Perceptual feature-based song genre classification using RANSAC

Arijit Ghosal*

CSE Department,
Institute of Technology and Marine Engineering,
24 Parganas, (South), West Bengal, India
Email: ghosal.arijit@yahoo.com
*Corresponding author

Rudrasis Chakraborty

Indian Statistical Institute,
203, Barrackpore Trunk Road,
Kolkata, West Bengal, India
Email: rudrasischa@gmail.com

Bibhas Chandra Dhara

IT Department,
Jadavpur University,
Kolkata, India
Email: bcdhara@gmail.com

Sanjoy Kumar Saha

CSE Department,
Jadavpur University,
Kolkata, India
Email: sks_ju@yahoo.co.in

Abstract: In the context of a content-based music retrieval system or archiving digital audio data, genre-based classification of song may serve as a fundamental step. In the earlier attempts, researchers have described the song content by a combination of different types of features. Such features include various frequency and time domain descriptors depicting the signal aspects. Perceptual aspects also have been combined along with. A listener perceives a song mostly in terms of its tempo (rhythm), periodicity, pitch and their variation and based on those recognises the genre of the song. Motivated by this observation, in this work, instead of dealing with wide range of features we have focused only on the perceptual aspect like melody and rhythm. In order to do so audio content is described based on pitch, tempo, amplitude variation pattern and periodicity. Dimensionality of descriptor vector is reduced and finally, random sample and consensus (RANSAC) is used as the classifier. Experimental result indicates the effectiveness of the proposed scheme.

Keywords: song genre; pitch; tempo; amplitude variation; random sample and consensus; RANSAC.

Reference to this paper should be made as follows: Ghosal, A., Chakraborty, R., Dhara, B.C. and Saha, S.K. (2015) 'Perceptual feature-based song genre classification using RANSAC', *Int. J. Computational Intelligence Studies*, Vol. 4, No. 1, pp.31–49.

Biographical notes: Arijit Ghosal is working as an Assistant Professor in the CSE Department of Institute of Technology and Marine Engineering, West Bengal, India. He is currently pursuing his PhD work at Jadavpur University, Kolkata, India.

Rudrasis Chakraborty received his undergraduate and post graduate degree in Computer Science and Engineering from Jadavpur University, Kolkata, India and Indian Statistical Institute, Kolkata, India respectively. He is currently registered as a graduate student at University of Florida, USA.

Bibhas Chandra Dhara is working as an Associate Professor in the Department of Information Technology, Jadavpur University, Kolkata, India. He obtained his PhD from Indian Statistical Institute, Kolkata. Signal processing and pattern recognition are his areas of research.

Sanjoy Kumar Saha received his PhD from Bengal Engineering and Science University, Howrah, India and currently working as an Associate Professor in the Department of Computer Science and Engineering in Jadavpur University, Kolkata, India. His area of research includes multimedia data processing.

This paper is a revised and expanded version of a paper entitled 'Genre based classification of song using perceptual features' presented at the International Conference on Advanced Computing, Networking, and Informatics (ICACNI 2013), Department of Computer Science, Central Institute of Technology, Raipur, India, 12–14 June 2013.

1 Introduction

Song genre is a conventional category that identifies pieces of song as belonging to certain style or set of conventions. A genre stands for group of songs sharing stylistic characteristics with each other. Basically, it is a tag created and used by human to categorise the collection of songs and differentiate each other. It is very tedious and difficult to classify a large collection manually. Moreover, the concept of genre is subjective in nature and there may exist overlapping. It makes the manual task further difficult. With electronic music distribution (EMD), as the music catalogues becoming huge, it is almost impossible to manually check each of them. Hence, automatic genre-based classification has gained impetus and it may act as the fundamental step for content-based music retrieval system and organising the digital library of song.

In this work, we present a novel scheme that relies on perceptual features and uses random sample and consensus (RANSAC) as classifier. Song genres are mainly characterised by timbral feature of the signal, rhythmic content and pitch content. Rhythmic aspect, pitch and periodicity are the features that resemble our perception and play important role in deciding the genre. This observation has motivated us for this work. Pitch and tempo-based perceptual features (Ghosal et al., 2013) are considered to capture the melody and rhythmic characteristics of music signal. The strength of the descriptor has been further enhanced by incorporating amplitude variation-based contextual feature and correlation-based periodicity (*CbP*) measure. The present work demonstrates the effectiveness of the proposed features by considering three classifiers namely multi-layer perceptron (MLP) network, support vector machines (SVMs) and RANSAC. It is observed that in all the cases high classification accuracy has been achieved and RANSAC outperforms the other two.

The rest of this paper is organised as follows. A survey of past work is presented in Section 2. Section 3 elaborates the proposed methodology. Experimental results are presented in Section 4. Comparison with an existing system is presented in Section 5. Finally, the concluding remarks is put in Section 6.

2 Previous work

Song genre recognition has been an active area of research over the last decade and is still growing rapidly. A genre-based song classification system broadly consists of two modules – *feature extraction* and *classification*. Past study indicates that a wide variety of features and classification schemes have been adopted by the researchers.

Solatu et al. (1998) proposed a strategy based on a temporal structure model. A set of abstract features are extracted to represent temporal structures. The cepstral coefficients computed from the signal are fed as the input to MLP network for training corresponding to each category of song. Once the network converges, the hidden layer weights are taken as abstract features for the particular category/genre and later these are used for classification.

Scheirer (1998) explored music and beat tracking in their work. Tempo is analysed and beats present in the signal are extracted by using bandpass filters and parallel comb filters. Laroche (2001) also worked with features based on tempo, beat, etc. An effort has been made by Foote and Uchihashi (2001) to represent rhythm with the help of beat spectrum. A signal is divided into frames and each frame is represented by its power spectrum obtained through short-term Fourier transform (STFT). A dissimilarity matrix for the signal is generated that summarises the distance between each pair of frames. By adding the diagonal elements of the said matrix, a measure of beat spectrum is obtained.

SVM is widely used as classifier. To deal with multiple classes, bi-class SVM is used in hierarchical manner. Whitman et al. (2001) and Xu et al. (2003) have worked with short-time features and SVM classifiers. Each frame of a signal are classified and based on the majority voting genre of the signal is decided. Features like zero crossing rates, spectrum power, Mel-frequency cepstral coefficients (MFCC), beat spectrum are used. West and Cox (2004) also relied on a SVM-based classification technique. Moreno et al. (2004) have considered symmetric Kullback-Leibler (KL) divergence-based kernel in their SVM classifier. In contrast to widely used SVM, Gaussian process model has been proposed by Markov and Matsui (2013) for song classification.

Logan and Salomon (2001) have modelled the song with Gaussian mixture model (GMM) and training is done using k-means algorithm. Tzanetakis and Cook (2002), and Aucouturier and Pachet (2004) have used k nearest neighbour (kNN) as classifier that uses GMM trained song features. Similar approach has also been adopted by

Jothilakshmi and Kathiresan (2012). Features denoting the timbral texture, rhythmic and pitch content of the song are used by Tzanetakis and Cook (2002), and Silla et al. (2007). Several classifiers like Naive-Bayesian, MLP, and SVM are used in the work of Silla et al. (2007). Finally, all the classifier's output are combined based on an ensembling method to decide the song genre.

Lin et al. (2004) have suggested significant repeating patterns (SRP) as a descriptor for rhythm/melody. Rhythm denotes a sequence of beats present in the music signal and taken as the perceptual feature. In order to find out the repeating pattern in the music, the algorithm presented by Hsu et al. (2001) has been used. Multivariate auto-regressive feature model is tried in the work of Meng et al. (2007). Lo and Lin (2010), and Madjarov et al. (2012) have dealt with feature sets representing timbre, rhythm and pitch content. Grimaldi et al. (2003) introduced discrete wavelet packet transform features. Altogether, 143-dimensional features are computed based on beat histogram and spectrum of 16 level wavelet decomposed signals. For classification, wrapper and filter methods are used.

Zhen and Xu (2010) have followed multi-modal approach. Along with the low level acoustic feature and corresponding social tags (music-tag and artist-tag) gathered from the web are used. Using latent dirichlet allocation (LDA), tags are analysed to correlate it to the genre. The acquired knowledge helps to classify the untagged songs using low-level features and it relies on 'interaction-based forward feature selection' approach.

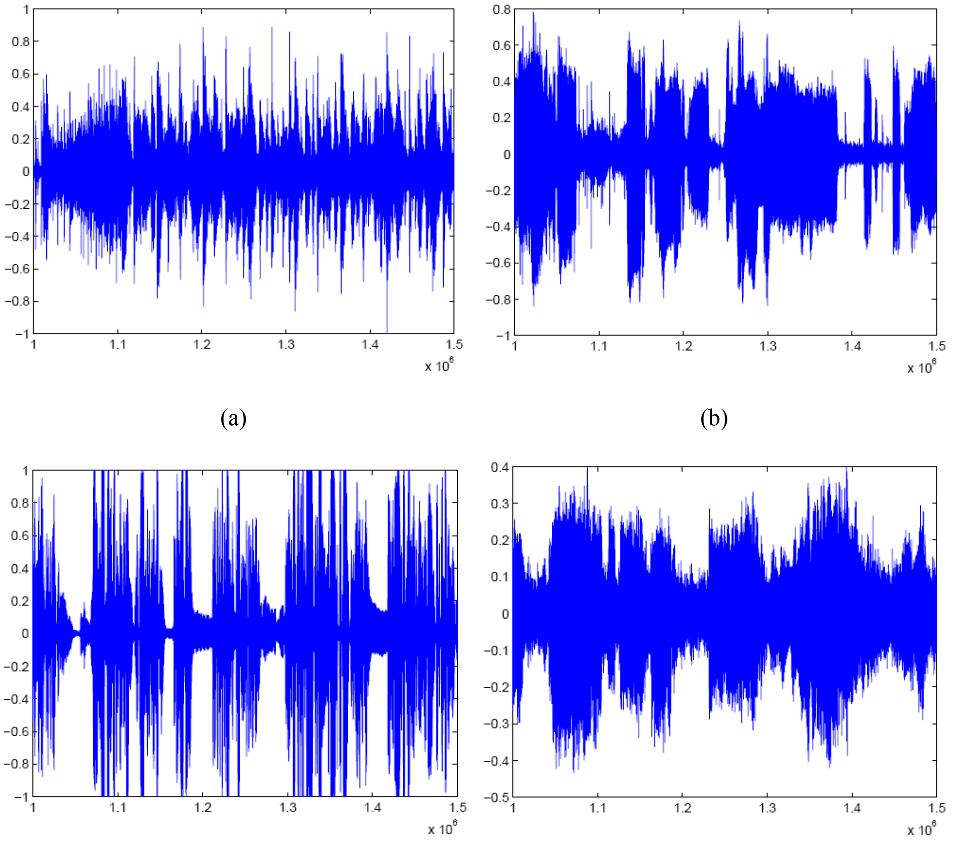
Octave-based spectral contrast (OSC) was tried in the work of Jiang et al. (2002). Spectral peak, spectral valley and their difference in various sub-bands are taken as features. Statistical spectrum of critical bands and rhythm histogram have been used by Lidy and Rauber (2005). MFCC in various forms have been utilised by Garcia et al. (2010). Simsekli (2010) has introduced the concept of melodic interval. Spectrogram-based description has been presented in the work of Costa et al. (2011). Non-negative tensor factorisation scheme has been deployed by Ren et al. (2010). Lee et al. (2011) have worked with long-term modulation spectral analysis of spectral an MFCC features to capture the rhythmic description. Finally, information fusion approach is followed for classification. Aryafar et al. (2013) have also considered MFCC and spectral analysis-based features and sparsity-eager SVM is used as classifier. Pitch contour-based melodic features are proposed by Salamon et al. (2012).

Past study reveals that wide variety of features are used as the descriptors for music signal. It is very difficult to obtain well defined mapping from feature space to song genre. A single feature may not discriminate multiple genres. This is the reason behind the use of features of various type to capture the signal pattern and to classify them mostly in a supervised manner. A listener identifies the genre of a song based on his perception and it has motivated us to go for perceptual features. Presence of various frequencies with their relative strength and variation pattern give rise to different perceptions. It has been observed that different perceptual features mostly based on the spectral analysis of the signal has drawn attention. In this work, we have focused on different types of the perceptual features to accomplish the task and finally, the classification is achieved using RANSAC.

3 Proposed methodology

Songs can be categorised into number of genres. Several categorisations might be possible, classical and non-classical is one of them. A crude genre categorisation can be pop, rock, jazz, Bhangra, classical, folk, Ghazal, etc. In this work, we took seven types of songs for automatic discrimination namely Bhangra (a North Indian genre), classical, folk, Ghazal (an Indian genre), jazz, pop and rock. A typical signal of each of the seven genres used in this work are given in Figure 1. In Section 3.1, extraction of several perceptual features used in this work are elaborated and the classification methodology is presented in Section 3.2.

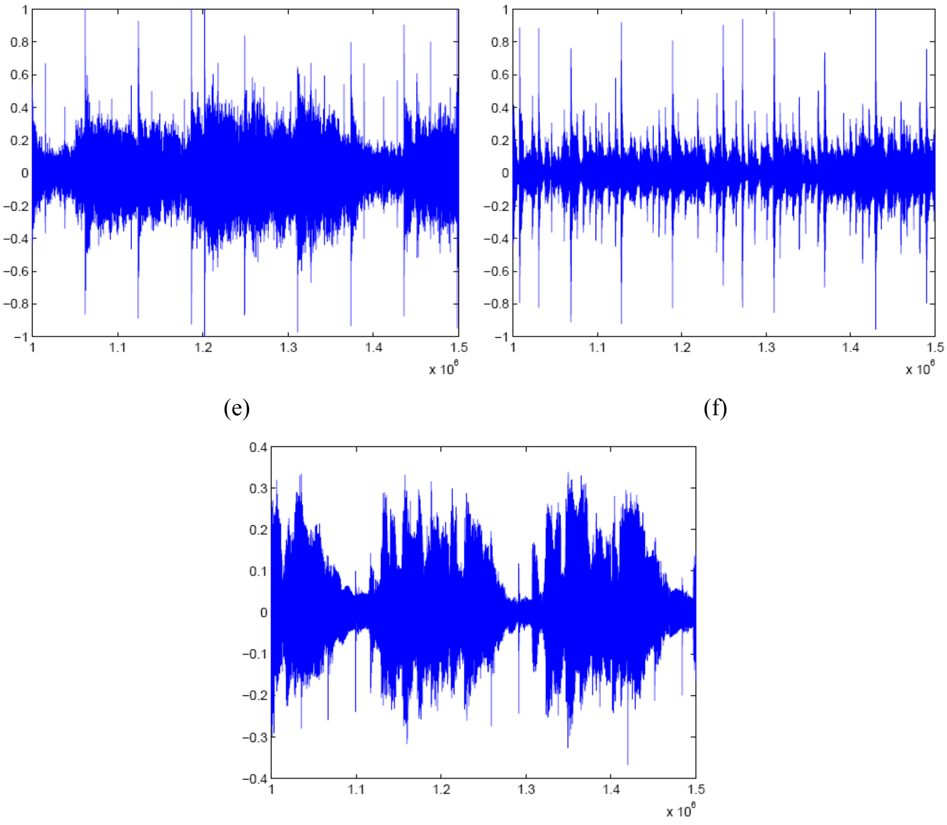
Figure 1 Plots of signal for (a) bhangra (b) classical (c) folk (d) ghazal (e) jazz (f) pop (g) rock genre song (time vs. amplitude) (see online version for colours)



(a) (b)

(c) (d)

Figure 1 Plots of signal for (a) bhangra (b) classical (c) folk (d) ghazal (e) jazz (f) pop (g) rock genre song (time vs. amplitude) (continued) (see online version for colours)



(e)

(f)

(g)

3.1 Extraction of features

In the works of Peeters (2004), West and Cox (2005), and Scaringella and Zoia (2005), temporal, energy, spectral shape and perceptual features have been used. Effectiveness of psychoacoustic features for genre classification has been mentioned in Lidy and Rauber (2005). Temporal features like zero-crossing rate, linear prediction coefficients, spectral shape features like fast Fourier transform (FFT) or MFCCs and perceptual features in the form of loudness, sharpness, spread have been used widely for genre classification. A combination of these different types of features might be possible as well and have been used in prior works. In this work, we have mainly focused on perceptual features. Melody and rhythm of a song are the two perceptual aspects that play major role in identifying the genre and proposed methodology utilises the same. For the genres like classical, Ghazal melody is the dominating property and so is rhythm for rock, jazz etc.

3.1.1 Pitch-based feature

Melody is formed by the linear succession of musical tones.

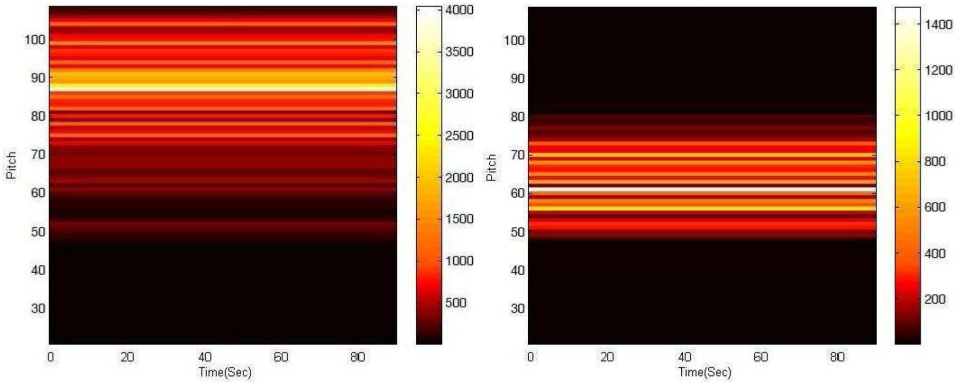
The presence of various tones along with the tonal strength reflects the melody. As a tone corresponds a frequency, energy distribution over the frequency can be taken as a measure for melody and it is computed as follows:

- decompose the signal into 88 frequency bands
- divide the signal of each sub band into short duration frames
- for each frame, compute short-time mean-square power (STMSP)
- for each band, compute average STMSP.

Pitch is the perceived form of frequency. Hence, the audio signal is decomposed into 88 frequency bands with centre frequency corresponding to pitches A0 to C8. These pitches correspond to consecutive MIDI pitches 21 to 108. For each sub-band, the average of the STMSP is computed. The average value denotes the energy localised in that frequency sub-band. In other words, it corresponds to the tonal strength represented by the sub-band. Considering all the sub-bands, 88-dimensional pitch feature vector (P) is formed. In order to extract the feature, chroma toolbox as suggested by Müller and Ewert (2011) is used.

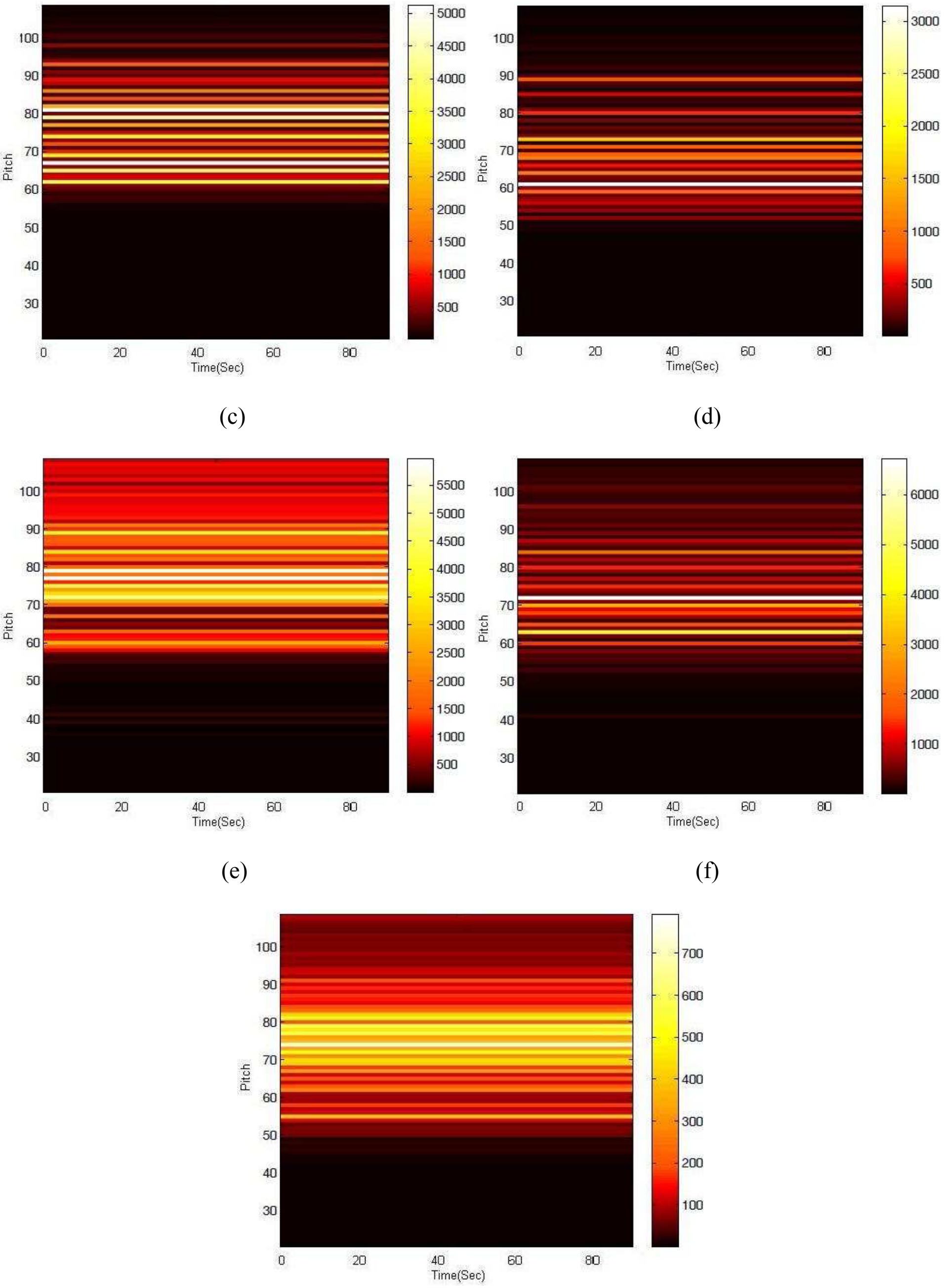
STMSP distribution over different frequency subbands is shown in Figure 2. STMSP for a band varies over time. For better understanding, the average value of STMSP for the bands are shown. For the melodic genres like classical and Ghazal, limited range of pitches are repeated and each of those has comparable strength. For others, higher bands are also present with considerable strength. It is evident that the variation of signal strength over the bands and presence of various tones (frequencies) differ from genre to genre. as a result, it can be utilised in discriminating the genres.

Figure 2 Samples of STMSP distribution over different sub-bands for various genres, (a) bhangra (b) classical (c) folk (d) ghazal (e) jazz (f) pop (g) rock (STMSP vs. pitch sub-band) (see online version for colours)



(a) (b)

Figure 2 Samples of STMSP distribution over different sub-bands for various genres, (a) bhangra (b) classical (c) folk (d) ghazal (e) jazz (f) pop (g) rock (STMSP vs. pitch sub-band) (continued) (see online version for colours)



(c)

(d)

(e)

(f)

(g)

3.1.2 Tempo-based feature

Melody is associated with pitch content of the signal. The rhythm or speed of a song is an important perceptual description and it is not reflected in melody. The duration pattern of the notes in a song reflects its rhythm. In a song, beats represent unbroken series of periodically occurring distinct short stimuli which are perceived as points in time. Beat interval is taken as the note duration. Tempo denotes how fast the beat is running and it represents the rhythmic property.

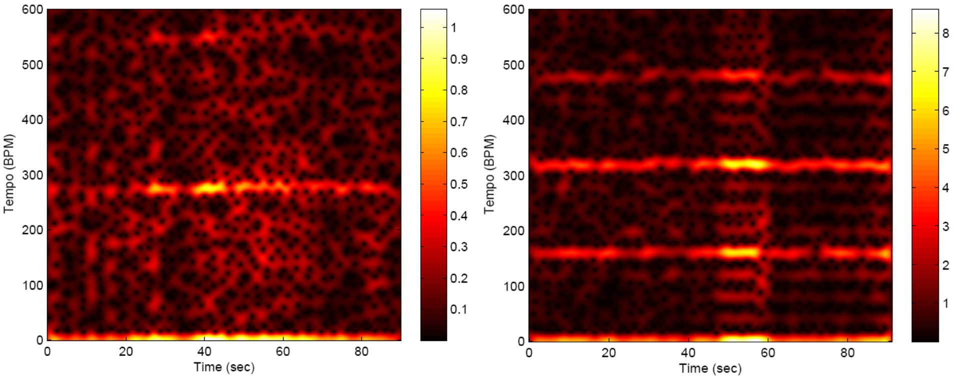
In order to extract the tempo-based features, at first novelty curve is computed from the signal to detect the onset following the methodology discussed in the work of Grosche and Müller (2011). Onset is the time position where a note is played and at that point of time change in signal properties like energy, pitch, spectral content occur. A spectrogram-based technique is used to obtain the novelty curve which represents the change in the spectral content. Peaks in the novelty curve are the candidate onsets. To obtain the tempogram, novelty curve is divided into non-overlapping tempo-window with 6 second duration each and STFT is applied on each window. Absolute values of the first 30 coefficients are extracted for each tempo-window. Mean and standard deviation of each such coefficient over all tempo-windows form the 60-dimensional feature vector $T < T_{0,1}, T_{1,1}, T_{0,2}, T_{1,2}, \cdots, T_{0,30}, T_{1,30} >$. $T_{0,i}$ and $T_{1,i}$ denotes the mean and standard deviation value of i^{th} Fourier coefficient respectively and i takes integer values ranging from 1 to 30. Thus, the major steps can be summarised as follows:

- compute the novelty curve from audio signal
- divide the novelty curve into non-overlapping tempo-window
- extract absolute value of first 30 Fourier coefficients from each tempo-window
- compute mean and standard deviation of each coefficient over all tempo windows.

The tempogram plot for classical and rock genre are shown in Figure 3. It is the time-tempo representation that encodes the local tempo of the song signal. Tempo is represented in terms of beat per minute (BPM). Intensity variation of different tempo (BPM) over time is shown. High rhythmic genres like rock, jazz maintain a strong continuity in tempo even at various BPM and it is not so for genres like classical. Thus, tempo-based features can well capture the rhythmic aspect and contribute in discriminating the genres.

Pitch and rhythmic content of a signal can help in discriminating the genres to a large extent. But there may exist an overlap among the genres in the said feature spaces. In terms of pitch content, classical and Ghazal have commonalities. Rock and Bhangra also reflect similarity among them. For classical genre, rhythmic pattern is different due to the absence of strong beats. But, genres like pop, rock, jazz, Bhangra possess high beats. Distinguishability among them arises out of the strength and periodicity of the beats. Still intermingling of different genres in the feature space may exist and it needs to be minimised by incorporating other features in addition to the spectral analysis-based description that we have already adopted. Variations in pitch and rhythm along with their strength also bear impact on the time-varying signal. It has motivated us to look for few other perceptual features like amplitude variation pattern and *CbP* which are derived from the envelope of the time-domain signal.

Figure 3 Tempogram plot for different genre, (a) classical (b) rock (time vs. tempo) (see online version for colours)



(a) (b)

3.1.3 *Amplitude variation pattern*

Two signals having the same loudness and pitch but having different waveforms will have different hearing perceptions. The shape or contour of a signal depends on its elementary components and can provide useful information for song genre identification. By tracing the amplitude variation pattern, the shape of the signal can be estimated. In case of an image, the distribution of intensity variation gives rise to a texture pattern. We extend the similar idea for song signal following the methodology of Ghosal et al. (2012). In order to compute the features the steps are as follows:

- smoothening of the signal
- obtain differential signal
- compute features based on the amplitude variation co-occurrence.

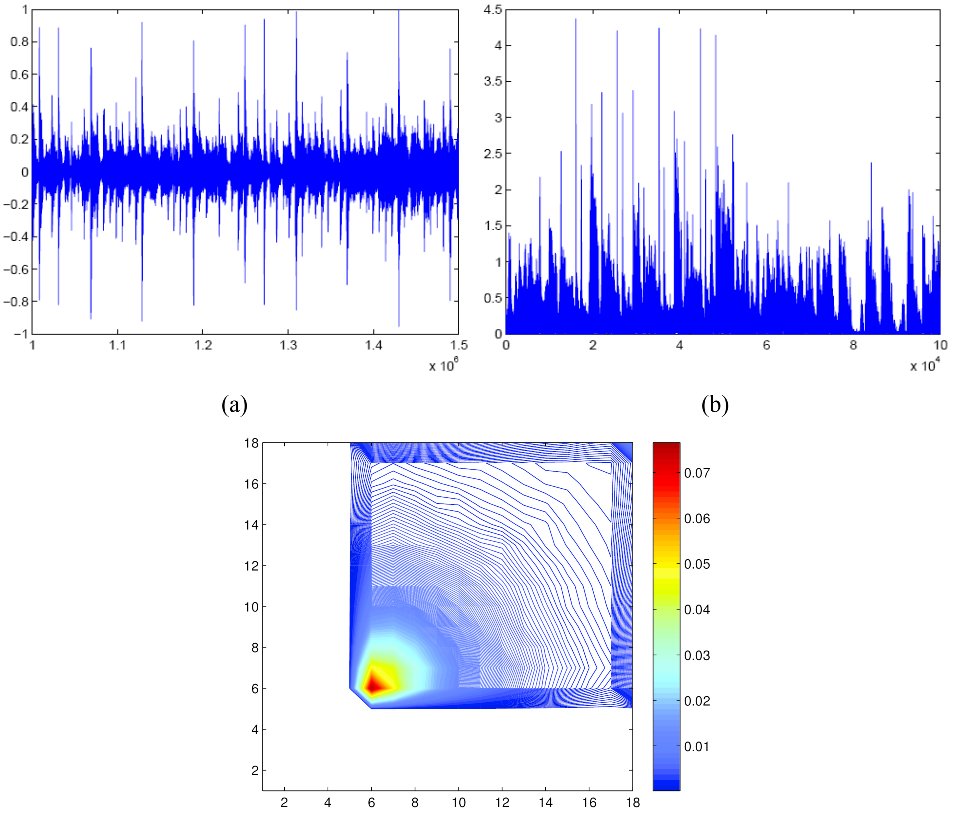
Smoothening of the signal is carried out by averaging the absolute signal amplitudes over a non-overlapping window of length 5 (i.e., the window consists of five samples). Let $\bar{\mathcal{S}}$ be the smoothened signal. The differential signal $\mathcal{D}$ is obtained by taking the absolute differences between two consecutive elements of $\bar{\mathcal{S}}$. Thus, i^{th} element of $\mathcal{D}$ is obtained as follows:

$$\mathcal{D}_i = abs\left(\bar{\mathcal{S}}_{i+1} - \bar{\mathcal{S}}_i\right).$$

The amplitude variation co-occurrence matrix $\mathcal{M}$ is computed from $\mathcal{D}$ as follows. The amplitude values of $\mathcal{D}$ are quantised considering $\mu \pm k \times \sigma$ where μ and σ are the mean and standard deviation of $\mathcal{D}$ is. In our work, k is varied from –2 to +2 with an increment of 0.25. Thus, altogether 18 quantisation levels are considered and 18 × 18 co-occurrence matrix $\mathcal{M}$ is formed where each matrix entry denotes the number of occurrence of a particular value pair in successive two positions. Uniformity in amplitude variation will be accumulated along the principal diagonal of matrix $\mathcal{M}$. Elements away from the diagonal denote the occurrence of patterns with diverging variation. Then, 18-dimensional feature vector, A is formed where an element, $A_i = \sum_r \mathcal{M}(r, r-i)$. In

Figure 4, an original signal, corresponding differential signal and the plot for co-occurrence matrix are shown.

Figure 4 Steps for generating amplitude variation pattern (a) original signal (b) differential signal and (c) co-occurrence matrix (see online version for colours)



(a)

(b)

(c)

3.1.4 Correlation-based periodicity

This is another perceptual features which captures the periodicity in the given signal. Signal *S* is first divided into non-overlapping frames. In our experiment, each frame consists of 100 samples. Then absolute Pearson's correlation coefficients (Walk and Rupp, 2010) between the frames are computed. For each frame maximum correlation value is noted. The average and standard deviation of such frame level maximum correlation values are taken as *CbP* feature set. The Pearson's correlation coefficient can take value between –1 to +1. If there is high absolute correlation value between two frames, then it is expected that those two frames are almost similar or in other words, there is a high periodicity. We have taken the absolute value of the coefficient as both positive and negative high correlation represents high periodicity. Mathematically, the *CbP* features can be represented as follows:

$$CbP_{mean} = \frac{\sum_{i=1}^{noF} \max_{j \neq i} \left\{ \left| \rho\left(Fr_i, Fr_j\right) \right| \right\}}{noF}$$

$$CbP_{std} = \sqrt{\frac{\sum_{i=1}^{noF} \left(\max_{j \neq i} \left\{ \left| \rho\left(Fr_i, Fr_j\right) \right| \right\} - CbP_{mean} \right)^2}{noF}}$$

noF denotes the number of non-overlapping frames in the signal. Fr_i denotes the vector consisting of the amplitude values of i^{th} frame. ρ denotes the Pearson's correlation coefficient value. The steps for computing *CbP* can be summarised as follows:

- divide the signal into non-overlapping frames
- for each frame, compute cross-correlation with others
- form a vector of size *noF* by taking maximum cross-correlation of each individual frame
- compute mean and standard deviation of the elements in the vector.

3.1.5 Merging and dimensionality reduction

The features computed are of high dimension. Pitch-based features (*P*) is of 88 dimension, tempo-based features (*T*) is of 60 dimension. Amplitude variation-based features (*A*) and *CbP* are of 18 and 2 dimension, respectively. Due to the high dimension of the feature vectors, the classification system's complexity would increase. In order to reduce the classifier's complexity, dimensionality feature vector is reduced following principal component analysis (PCA) (David, 2000).

PCA is an orthogonal transformation to un-correlate the features. It can be viewed as a feature extraction scheme where the un-correlated features are a linear combination of the original feature sets. Like any feature extraction scheme, here also the original feature set is changed. The first principal component denotes the direction with the highest variance. The succeeding components also have the highest variances but they are orthogonal to the previous components.

In order to study the utility of the features, we have combined *P* and *T* to form a 148 dimensional feature vector. It is referred as F_1. Subsequently, *A* and *CbP* is combined with F_1 to form 168 dimensional vector F_2. In both the cases, PCA is applied to bring down the dimension. In our experiment, the first 35 eigen vectors are used for classification. Corresponding to F_1 and F_2 reduced vectors obtained after PCA are referred as $\mathcal{F}_1, \mathcal{F}_2,$ respectively.

3.2 Classification

Song clips are represented by the feature vector and then the subsequent task is to identify the genre. In order to do so, we have carried out the experiment with three classifiers namely, MLP network, SVMs and RANSAC (Fischler and Bolles, 1981). Among these MLP and SVM are widely used in audio signal classification.

In MLP network, number of nodes in the input layer is same as the feature dimension and the output layer has number of nodes same as the number of classes (genres). In between one or more hidden layers are there. Determining the number of hidden layers (l) and number of nodes in the hidden layer (m) are two very important issues. In our experiment, we have considered only one hidden layer (i.e., $l = 1$). Number of hidden nodes is determined experimentally by varying it from 4 to 10 with ten-fold cross validation. By studying the performance, m corresponding to the maximum test accuracy is considered. In case of SVM, Radial Basis function kernel is used. Two parameters namely cost (c) and Gamma (g) need to be tuned for this RBF kernel. Extensive search is carried out between [–20, 20] with 0.5 as increment to obtain the best c and g values using ten-fold cross validation as earlier.

Although MLP network or SVM-based classification schemes are quite popular, they suffer from certain shortcomings also. In the context of genre-based classification of song, variation present even with in a genre degrades the performance of neural network-based classification. SVM is robust but the tuning of parameters for optimal performance is very critical and time consuming. It has motivated us to look for a robust estimator capable of modelling the data satisfactorily even in presence of diversity or outliers. RANSAC appears as a suitable alternative to fulfil the requirement.

RANSAC estimates the parameters of a certain model iteratively from a set of data despite the presence of large number of outliers. There are two major steps namely, *hypothesise* and *test*. During hypothesise phase, estimation is made based on a randomly selected minimal sample set. During the test phase, rest of the elements in the dataset are verified with the model. Elements consistent with the model estimated so far are included in the sample set for the next iteration. An element is consistent, i.e., taken as inlier if its distance from the evolving model lies with in a threshold. Thus, in each iteration a new model evolves and the best fitted one is considered as the final estimate. The major strength of RANSAC over other estimators lies in the fact that the estimation is made based on inliers, i.e., whose distribution can be explained by a set of model parameters. In our experiment, the threshold is taken as 0.02 which is the suggested default value. Presence of sizable inliers in the dataset is sufficient to obtain a reasonably good model and it can work satisfactorily even with outliers amounting to 50%of entire dataset (Fischler and Bolles, 1981). In order to classify the songs, first of all, model for each genre is estimated by using a subset of songs of each genre and for subsequent classification the evolved model is used.

4 Experimental results

In order to carry out the experiment, we have formed a song database with 490 files of seven genres like Bhangra (a North Indian genre and is rhythmic), classical, folk, Ghazal (an Indian genre of semi-classical type), jazz, pop and rock. The classical songs are mostly Indian. Thus, the collection reflects enough variety due to presence of Indian and Western songs. In each of the seven genres, there are 70 data files collected from different sources over the internet. Each file is of 60–90 seconds duration sampled at 22,050 Hz. Samples are quantised in 16−bit and of type mono. The dataset is chosen carefully so that each dominantly belongs to only one genre and is manually assigned a genre by the domain expert for evaluating the system performance.

All the two feature sets consist of 490 points and seven classes (i.e., seven song genres). These feature sets are randomised and z-score normalised. For all the two feature sets, the number of hidden nodes for MLP network is experimentally selected as nine by cross-validation scheme. In case of SVM classifier, cost (*c*) and gamma (*g*) values are tuned by grid search technique of LibSVM package (Chang and Lin, 2011). For feature set $\mathcal{F}_1$, the values of *c* and *g* are determined as 20.03 and 0.25, respectively. In case of feature set $\mathcal{F}_2$, the values are 29.26 and 0.46.

The overall classification accuracy for different feature sets and classifiers are shown in Table 1. It may be noted that using feature set $\mathcal{F}_2$, substantial improvement (with respect to $\mathcal{F}_1$) has been achieved for all the classifiers. It indicates that the inclusion of features based on amplitude variation (*A*) and *CbP* enhances the strength. Furthermore, RANSAC as classifier has outperformed the other two because of its strength in handling the outliers.

Table 1 Overall classification accuracy (in %)

Classification scheme	*Feature set* $\mathcal{F}_1$	*Feature set* $\mathcal{F}_2$
MLP	84.69	92.86
SVM	86.94	95.10
RANSAC	89.39	97.55

Table 2 Confusion matrix for classification using $\mathcal{F}_1$ and *RANSAC*

Genre	*Bhangra*	*Ghazal*	*Classical*	*Folk*	*Jazz*	*Pop*	*Rock*
Bhangra	*66*	0	0	1	0	2	1
Ghazal	0	*70*	0	0	0	0	0
Classical	0	0	*70*	0	0	0	0
Folk	2	3	1	*62*	2	0	0
Jazz	0	1	1	2	*53*	1	12
Pop	4	3	0	3	0	*60*	0
Rock	0	2	0	1	6	4	*57*

Table 3 Confusion matrix for classification using $\mathcal{F}_2$ and *RANSAC*

Genre	*Bhangra*	*Ghazal*	*Classical*	*Folk*	*Jazz*	*Pop*	*Rock*
Bhangra	*69*	1	0	0	0	0	0
Ghazal	0	*70*	0	0	0	0	0
Classical	0	0	*70*	0	0	0	0
Folk	0	0	0	*70*	0	0	0
Jazz	0	1	0	1	*65*	0	3
Pop	1	1	0	1	0	*66*	1
Rock	0	0	0	0	1	1	*68*

The confusion matrices for classification using different feature sets ($\mathcal{F}_1$ and $\mathcal{F}_2$) and RANSAC are shown in Tables 2 and 3. It is observed that classical and Ghazal genres are correctly classified. On the other hand, jazz has the highest misclassification and it is mostly confused with the genre rock. Because of the overlap of various genres in the feature space, the intermingling occurs. Moreover, it happens even in case of manual judgment also. As the amplitude variation and periodicity-based features are introduced, misclassification is substantially reduced and it indicates their effectiveness.

5 Comparison

The performance of the proposed methodology (using feature set $\mathcal{F}_2$ and RANSAC) is compared with that of the works of Wang et al. (2008), Lee et al. (2009) and Lee et al. (2011). All the systems are implemented and applied on the same data set used in this work. The comparative results in terms of classification accuracy is presented in Table 4.

In the work of Wang et al. (2008), MFCC and four other MPEG-7 features (spectral centroid, spectral flux, zero crossing rate, spectral roll off) are computed for the frames of short duration. A group of frames are gathered to form a segment. Average and variance of the frame level features are the descriptors for the segment. Random forest (RF) and MLP are used independently to classify the segments. Finally, based on weighted voting fusion strategy, the result of the two classifiers on each segment is combined to put a label. An audio clip consists of multiple segments. Most frequently labelled genre over all segments is taken as the genre for the audio clip.

In the work of Lee et al. (2009), features based on MFCC, OSC and normalised audio spectral envelope (NASE) are extracted based on short-term power spectrum. To generate MFCC-based features, the spectrum is divided into Melscale sub bands. On the other hand, octave-scale filters are used to divide the power spectrum for computing OSC-based features. To extract NASE-based feature, the spectrum is divided into logarithmic spaced sub-bands. First 20 coefficients of MFCC, Spectral peak and spectral valley of top 20 dominating octave-scale bands and 19 dimensional NASE coefficients are computed for each frame. Finally, mean and variance of each feature over the frames form 118-dimensional descriptor for the audio clip. For classification nearest neighbour (NN) classifier is used.

Table 4 Comparison of performance in terms of classification accuracy (in %)

Method	*Bhangra*	*Ghazal*	*Classical*	*Folk*	*Jazz*	*Pop*	*Rock*	*Overall*
Method by Wang et al. (2008)	76.81	80.00	98.41	96.15	72.93	74.17	78.23	82.39
Method by Lee et al. (2009)	94.29	100.00	97.55	97.14	75.71	82.86	88.57	90.20
Method by Lee et al. (2011)	71.43	68.57	88.57	80.00	48.57	80.00	88.57	75.10
Proposed method ($\mathcal{F}_2$ + RANSAC)	98.57	100.00	100.00	100.00	92.85	94.29	97.14	97.55

Table 5 Confusion matrix for classification using the method proposed by Lee et al. (2009)

Genre	*Bhangra*	*Ghazal*	*Classical*	*Folk*	*Jazz*	*Pop*	*Rock*
Bhangra	65	0	0	0	0	5	0
Ghazal	0	70	0	0	0	0	0
Classical	0	1	68	0	0	1	0
Folk	0	0	0	66	1	2	1
Jazz	3	0	0	2	53	3	9
Pop	7	0	0	1	1	58	3
Rock	0	0	0	0	8	0	62

MFCC, OSC and NASE-based frame level features are again considered in the work of Lee et al. (2011). Instead of working with mean and variance of frame level features, long-term modulation analysis is applied on MFCC, OSC and NASE-based features to gather the time varying behaviour of the signals. Frame level features are concatenated to form a long duration window (texture window). By applying FFT on the window, modulation spectrogram is obtained. By time averaging the magnitude of all such spectrogram, the signature for the signal is generated. For classification non-parametric discriminant analysis (NDA) is applied.

Table 4 clearly shows that the proposed scheme performs better than the others. All the other three systems under consideration mostly rely on MFCC and statistical features based on different variants of spectrum analysis. On the contrary, proposed methodology considers the features that correlate with the human perception. The categorisation of the song into various genres is highly perceptual. Proposed methodology focuses on two major perceptual aspect like melody and rhythm. Genres like pop, jazz, rock, Bhangra has similar behaviour in terms of rhythm. On the other hand, classical, ghazal overlaps in terms of melody. Thus, training becomes crucial for classification. In this context, RANSAC contributes significantly because of its strength in handling the outliers. Among the three systems with which comparison is made, the performance of the system proposed by Lee et al. (2009) is closest to the proposed methodology. Confusion matrices of the proposed system and system of Lee et al. (2009) are shown in Tables 3 and 5, respectively. It also clearly indicates the superiority of the proposed methodology in discriminating the confusing genres.

6 Conclusions

In this work, a novel scheme for song genre classification is presented. Instead of considering a wide variety of features proposed scheme judiciously uses the features in a focused manner and considers only the perceptual features to capture the melodic and rhythmic aspects. Audio content is described by extracting the features based on pitch and tempo. Amplitude variation pattern of the signal and *CbP* are directly computed from the time-domain signal which helps in discriminating the genres which are prone to get misjudged. PCA is used to reduce the dimensionality of feature vector and RANSAC acts as the classifier. Experimental result establishes the strength of the features and also the utility of RANSAC as classifier in comparison to other classification technique.

Comparison with few other state-of-the-art techniques indicates that the performance of the proposed methodology is substantially better.

References

Aryafar, K., Adams, T.R. and Shokoufandeh, A. (2013) 'Content-based music genre classification using sparse approximation techniques', *International Conf. on Pattern Recognition*.

Aucouturier, J-J. and Pachet, F. (2004) 'Improving timbre similarity: how high is the sky?', *Journal of Negative Results in Speech and Audio Sciences*, Vol. 1, No. 1, pp.1–13.

Chang, C-C. and Lin, C-J. (2011) 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp.27:1–27:27.

Costa, Y.M.G., Oliveira, L.S., Koreich, A.L. and Gouyon, F. (2011) 'Music genre recognition using spectrograms', *Intl. Conf. on Systems, Signal and Image Processing*, pp.1–4.

David, L. (2000) *Linear Algebra and It's Applications*, pp.441–486, Addison-Wesley, New York.

Fischler, M.A. and Bolles, R.C. (1981) 'Random sample consensus: a paradigm for model for model fitting with applications to image analysis and automated cartography', *ACM Communications*, Vol. 24, No. 6, pp.381–395.

Foote, J.T. and Uchihashi, S. (2001) 'The beat spectrum: a new approach to rhythmic analysis', *IEEE International Conference on Multimedia and Expo (ICME)*.

Garcia, D.G., Garcia, J.A., Hernandez, E.P. and Maria, F.D. (2010) 'Music genre classification using temporal structure of songs', *IEEE Intl. Workshop on Machine Learning for Signal Processing*, pp.266–271.

Ghosal, A., Chakraborty, R., Dhara, B.C. and Saha, S.K. (2012) 'Music classification based on mfcc variants and amplitude variation pattern: a hierarchical approach', *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 5, No. 1, pp.131–150.

Ghosal, A., Chakraborty, R., Dhara, B.C. and Saha, S.K. (2013) 'Genre based classification of song using perceptual features', *International Conference on Advanced Computing, Networking, and Informatics (ICACNI-2013)*.

Grimaldi, M., Cunningham, P. and Kokaram, A. (2003) 'An evaluation of alternative feature selection strategies and ensemble techniques for classifying music', *Workshop on Multimedia Discovery and Mining*.

Grosche, P. and Müller, M. (2011) 'Tempogram toolbox: MATLAB tempo and pulse analysis of music recordings', *12th International Conference on Music Information Retrieval (ISMIR, Late-Breaking Contribution)*, Miami, USA.

Hsu, J.L., Liu, C.C. and Chen, A.L. (2001) 'Discovering nontrivial repeating patterns in music data', *IEEE Transactions on Multimedia*, Vol. 3, No. 3, pp.311–325.

Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H. and Cai, L.H. (2002) 'Music type classification by spectral contrast feature', *IEEE International Conference on Multimedia and Expo (ICME)*, pp.113–116.

Jothilakshmi, S. and Kathiresan, N. (2012) 'Automatic music genre classification for Indian music', *International Conf. on Software and Computer Applications*, pp.55–59.

Laroche, J. (2001) 'Estimating tempo, swing and beat locations in audio recordings', *Workshop on Appln. of Sig. Proc. to Audio and Acoustics (WASPAA)*, pp.135–138.

Lee, C.H., Shih, J.L., Yu, K.M. and Lin, H.S. (2009) 'Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features', *IEEE Transactions on Multimedia*, Vol. 11, No. 4, pp.670–682.

Lee, C-H., Chou, C-H. and Fang, J-C. (2011) 'Automatic music genre classification using modulation spectral features and nonparametric discriminant analysis', *Journal of Information Technology and Applications*, Vol. 5, No. 2, pp.75–82.

Lidy, T. and Rauber, A. (2005) 'Evaluation of feature extractors and psycho-acoustic transformations for music genre classification', *Proceedings of the 6th Int. Symposium on Music Information Retrieval*.

Lin, C.R., Liu, N.H., Wu, Y.H. and Chen, A.L.P. (2004) 'Music classification using significant repeating patterns', *Database Systems for Advanced Applications*, Vol. 2973, pp.506–518.

Lo, Y.L. and Lin, Y.C. (2010) 'Content-based music classification', *Intl. Conf. on Computer Science and Information Technology*, Vol. 2, pp.112–116.

Logan, B. and Salomon, A. (2001) 'A music similarity function based on signal analysis', *ICME 2001*.

Madjarov, G., Pesanski, G., Spasovski, D. and Gjorgjevikj, D. (2012) 'Automatic music classification into genres', *ICT Innovations 2012, Web Proceedings*, pp.623–631.

Markov, K. and Matsui, T. (2013) 'Music genre classification using Gaussian process models', *IEEE International Workshop on Machine Learning for Signal Processing*, pp.1–6.

Meng, A., Ahrendt, P., Larsen, J. and Hansen, L.K. (2007) 'Temporal feature integration for music genre classification', *IEEE Transactions on Speech and Audio Processing*, Vol. 15, No. 5, pp.1654–1664.

Moreno, P.J., Ho, P.P. and Vasconcelos, N. (2004) 'A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications', in Thrun, S., Saul, L. and Scholkopf, B. (Eds.): *Advances in Neural Information Processing Systems*, p.16.

Müller, M. and Ewert, S. (2011) 'Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features', *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

Peeters, G. (2004) 'A large set of audio features for sound description (similarity and classification)', CUIDADO project, CUIDADO I.S.T. Project Report.

Ren, J.M., Chen, Z.S. and Jang, J.S.R. (2010) 'On the use of sequential patterns mining as temporal features for music genre classification', *IEEE Conf. on Acoustics, Speech and Signal Processing*, pp.2294–2297.

Salamon, J., Rocha, B. and Gomez, E. (2012) 'Musical genre classification using melody features extracted from polyphonic music signals', *IEEE Conf. on Acoustics, Speech and Signal Processing*, pp.81–84.

Scaringella, N. and Zoia, G. (2005) 'On the modeling of time information for automatic genre recognition systems in audio signals', *Proceedings of the 6th International Symposium on Music Information Retrieval*, pp.666–671.

Scheirer, E. (1998) 'Tempo and beat analysis of acoustic musical signals', *Journal of Acoustical Society of America*, Vol. 103, No. 1, pp.588–601.

Silla, C.N., Kaestner, C.A.A. and Koerich, A.L. (2007), 'Automatic music genre classification using ensemble of classifiers', *IEEE International Conference on Systems, Man and Cybernetics, 2007, ISIC*, IEEE, pp.1687–1692.

Simsekli, U. (2010) 'Automatic music genre classification using bass lines', *Intl. Conf. on Pattern Recognition*, pp.4137–4140.

Solatu, H., Schultz, T., Westphal, M. and Waibel, A. (1998) 'Recognition of music types', *IEEE Conf. on Acoustics, Speech and Signal Processing*, pp.1137–1140.

Tzanetakis, G. and Cook, P. (2002) 'Musical genre classification of audio signals', *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp.293–302.

Walk, M.J. and Rupp, A.A. (2010) 'Pearson product-moment correlation co-efficient', *Encyclopedia of Research Design*.

Wang, L., Huang, S., Wang, S., Liang, J. and Xu, B. (2008) 'Music genre classification based on multiple classifier fusion', *International Conf. on Natural Computation*, Vol. 5, pp.580–583.

West, K. and Cox, S. (2004) 'Features and classifiers for the automatic classification of musical audio signals', *International Symposium on Music Information Retrieval*.

West, K. and Cox, S. (2005) 'Finding an optimal segmentation for audio genre classification', *Proceedings of the 6th Int. Symposium on Music Information Retrieval*.

Whitman, B., Flake, G. and Lawrence, S. (2001) 'Artist detection in music with minnow match', *IEEE Workshop on Neural Networks for Signal Processing*, pp.559–568.

Xu, C., Maddage, N.C., Shao, X., Cao, F. and Tian, Q. (2003) 'Musical genre classification using support vector machines', *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Vol. 5, pp.v-429–v-432.

Zhen, C. and Xu, J. (2010) 'Multi-modal music genre classification approach', *Intl. Conf. on Computer Science and Information Technology*, Vol. 8, pp.398–402.