Music Classification based on MFCC Variants and Amplitude Variation Pattern: A Hierarchical Approach

Arijit Ghosal[1], Rudrasis Chakraborty[2], Bibhas Chandra Dhara[3] and Sanjoy Kumar Saha[3]

[1]*Institute of Technology and Marine Engg., 24 Parganas (South), West Bengal, India*
[2]*Indian Statistical Institute, Kolkata, India*
[3] *Jadavpur University, Kolkata, India*
ghosal.arijit@yahoo.com, rudrasischa@ieee.org, bcdhara@gmail.com, sks_ju@yahoo.co.in

Abstract

In this work, we have presented a hierarchical scheme for classifying music data. Instead of dealing with large variety of features, proposed scheme relies on MFCC and its variants which are introduced at the different stages to satisfy the need. At the top level music is classified as song (music with voice) and instrumental (music without voice) based on MFCC. Subsequently, instrumental signals and songs are classified based on instrument type and genres respectively. Hierarchical approach has been followed for such detailed categorization. Using two-stage process, instrumental signals are identified as one of the four types namely, string, woodwind, percussion or keyboard. Wavelet and MFCC based features are used for this purpose. For song classification, at first level signals are categorized as classical or non-classical(popular) ones by capturing the MFCC pattern present in the high sub-band of wavelet decomposed signal. At second level, we consider the task of further classification of popular songs into various genres like Pop, Jazz, Bhangra (an Indian genre) based on amplitude variation pattern. RANSAC has been utilized as the classifier at all stages. Experimental result indicates the effectiveness of the proposed schemes.

Keywords: *Song Classification, Music Retrieval, Audio Classification, MFCC, RANSAC.*

1. Introduction

The development in multimedia technology has led to the enormous growth in the data volume of various media including audio. But, fast accessing and retrieving the desired piece of data is an important issue. In the field of image and video lots of work have been carried out in comparison to audio domain [46]. An efficient audio classification system can serve as the foundation for various applications like audio indexing, content based audio retrieval [42], audio content description, and music genre classification [43].

Over a prolonged period, audio signal processing activity centered on speech processing, speaker recognition, speech-music discrimination etc. But applications like music retrieval have come up strongly and music classification has gained high impetus over the last decade. An audio classification system consists of two important modules like feature extraction and classification. Characteristics like musical structure, rhythm, tempo, melody can be used to discriminate the music types. But, it is not easy to extract

the features representing the characteristics. Past work indicates significant variety in their approaches.

In the context of music retrieval system, at first level it is necessary to classify them as music without voice i.e. instrumental and music with voice i.e. song. Such classification bears significant importance in the context of retrieval system dealing with multilingual songs where for heterogeneous audience an instrumental is preferred over music with voice. Moreover, locating singing voice in an audio track is one of the applications where song/instrumental discrimination becomes essential. A few works [2, 47] have been reported in this direction. Berenzweig et al. [2] have relied on the fact that song will have features of speech embedded in it and speech-trained model is used to detect the song. The success of the scheme relies on the availability of suitable speech recognizer. Zhang et al. [47] have considered four features based on average ZCR and fundamental frequency. Threshold for each features are considered to characterize the music signals and finally the decision is taken based on heuristic approach. The success of the schemes heavily depends on the proper selection of thresholds. Music retrieval being an important application, it is very crucial to differentiate the song and instrumental music. Only after such recognition, tasks of identifying the instrument type, genre based song classification become meaningful.

Automatic recognition of instrument or its type like string, woodwind, keyboard is an important issue. In several works [9, 20], isolated musical notes have been considered as input to the system. But, in the signal arising out of a performance, the notes are not separated [29]. Thus, note based schemes have limited scope. On the other hand recognition of musical instruments in a polyphonic, multi-instrumental music is a difficult challenge and a successful recognition system for a single instrument music may help in addressing the case for multi-instrumental music [29]. Study indicates a few efforts [8] are directed in dealing with multi-instrumental signal and most of the works are still restricted to solo music. A comprehensive study made by Deng [6] indicates that a wide variety of features and classification schemes have been reported by the researchers. Mel Frequency Cepstral Coefficient (MFCC) has been used in different manner in number of systems. Marques et al. [33] have presented a MFCC based system using Gaussian Mixture Model and SVM as the classifier. MPEG-7 audio features have been tried by many researchers [35, 23]. Brown et al. [3] have relied on MFCC, spectral centroid, auto correlation coefficients and adopted Bayes decision rules for classification. Agostini et al. [1] have dealt with timbre classification based on spectral features. A set of 62-dimensional temporal, spectral, harmonic and perceptual features is used by Livshin et al [29]. The dimension is subsequently reduced to 20 and k-NN classification is used for recognition. Kaminskyj et al. [22] have initially considered 710 features including MFCC, rms, spectral centroid, amplitude envelope and dimensionality is reduced by performing PCA. Finally, k-NN classifier is used. The branch and bound search technique and non negative matrix factorization have been considered by Benetos et al. [23] respectively for feature selection and classification.

Comparatively, a lot of work has been done towards classification of song. Songs may be classified based on different characteristics. But, genre based classification resembles the human perception most. Genre is a subjective concept and depends on individual perception. Hence, it is quite difficult to define and in the context of automated classification it can broadly be thought of as the rhythmic patterns and pitch distributions inherent in the signal [7].

Solatu et al. [40] have used temporal structures and HMM. Tempo and beat tracking have been explored in [25]. As a global representation for rhythm, beat spectrum has

been presented in [13]. Lin et al. [28] have suggested Significant Repeating Patterns (SRP) as a descriptor for rhythm/melody. Researchers [43, 30] have worked with feature sets representing timbre, rhythm and pitch content. Grimaldi et al. [18] have dealt with 143 dimensional features computed based on beat histogram and spectrum of 16 level wavelet decomposed signals. Octave-based spectral contrast has been tried in [21]. It considers spectral peak, spectral valley and their difference in different sub-bands. Statistical spectrum descriptors based on 24 critical bands and rhythm histogram features have been used by Lidy and Rauber [27]. Lee et al. [26] have considered commonly used static features computed over the frames and further combined the transitional aspect by considering the variation of those static features. MFCC in various forms have been utilized by many researchers [48, 14]. Simsekli [39] have introduced the concept of melodic interval histogram as the descriptor. Spectrogram based description has been presented in [5]. Non negative tensor factorization scheme has been deployed in [37].

For classification also, various techniques have been tried by the researchers. Threshold based techniques [38], Neural Network [34], clustering [41] have been deployed for song classification. A heuristic rule based scheme has been proposed in [45]. K-Nearest Neighbors and GMM have been used in [43]. Bayes decision rules [9], SVM based scheme [5] have also been tried. Perceptually weighted Euclidean distance has been presented in [39] and finally relied on KNN for classification. Linear Discriminant Analysis (LDA) has been tried in [49, 26]. Langlois et al. [24] have followed language modeling approach to model the genre and a two stage clustering technique is used to classify the songs. A fuzzy rule based scheme [10] has also come up. Ren et al. [37] have presented a scheme combining string tokenization method and data mining technique. Thus, it is quite apparent that wide ranges of methodology have been tried by the researchers.

In this work, at first level we have discriminated a music signal as music with voice (song) or without voice (instrumental). After that, hierarchical scheme has been followed for each type i.e. song and instrumental to classify them. Instrumentals are classified according to the type of instruments whereas genre based classification has been achieved for song. Previous work reflects that the common practice is to consider a wide variety of features for describing the signals. The role played by different types of descriptor in a high dimensional feature vector is very difficult to understand. It has motivated us to go for a hierarchical approach and judicious dealing of the descriptors at each level focusing on the need. The organization of the paper is as follows. The introduction is followed by the description of proposed methodology in section 2. Experimental results are presented in section 3. Concluding remarks are put in to section 4.

2. Proposed Methodology

In this work, we have first accomplished the task of discriminating music with voice (song) and music without voice (instrumental). Subsequently, instrument type based classification of instrument signal and genre based song classification have been considered. The schematic diagram of the proposed methodology has been shown in Figure 1. Feature computation and classification schemes are two major modules of the system which are elaborated in section 2.1 and 2.2.

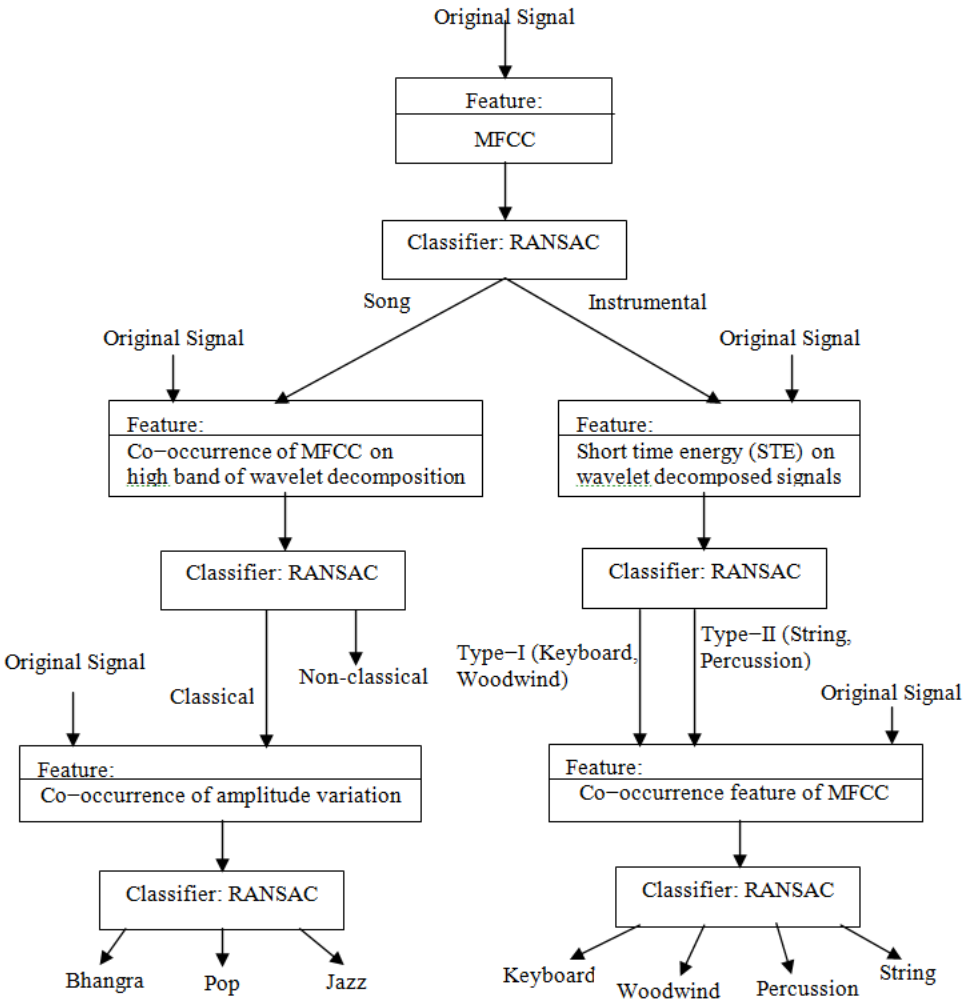


Figure 1. Schematic Diagram of Proposed Methodology

2.1. Computation of Features

In this work, we have followed a hierarchical scheme for classifying the music signals in various categories and sub-categories. As it has been indicated in section 1, we have relied on features of particular type to capture the suitable characteristics of the signal relevant at various stages. Computation of features for different stages is elaborated in the following subsections.

2.1.1. Features for Song-Instrumental Classification: It has been indicated in [47] that unlike song, the spectrogram of instrumental music reflects a stable frequency peaks. In case of song, because of the human voice, such stability is not visible. The same observation has motivated us to look into frequency domain. In case of instrumental music, ideally the spectral power is confined around few frequency ranges. Whereas, for song, it is distributed over a wider range of frequency. Song is further complex signal as it is normally accompanied by instrumental music also. Considering all these aspects, we have relied on cepstrum based feature as presented in [15]. The

technique is particularly good at separating the components of complex signals made up of several simultaneous but different elements combined together.

Mel-frequency cepstral coefficients are short term spectral based features used by many researchers for speech recognition [44], retrieval system [12], music summarization [32], speech/music discrimination [31]. The strength of MFCC lies in its ability for compact representation of amplitude spectrum. The steps for computing MFCC as elaborated in [36] have been shown in Figure 2.

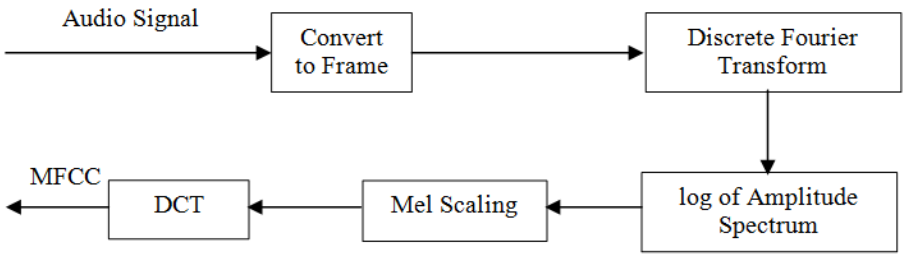


Figure 2. Steps for Computing MFCC

Audio signal is first divided into number of frames of fixed duration. Frames may consist of samples with an overlap with the previous frame. To minimize the discontinuity at the beginning and end of the frame a windowing function (Hamming window is the most widely used one) is also applied on the frame. Amplitude spectrum for each (windowed) frame is obtained by applying Discrete Fourier Transform (DFT). As the relation between perceived loudness and amplitude spectrum is more logarithmic than linear, logarithm of amplitudes is taken. Thus, N- dimensional spectrum is obtained where N is the frame size. The spectrum is smoothened to make it perceptually meaningful. The simplest way of doing this to consider the average spectrum over the frequency bins. But equi-spaced bin over the frequency scale does not conform the human auditory system as the perceived frequency and the signal frequency are not linearly related. It has led to the development of Mel frequency. The relation can be expressed as follows.

$$f_m = 2595 * log_{10}(1 + \frac{f}{100})$$

where, f and f_m are signal frequency and corresponding Mel frequency respectively. The mapping is approximately linear below 1kHz and logarithmic above. Thus, logarithm of amplitude spectrum obtained after DFT is mapped on to Mel-frequency scale and smoothened by considering the bins over the Mel-scale. The elements in the smoothened Mel-spectra vector are highly correlated. To decorrelate and to reduce the number of parameters DCT is performed to obtain Mel-frequency cepstral coefficients and first 13 coefficients are taken as features for the frame. Thus, the coefficients c[n] can be represented as

$$c[n] = \frac{1}{L} \sum_{k=0}^{L-1} | X[k] |^{j\frac{2\pi}{L}kn}$$

where, $0 < n < L-1$, X[k] be the Smoothened Mel-spectrum of input signal x[n] and L is the number of elements in the smoothened Mel-spectra vector.

After computing the MFCCs for all the frames, the vector comprising of the average value corresponding to each coefficient forms the feature descriptor. It may be noted that each Mel-frequency cepstral coefficient, c[n] is the weighted sum of all spectral component obtained after DFT and the weight varies also with n. As a result, even if a limited number of coefficients are taken as features, signature of the complete frequency spectrum is still embedded in them. Thus, MFCC provides a compact representation of the amplitude spectrum of a signal. In our experiment, we have considered first 13 coefficients.

2.1.2. Features for Instrumental Classification: The proposed scheme deals with recorded signals of single instrument. A hierarchical framework is presented to classify the signal according to the type of instrument used in generating the music. Instruments are commonly categorized as follows.

- String (Violin, Guitar etc.)
- Woodwind (Flute, Saxophone etc.)
- Percussion (Drum, Tabla etc.)
- Keyboard (Piano, Organ etc.)

Keyboard instruments have similarity in their mechanism with other category also. Piano uses string but unlike string instruments, here the string is struck by hammer. On the other hand, organ is similar to woodwind type instruments as they rely on combination of tubes with different apertures to generate the notes.

Sound produced by different instruments bear different acoustics. Sound envelopes produced by a note may reflect signature of the instrument. Shape of the envelope is determined by the variation in sound level of the note and represents the timbral characteristics. The envelope includes attack i.e. time from silence to peak, sustain i.e. time length for which the amplitude level is maintained and decay i.e. time the sound fades to silence. As in a continuous signal, it is difficult to isolate a note, a higher level features are designed that can exploit the underlying characteristics. It has already been indicated in [6] that finding the right features which can be directly used in classifying the instruments is quite difficult. As we perceive, sound generated by a string or percussion instrument persists longer till it gradually fades away completely and it is not so for a conventional keyboard or woodwind type instrument. This observation has motivated us to classify the signals first in to two groups namely type-I and type-II. Type-I consists of keyboard and woodwind whereas the type-II consists of string and percussion. At subsequent level, we take up the task of classifying the individual groups. Thus, the broad steps are as follows.

- Classify the instrumental as type-I or type-II based on wavelet based features
- Sub-classify each type based on MFCC

The computations of the features are detailed as follows.

Wavelet Based Feature: At the first level, we have opted for features that can reflect the difference in the sound envelope of the two groups of instruments as discussed earlier. Basically, the envelope is formed by the variation in amplitude. It has motivated us to look for wavelet based feature. In an image, intensity variation generates the texture pattern and wavelet based texture feature is successfully used. Same concept is also being deployed here. Audio signal is decomposed following Haar Wavelet

transform [17]. As it has been shown in Figure 3, a signal is first decomposed in low (L_1)

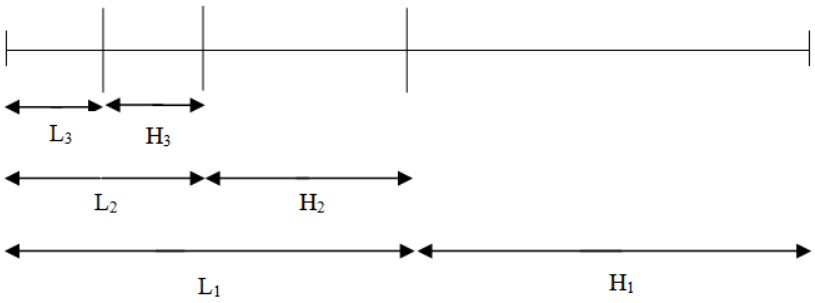
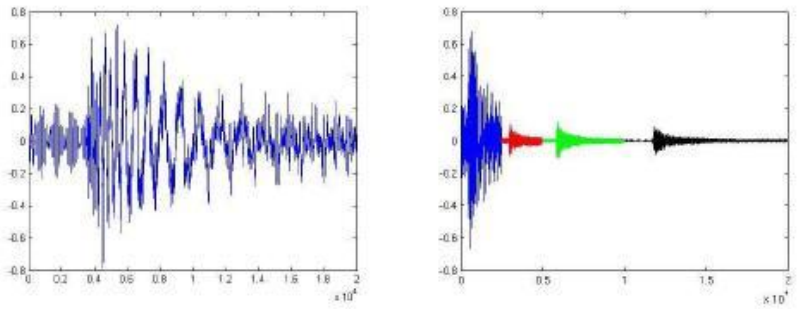
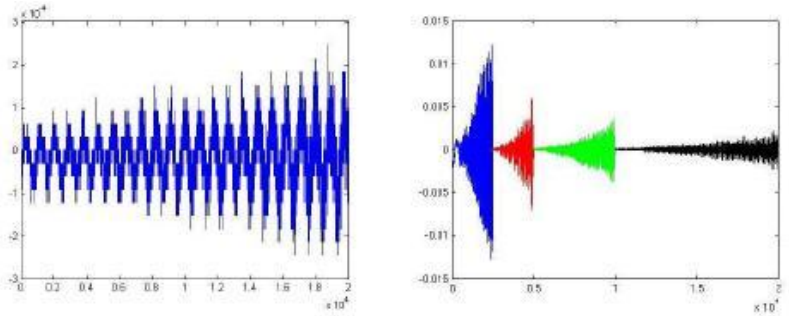


Figure 3. Schematic Diagram for Wavelet Decomposition

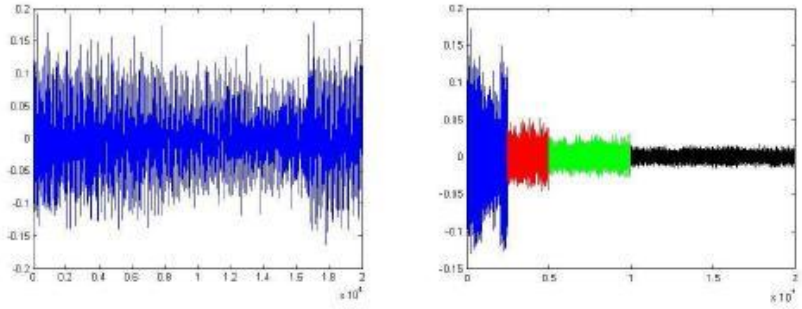
and High (H_1) bands. Low band is successively decomposed giving rise to L_2 and H_2 and so on. In general, high band contains the variation details at each level. Wavelet decomposed signals (after 3rd level of decomposition) for different types of instruments have been shown in Fig. 4. Sustain phase of audio envelope is mostly reflected in low and. On the other hand, amplitude variations during attack and decay have substantial impact on the high bands. A fast attack or decay will give rise to sharp change in amplitude in the high band and a steady rise or fall is reflected by uniform amplitude in high bands. As it appears in Figure 4, the high bands show discriminating characteristics for the two groups of instruments. There is a uniform variation of amplitudes for the first group of instruments. On the other hand, for the second group a noticeable phase of uniform amplitude without much variation is reflected.



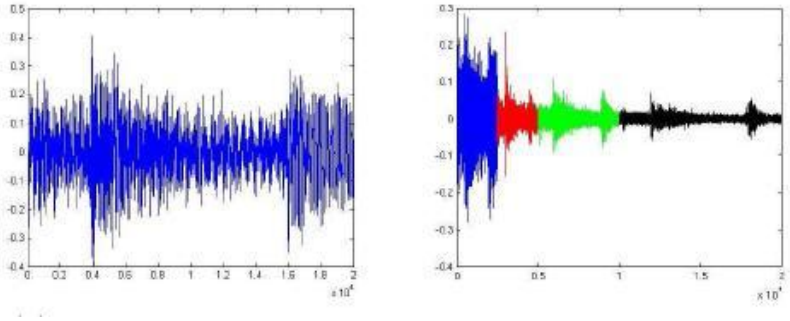
(a) Keyboard Signal and Signal after Wavelet Decomposition



(b) Woodwind signal and signal after wavelet decomposition



(c) String signal and signal after wavelet decomposition



(d) Percussion signal and signal after wavelet decomposition

Figure 4. Signal of Different Instruments and Corresponding Signal after Wavelet Decomposition

Features are computed based on Short time energy (STE) for the decomposed signals in H_1, H_2, H_3 and L_3 bands. For each band, signal is first divided into frames consisting of 400 samples. For each frame, E_i, the short time energy is computed as follows.

$$E_i = \frac{1}{n}\sum_{m=0}^{n-1}[x_i(m)]^2$$

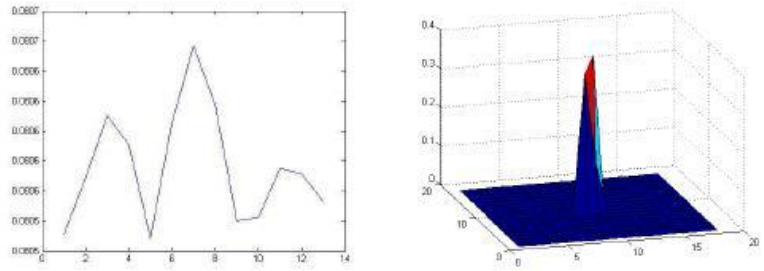
Finally, the average and standard deviation of E_is of all frames in the band are taken as feature. Thus 8-dimensional wavelet based feature is formed to use in the first level of classification.

MFCC Based Feature: For the second stage, in order to discriminate the instrument types within the groups, we have considered Mel Frequency Cepstral Coefficients (MFCC) based features. As the instruments within each type differ in terms of distribution of spectral power, such features can distinguish them. In order to capture adequate representation of signal variety required at this stage, in addition to the coefficients, we have further developed a variant of MFCC. The steps for computing the features as follows.

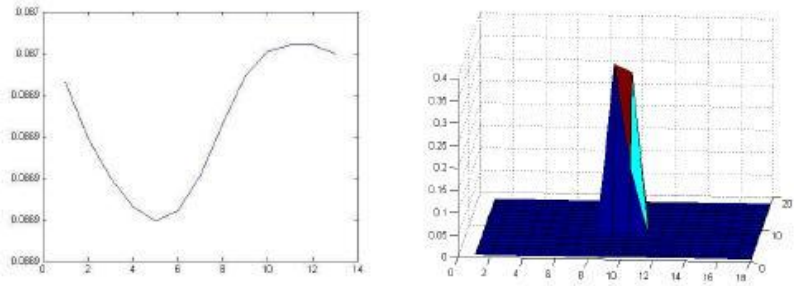
- Compute MFCC for the signal
- Compute features based on MFCC co-occurrence matrix
- Perform feature reduction

As described in section 2, signal is divided in to frames and MFCCs are computed. First 13 coefficients are considered for each frame. For each coefficient, average is taken over all frames to obtain the 13-dimensional feature. The plot of MFCC coefficients for different signals has been shown in Fig. 5. It clearly shows that the plots for a keyboard and woodwind are quite distinctive and same is also observed for a string and percussion instrument.

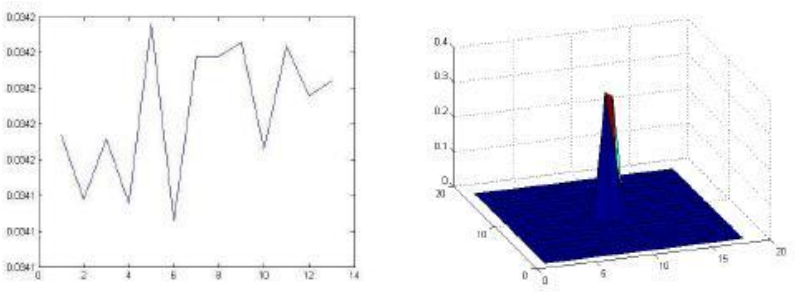
The occurrence pattern of each coefficient in MFCC in different frames provides detailed description of the signal. It has motivated us to capture the pattern in the form of a co-occurrence matrix. First of all, the coefficients values are quantized considering $\mu_i \pm k*\sigma_i$ as different quantization levels, μ_i and σ_i are the mean and standard deviation of i-th coefficient of all the frames. In our experiment k is varies from −2 to +2 with 0.25 as step size. Two other bins for values larger than $\mu_i+2\sigma_i$ and less than $\mu_i-2\sigma_i$ are also considered. Thus, coefficient values are mapped on to 18 different bins. For each coefficient, a 18×18 co-occurrence matrix, M_i is formed where an element denotes number of occurrences of a particular value (bin) pair of the i-th coefficient in two successive frames. Sample 3-D plots of co-occurrence matrices for the first MFCC coefficient for different instruments have been shown in Fig. 5. For type-I, the coefficient varies over a wider range (more number of bins) for Woodwind instruments and in case of type-II same is reflected for percussion. Finally, co-occurrence matrix based statistical measures [19] namely *energy*, *entropy*, *homogeneity*, *contrast*, *correlation* are computed which represent the pattern of the coefficients. Thus, for 13 coefficients altogether 65 such features are obtained. These features along with the coefficients form 78 dimensional features.



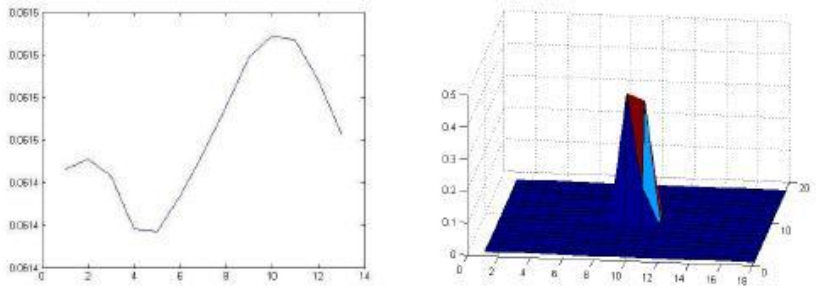
(a) MFCC plot and 3-D plot of co-occurrence of one sample coefficient of Keyboard signal



(b) MFCC plot and 3-D plot of co-occurrence of one sample coefficient of Woodwind signal



(c) MFCC plot and 3-D plot of co-occurrence of one sample coefficient of String signal



(d) MFCC plot and 3-D plot of co-occurrence of one sample coefficient of percussion signal

Figure 5. MFCC Plot and Sample Co-occurrence Matrix for Different Instrument Signals Shown in Figure 4

The feature vector of dimension 78 is high enough. Moreover, all the elements may not contribute significantly towards the discriminating process. Hence, Principal Component Analysis (PCA) is carried to reduce the dimensionality of the feature vector. In our experiment, top 20 features have been considered.

2.1.3. Features for Song Classification: Songs can be categorized in to number of genres. They differ not only in terms of rhythm, melody, beats, notes etc. but also depends on individual perception. Thus, it becomes a challenging task to classify the songs automatically based on the content inherent in the signal. Possibly, it has influenced the researchers to combine wide range of features covering different aspects of musical attributes. In this work, we have followed a hierarchical approach and restricted ourselves to conceive the features essential to meet the specific need. The methodology consists of the following steps.

- Classify the songs as classical/non-classical (popular) song
- Classify non-classical songs based on genre

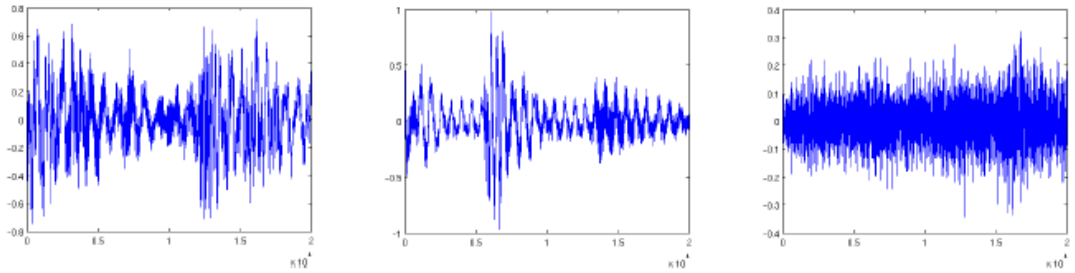
It has been observed that non-classical songs like Pop, Jazz, Bhangra (a north Indian genre) are characterized by the frequent and strong presence of beat and it is well reflected in corresponding signals. We have tried to discriminate classical and non-classical song relying on this observation. MFCC based features have been utilized for this first level of discrimination. As the popular song occupies the major share in music

collection and offers variety, it is necessary to further classify them. More likely that detailed descriptor is to be formulated to serve the purpose. We have dealt with feature based on the variation pattern of the signal amplitude.

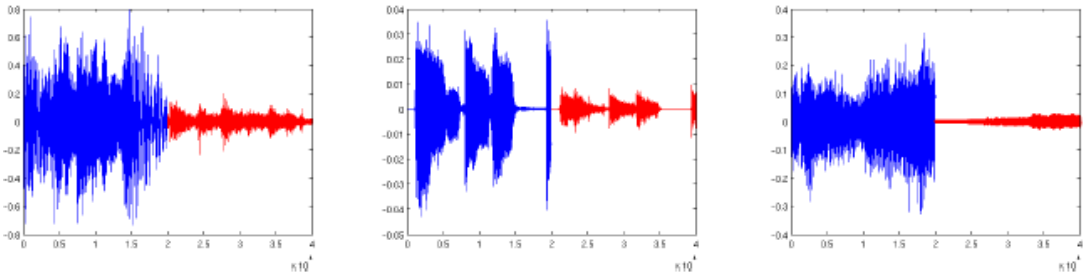
MFCC Based Feature: In non-classical/popular songs, occurrences of beats give rise to sudden change of amplitude in the signal. Depending on the beat strength and frequency, change in amplitude forms a prominent signature for popular genres whereas such pattern is very weak in case of classical song as shown in Fig. 6(a). To capture the pattern, we have first decomposed the signal using Haar wavelet transform. The decomposed signals are shown in Fig. 6(b) where left part in each represents the low sub-band and high sub-band is in the right. It may be noted that the low sub-band represents a smoothened version of the original signal whereas the amplitude variation pattern is well emphasized in the high sub-band. It has motivated us to focus only on the high sub-band for feature extraction. Thus, the steps for computation of features as presented in [16] are as follows.

- Perform wavelet decomposition on the signal
- Compute MFCC for high sub-band
- Compute features based on MFCC co-occurrence matrix
- Perform feature reduction

The high signal is first broken into number of equal sized frames and for each frame 13 dimensional MFCC coefficients are computed. MFCC obtained after taking average over all frames have been shown in Fig. 7 for different types of song and it is apparent that the plots are different for classical and non-classical song. In case of classical song, strength of the coefficients varies over a wide range and higher order coefficients are quite prominent. For the non-classical songs the coefficients are of moderate strength and confined within a range. It seems that the intervention of significant and periodic beats have smoothened the coefficients which was absent in case of classical song.



(a) Signal of Bhangra, Pop and Classical song



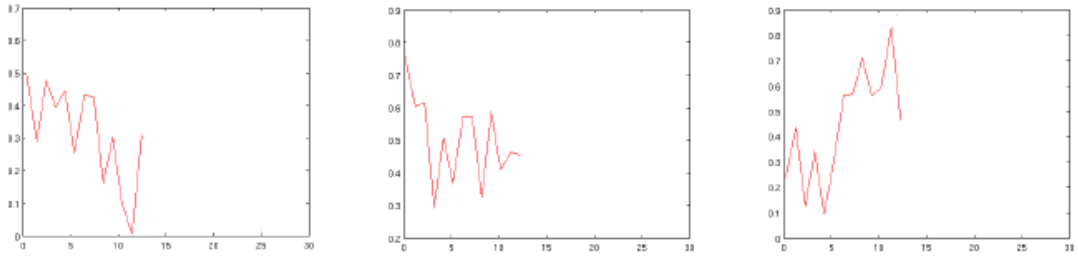
(b) Signal after Wavelet decomposition of corresponding signal shown in (a)

Figure 6. Signal of different types of song and corresponding signal after wavelet decomposition

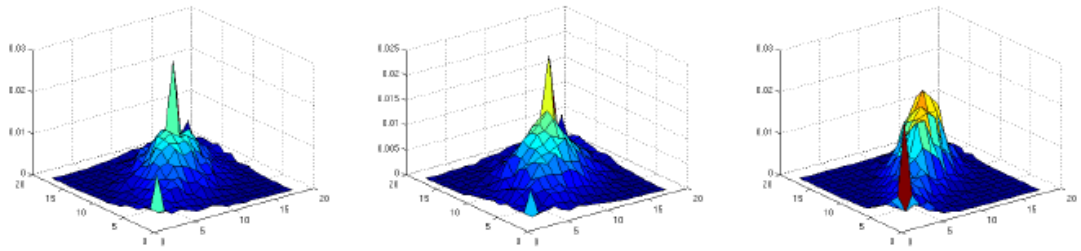
MFCCs provide a general description of the signal. In order to obtain a detailed representation, we have further designed additional features based on the co-occurrences of each coefficient over the frames as discussed in section 2.1.2. Sample 3-D plot of co-occurrence matrices for a particular coefficient for different types of song has been shown in Fig. 7(b). It clearly indicates that the co-occurrence pattern is quite distinct for classical song. As it has been mentioned in section 2.1.2, here also we perform PCA based dimensionality reduction on the MFCC and co-occurrence matrix based features to form 20 dimensional feature vector.

Amplitude Variation Pattern: In order to classify the non-classical song based on their genre, detailed signature is required that can capture the finer aspects. Genres differ in terms of one or more attributes like rhythm, tempo, and melody. Such attributes being quite perception dependent, it is quite difficult to find a measure to quantify them properly. The similar situation may be the measuring of image texture and we have drawn our inspiration from the same. In case of an image, the distribution of intensity variation gives rise to a texture pattern. Similarly, we look forward to quantify the patterns by looking in to the amplitude variation in case of an audio signal. The pattern and frequency of amplitude variation in the signal bear the impact of attributes like rhythm, tempo and melody. The feature extraction steps are as follows.

- Smoothening of the signal
- Obtain differential signal
- Compute features based on co-occurrence of amplitude variation
- Perform feature reduction



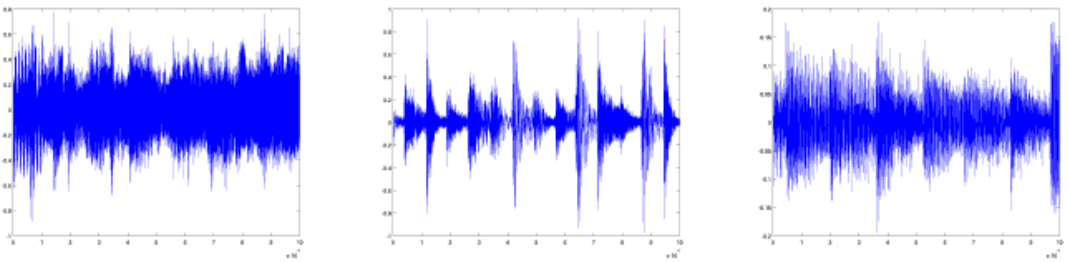
(a) MFCC plot for high sub-band signal of different song type: Bhangra, Pop and Classical



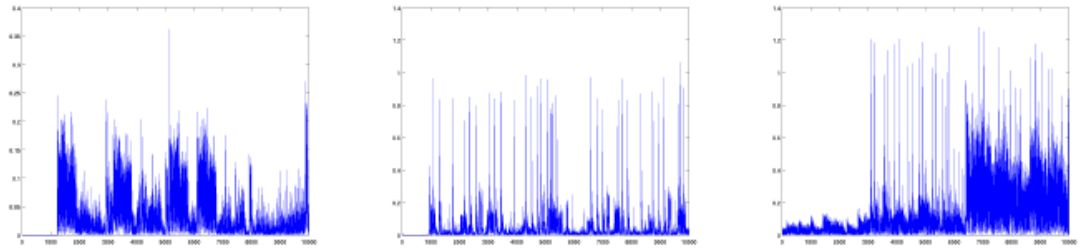
(b) Sample 3-D plot of MFCC co-occurrence matrix of different song type: Bhangra, Pop and Classical

Figure 7. Plots corresponding to different type of song signals in Figure 4

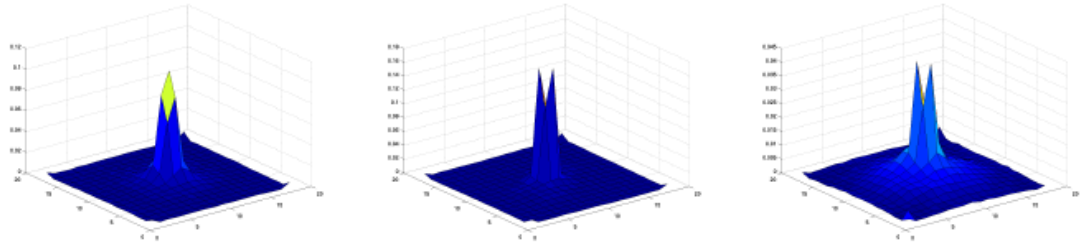
The common practice is to break the signal into frames of specific size and then the features are computed for each frame. As the size and frequency of repeating pattern in a signal is not known, selection of frame size becomes an important issue. It is more so at this stage because the classification heavily depends on such pattern and its frequency. To combat the effect researchers have dealt with dynamic frame size also [4]. But, we have dealt with the whole audio clipping without breaking it. As a result, the stationery distribution of the repeating patterns can be readily obtained. Keeping the behavior of human auditory system in mind and to minimize the effect of noise, the signal is first smoothened. The signal is broken into a time slice of 5*ms*. Without compromising the signal behavior significantly, every slice is replaced by a single sample with average amplitude computed over the slice. To obtain the differential signal, we compute $y(t) = | x(t) - x(t-1) |$ where, y(t) and x(t) are the t-th sample in the differential and smoothened signal respectively. $x(t-1)$ is the $(t-1)$-th sample in smoothened signal. Fig. 8 (a) and (b) show the original signal and differential signal for different types of genre under consideration. As the differential signal emphasizes the amplitude variation, the distinguishibility among the genres gets enhanced. Once the differential signal is obtained, 18×18 co-occurrence matrix of amplitude variation, A is formed following the similar technique as in case of MFCC based features. The matrix reflects the distribution of variation pattern as shown in Fig. 8 (c). Uniformity in amplitude variation will be accumulated along the Principal diagonal of the matrix. Elements away from the diagonal denote the occurrence of patterns with diverging variation. Finally, 18 dimensional feature vector $< s_0, s_1, \ldots, s_{17} >$ is formed where $s_i = \sum_r \sum_c A(r, c)$ such that $r - c = i$.



(a) Signal of Bhangra, Pop and Jazz song



(b) Differential signal showing amplitude variation corresponding to signals in (a)



(c) 3-D Plot of overall amplitude variation pattern corresponding to the signals in (b)

Figure 8. Different stages for measuring amplitude variation pattern

2.2. Classification

In audio classification, the variety present in the database poses the major challenge. Even though, we are dealing with only songs, a single genre may offer sufficient variation. In this work, the non-classical category includes multiple genres that further adds to the variety.

Neural network or SVM based classification schemes are quite popular. But, in the problem under consideration, variation present even within a class poses problem for neural network based classification. SVM is robust but the tuning of parameters for optimal performance is very critical. It has motivated us to look for a robust estimator capable of handling the diversity of data and can model the data satisfactorily. RANdom Sample And Consensus (RANSAC) [11] appears as a suitable alternative to fulfill the requirement.

RANSAC is an iterative method to estimate the parameters of a certain model from a set of data contaminated by large number of outliers. 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. At each iteration, a point is taken as inlier if its distance from the evolving model lies with in a threshold. In our experiment, the threshold is taken as 0.02 which is the suggested default value. RANSAC can produce reasonably good model provided a data set contains a sizable amount of inliers. It may be noted that RANSAC can work satisfactorily even with outliers amounting to 50% of entire data set. Using a subset of each category of data, model parameters are first estimated and the evolved model is used for the subsequent classification.

3. Experimental Result

In order to carry out the experiments, we have formed a music database consisting of 837 audio files. Each file contains audio of 30-45 second duration. Files are obtained from CD recordings, recordings of live programs and downloaded from various sites in Internet. The database reflects variety in terms of genre, language and quality as it contains noisy data. 501 files correspond to instrumental type with the break up as follows. 129 files represent keyboard type. 123 files for each woodwind and percussion type instrument and remaining 126 files correspond to string type instrument. In our collection we have 336 song files with the following break up. 193 files correspond to non-classical song of which 64, 62 and 67 files belong to the category of Bhangra (north Indian genre), Pop and Jazz respectively. Remaining 143 files correspond to Indian and Western classical song. Sampling frequency for the data is 22050 Hz. Samples are of 16-bits and of type mono.

Table 1. Music Classification Accuracy (in %)

Classific. Scheme	Instrumental	Song	Overall
MLP	68.46	94.04	78.72
SVM	81.83	86.90	83.86
RANSAC	94.01	90.77	92.71

For classification, apart from RANSAC classifier we have also tried with Multilayer Perceptron Network (MLP) and SVM based classifier. In MLP we have considered a single hidden layer with $\frac{n_i + n_0}{2} + 1$ nodes, where n_i and n_o are number of nodes in the input and output layers respectively. n_i is same as the dimension of the feature vectors and n_o is taken as the number of classification labels. For SVM, we have considered Radial Basis Function (RBF). Optimal values for cost parameter that governs the number of support vectors and kernel width are obtained by grid search over a wide range. For all the classification schemes, 50% of each category of data set has been used for training (estimating the model in case of RANSAC) and remaining data has been used for testing. All the experiments are repeated once again by reverting the test and training data. Finally, average classification accuracy has been considered.

Table 2. Instrumental Classification Accuracy (in %) at First Stage

Classific. Scheme	Type-I (Keyboard and Woodwind)	Type-II (String and Percussion)
MLP	81.74	85.94
SVM	88.49	85.54
RANSAC	92.06	94.78

Table 3. Instrumental Classification Accuracy (in %) at Second Stage

Classific. Scheme	Keyboard	Woodwind	String	Percussion
MLP	81.40	76.42	71.43	75.60
SVM	82.94	78.86	73.80	90.24
RANSAC	96.12	93.49	87.80	98.41

In our experiment, 13-dimensional MFCC have been used for classifying the music signals into two groups namely instrumental and song. Table 1 shows the classification accuracy of various schemes. Proposed RANSAC scheme performs quite satisfactorily. Instrumental signals are then classified following a two stage scheme. At first level those are categorized as type-I and type-II using wavelet based 8-dimensional features and subsequently each type is further sub categorized using 13-dimensional MFCC features. The performance of MLP, SVM and RANSAC based classification scheme for the two stages have been presented in table 2 and 3. In the cases, proposed methodology provides better classification.

Table 4. Song Classification Accuracy (in %) at First Stage

Classific. Scheme	Only MFCC Co-eff		MFCC Co-eff. and Co-occ. Matrix based features (after PCA)	
	Non-classical	Classical	Non-classical	Classical
MLP	73.58	76.92	78.24	79.72
SVM	75.13	84.62	89.64	90.21
RANSAC	86.53	87.41	94.82	92.31

Table 5. Song Classification Accuracy (in %) at Second Stage

Classific. Scheme	Bhangra	Pop	Jazz
MLP	73.44	69.35	67.16
RANSAC	87.50	85.48	82.09

Songs are first classified as classical or non-classical one using 20-dimensional feature vector obtained after applying PCA on 78-dimensional descriptor based on MFCC variants. Non-classical songs are then further classified as Bhangra, Pop and Jazz following 18-dimensional vector depicting amplitude variation pattern. Average classification accuracy for the two stages has been shown in Table 4 and 5. RANSAC based classification outperforms MLP network and SVM based schemes. At the second stage, it requires 3 class categorization. As SVM based methodology would have required hierarchical scheme, we have avoided it. From the experimental result it is

clear that proposed hierarchical methodology is quite effective in classifying the music signal.

4. Conclusion

In this work, we have presented a hierarchical scheme to classify the music signals in various categories. Most of the systems have relied on a wide variety of features combined together. On the contrary, the proposed scheme has dealt with the features of specific type and of very low dimension to cater the need at various stages. At the top level music signal is categorized as instrumental (music without voice) and song (music with voice) based on MFCC. Instrumental signals are classified according to the instrument type following a two stage scheme. Wavelet and MFCC based features have been utilized for the same. Songs have been first categorized as classical and non-classical ones by devising a detailed descriptor based on the co-occurrence pattern of MFCC coefficients. The descriptor has been computed using only the high sub-band component of the wavelet decomposed signal. Genre based classification of non-classical songs has been achieved using the amplitude variation pattern computed based on the differential signal. At each stage, RANSAC has been utilized as the classifier. Experimental result clearly indicates the utility of the proposed features and the classification scheme.

Acknowledgement

The work is partially supported by the facilities created under DST-PURSE program in Computer Science and Engineering Department of Jadavpur University, India.

References

[1] G. Agostini, M. Longari, and E. Poolastri. Musical instrument timbres classification with spectral features. *EURASIP Journal Appl. Signal Process*., vol. 1, **(2003)**, pp. 5–14.

[2] A. L. Berenzweig and D. P. W. Ellis. Locating singing voice segments within music signals. In *IEEE Workshop on Appln. of Sig. Proc. to Audio and Acoustics*, **(2001)**, pp. 119– 122.

[3] J. C. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *Journal of Acoustic Soc. America*, vol. 109, issue 3, **(2001)**, pp. 1064–1072.

[4] S. Hao Chen, S. Huang Chen, and R. C. Guido. Music genre classification algorithm based on dynamic frame analysis and support vector machine. In *IEEE Intl. Symposium on Multimedia*, **(2010)**, pp. 357–361.

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

[6] J. D. Deng, C. Simmermacher, and S. Cranefield. A study on feature analysis for musical instrument classification. *IEEE Trans. on System, Man and Cybernatics* – Part B, vol. 38, no. 2, **(2008)**, pp. 429–438.

[7] W. J. Dowling and D. L. Harwood. *Music Cognition*. Academic Press, Inc., **(1986)**.

[8] J. Eggink and G. J. Brown. Instrument recognition in accompanied sonatas and concertos. In *IEEE Conf. on Acoustics, Speech and Signal Processing*, **(2004)**, pp. 217–220.

[9] A. Eronen. comparison of features for musical instrument recognition. In *IEEE Workshop Appl. Signal Process. Audio Acoust*., **(2001)**, pp. 19–22.

[10] F. Fernandez, F. Chavez, R. Alcala, and F. Herrera. Musical genre classification by means of fuzzy rule-based systems: A preliminary approach. In *IEEE Congress on Evolutionery Computing*, **(2011)**, pp. 2571–2577.

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model for model fitting with applications to image analysis and automated cartography. *ACM Communications*, vol. 24, **(1981)**, pp. 381–395.

[12] J. T. Foote. Content-based retrieval of music and audio. In *SPIE*, **(1997)**, pp. 138–147.

[13] J. T. Foote and S. Uchihashi. The beat spectrum: A new aprrpoach to rhythmic analysis. In *ICME*, **(2001)**.

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

[15] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha. Instrumental/song classification of music signal using ransac. In *3rd Intl. Conf. on Electronic Computer Technology*, India, IEEE CS Press, **(2011)**.

[16] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha. Song classification: Classical and non-classical discrimination using mfcc co-occurrence based features. In *Intl. Conf. on Signal Processing, Image Processing and Pattern Recognition*, Korea, Springer, **(2011)**.

[17] C. R. Gonzalez and E. R. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall Inc., NJ, USA, **(2006)**.

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

[19] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision (Vol-I)*. Addision-Wesley, **(1992)**.

[20] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *New Music Research*, **(2000)**.

[21] D-N Jiang, L Lu, H-J Zhang, J-H Tao, and L-H Cai. Music type classification by spectral contrast feature. In *ICME*, **(2002)**.

[22] L. Kaminskyj and T. Czaszejko. Automatic recognition of isolated monophonic musical instrument using knnc. *J. Intell. Inf. Syst.*, vol. 24, no. 2-3, **(2005)**, pp. 199–221.

[23] E. Benetos M. Kotti and C. Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *ICASSP*, **(2006)**.

[24] T. Langlois and G. Marques. Automatic music genre classification using a hierarchical clustering and a language model approach. In *Intl. Conf. on Advances in Multimedia*, **(2009)**.

[25] J. Laroche. Estimating tempo, swing and beat locations in audio recordings. In *Workshop on Appln. of Sig. Proc. to Audio and Acoustics (WASPAA)*, **(2001)**.

[26] C-H Lee, H-S Lin, C-H Chou, and J-L Shih. Modulation spectral analysis of static and transitional information of cepstral and spectral features for music genre classification. In *Intl. Conf. on Intelligent Hiding and Multimedia Signal Processing*, **(2009)**, pp. 1030–1033.

[27] T. Lidy and A. Rauber. Evaluation of feature extractor and psychoacoustics transformations for music genre classification. In *Int. Conf. ISMIR*, (2005), pp. 34–41.

[28] C. R. Lin, N. H. Liu, Y. H. wu, and A. L. P. Chen. Music classification using significant repeating patterns. In *LNCS*, vol. 2973, **(2004)**, pp. 506–518.

[29] A. A. Livshin and X. Rodet. Musical instrument identification in continuous recordings. In *Intl. Conf. Digital Audio Effects*, **(2004)**, pp. 222–226.

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

[31] B. Logan. Mel frequency cepstral coefficients for music modelling. In *Intl. Symposium on Music Information Retrieval*, **(2000)**.

[32] B. Logan and S. Chu. Music summarization using key phrases. In *IEEE Conf. on Acoustics, Speech and Signal Processing*, **(2000)**.

[33] J. Marques and P. Moreno. A study of musical instrument classification using gaussian mixture model and support vector machines. *Compaq Comput. Corp., Tech. Rep. CRL 99/4*, **(1999)**.

[34] B. Matityaho and M. Furst. Classification of music type by a multilayer neural network. *Journal of the Acoustical Society of America*, vol. 95, **(1994)**.

[35] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of mpeg-7. In *Int. Comput. Music Conf.*, **(2000)**, pp. 166–169.

[36] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prectice-Hall, **(1993)**.

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

[38] J. Saunders. Real-time discrimination of broadcast speech/music. In *IEEE Conf. on Acoustics, Speech, Signal Processing*, **(1996)**, pp. 993–996.

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

[40] H. Solatu, T. Schultz, M. Westphal, and A. Waibel. Recognition of music types. In *IEEE Conf. on Acoustics, Speech and Signal Processing*, **(1998)**, pp. 1137–1140.

[41] W-H. Tsai and D-F. Bao. Clustering music recordings based on genres. In *Intl. Conf. on Information Science and Applications*, **(2010)**, pp. 1–5.

[42] Y. H. Tseng. Content-based retrieval for music classification. In *ACM SIGIR Int. Conf. res. Develop. Inf. Retrieval*, **(1999)**, pp. 176–182.

[43] G. Tzanetakis and P. Cook. Music genre classification of audio signals. *IEEE Trans. on Speech Audio Processing*, vol. 10, no. 5, **(2002)**, pp. 293–302.

[44] W. Walker, P. Lamere, P. Kwok, B. Raj, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: A flexible open source framework for speech recognition. *cmus-phinx.sourceforge.net /sphinx4/doc/Sphinx4Whitepaper.pdf*, **(2004)**.

[45] T. Zhang and C. C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, **(2001)**, pp. 27–36.

[46] T. Zhang and C. C. Jay Kuo. Content-based classification and retrieval of audio. In *SPIE Conf. on Advanced Signal Processing Algorithms, Architectures and Implementations VIII*, **(1998)**.

[47] T. Zhang and C. C. Jay Kuo. *Content-based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer Academic, **(2001)**.

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

[49] C. Zhen and J. Xu. Solely tag-based music genre classification. In *Intl. Conf. on Web Information Systems and Mining*, **(2011)**, pp. 20–24.

Authors

Arijit Ghosal

Arijit Ghosal has received his B. Tech degree in Information Technology and M. Tech degree in Software Engineering in 2006 and 2008 respectively from West Bengal University of Technology, India. He is currently working as Assistant Professor in Institute of Technology and Marine Engineering, Kolkata, India. His research interest is in the area of Signal Processing and Pattern Recognition.

Rudrasis Chakraborty

Rudrasis Chakraborty has completed B.E. in Computer Science and Engineering from Jadavpur University, India in 2010. He is currently pursuing Master degree curriculum in Computer Science at Indian Statistical Institute, Kolkata, India. His research interest includes Pattern Recognition, Artificial Intelligence and Signal Processing.

Bibhas Chandra Dhara

Bibhas Chandra Dhara received B.Sc. (Hons) degree in Mathematics and B.Tech degree in Computer Science and Engineering from University of Calcutta, India in 1997 and 2000, respectively. He earned M.Tech and Ph.D both in Computer Science from Indian Statistical Institute in 2002 and 2008, respectively. Currently, he is working as Assistant Professor in the Department of Information Technology, Jadavpur University, India. His research

area and interest include Image Processing, Video Processing and Audio Processing.

Sanjoy Kumar Saha

Sanjoy Kumar Saha Received his B.E. and M.E. Degree in Electronics and Tele-communication Engineering from Jadavpur University, West Bengal, India in 1990 and 1992 respectively and obtained his Ph.D from Bengal Engineering and Science University, India in 2006. Currently, he is working as Associate Professor in Computer Science and Engineering Department of Jadavpur University, India. His research interests are in the area of Image Processing, Video Processing, Multimedia Data Retrieval and Pattern Recognition.