. Moreno-Seco, J. Iñesta, P. de León, and L. Micó, "Compar- ison of classifier fusion methods for classification in pattern recognition tasks," in Proc. of the 2006 joint IAPR Int. Conf. on Structural, Syntactic, and Statistical Pattern Recognition ((SSPR'06/SPR'06), pp. 705–713, Aug. 2006.

Dhanith Chathuranga is currently a fourth year computer science undergraduate at University of Colombo School of Computing(UCSC), Sri Lanka. His research interests include Multimedia Information Management, Audio Signal Processing, Music Information Retrieval, and Machine Learning.

Dr Lakshman Jayaratne - (Ph.D. (UWS), B.Sc.(SL), MACS, MCS(SL), MIEEE) obtained his B.Sc (Hons) in Computer Science from the University of Colombo, Sri Lanka in 1992. He obtained his PhD degree in Information Technology in 2006 from the University of Western Sydney, Sydney, Australia. He is working as a Senior Lecturer at the University of Colombo School of Computing (UCSC), University of Colombo. He has wide experience in actively engaging in IT consultancies for public and private sector organizations in Sri Lanka. At present, he is working as a Research Advisor to Ministry of Defense, Sri Lanka. Also he is the present President of Chapter of Sri Lanka, IEEE. His research interest includes Multimedia Information Management, Multimedia Databases, Intelligent Human-Web Interaction, Web Information Management and Retrieval, and Web Search Optimization