On Multi-aspect Classification of Music Data

Thesis submitted by

Rajib Sarkar

Doctor of Philosophy (Engineering)

Department of Computer Science and Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India

2019

JADAVPUR UNIVERSITY
KOLKATA – 700032, INDIA

INDEX NO. **288/14/E**

1. Title of the thesis : **On Multi-aspect Classification of Music Data**

2. Name, Designation & Institution of the Supervisor : **Prof. Sanjoy Kumar Saha, Professor, Department of Computer Science and Engineering, Jadavpur University.**

3. List of Publication :

   (a) Rajib Sarkar, Soumya Kanti Naskar, and Sanjoy Kumar Saha. *Raga identification from hindustani classical music signal using compositional properties.* Computing and Visualization in Science, Springer, *pages 1–12,* 2017.

   (b) Rajib Sarkar and Sanjoy Kumar Saha. *Singer wise classification of song data using mfcc and spectrogram based vocal-print.* Communicated to International Journal of Pattern Recognition and Artificial Intelligence, World Scientific.

   (c) Rajib Sarkar, Sombuddha Choudhury, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. *Recognition of emotion in music based on deep convolutional neural network.* Revision submitted to Multimedia Tools and Applications, Springer.

   (d) R Sarkar and S K Saha. *Music genre classification using EMD and pitch based feature.* In Proceedings of the International Conference on Advances in Pattern Recognition, pages 1–6, 2015.

   (e) R. Sarkar and S. K. Saha. *Singer based classification of song dataset using vocal signature inherent in signal.* In Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, pages 1–4, 2015.

   (f) R Sarkar, N Biswas, and S Chakraborty. *Music genre classification using frequency domain features.* In Proceedings of the International Conference on Emerging Applications of Information Technology, 2018.

   (g) Rajib Sarkar, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. *Emotion based categorization of music using low level features and agglomerative clustering.* In Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, pages 506–516, 2018.

4. List of Patents : None.

5. List of Presentations in National/International/Conferences/Workshops :

   (a) R Sarkar and S K Saha. *Music genre classification using EMD and pitch based feature.* In Proceedings of the International Conference on Advances in Pattern Recognition, pages 1–6, 2015.

   (b) R. Sarkar and S. K. Saha. *Singer based classification of song dataset using vocal signature inherent in signal.* In Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, pages 1–4, 2015.

   (c) R Sarkar, N Biswas, and S Chakraborty. *Music genre classification using frequency domain features.* In Proceedings of the International Conference on Emerging Applications of Information Technology, 2018.

   (d) Rajib Sarkar, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. *Emotion based categorization of music using low level features and agglomerative clustering.* In Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, pages 506–516, 2018.

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **"On Multi-aspect Classification of Music Data"** submitted by Shri **Rajib Sarkar** who got his name registered on **25/03/2014** for the award of Ph.D. (Engg.) degree of Jadavpur University is absolutely based upon his own work under the supervision of **Prof. Sanjoy Kumar Saha** and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

_____

Signature of the Supervisior
and date with Office Seal

"... To the memory of my mother, With love and eternal appreciation. You are gone but your belief in me has made this journey possible..."

Acknowledgements

At the outset I would acknowledge the patient efforts of my supervisor Prof. Sanjoy Kumar Saha in deciding upon the problem which finally has led to this thesis. Without his iterative feedback and suggestions it would not have been possible to come up with this direction of research. I am grateful to him for his painstaking review of my work and meaningful suggestions which ultimately has given shape to this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. I am thankful to Prof. Chandan Mazumdar and the Center for Distributed Computing group, Jadavpur University, for including me in their fraternity and extending all possible resources.

I express my sincere gratitude to Sombuddha, Nimagna, Saikat, Sourajit and Aneek for their practical ideas and cooperation. I also thank Anjishnu, Soumya and Dibyadwati for their help in making me understand the characteristics of Indian classical music. Thanks are also due to the University Grants Commission (UGC) for NET-JRF fellowship to carry out my entire work without any interruption and successfully bring it to execution. Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this thesis and my life in general.

Rajib Sarkar

Abstract

Automated classification of music signal is an active area of research. It can act as the fundamental step for various applications like archival, indexing and retrieval of music data. In this work we have presented an automated system to classify a music signal based on various aspects like its *genre*, *singer*, *emotion*. For *Hindusthani* classical music, *raga* being a crucial property we have worked on *raga* identification also. The system will have two major modules like *feature extraction* and *classification*. Current work focuses mostly on the extraction of suitable low level features that can depict the characteristics meaningfully. For classification we have relied on conventional classifiers.

In broad sense *genre* reflects the style. To capture the characteristics of different genres, music signal is first decomposed to extract the component reflecting the desired degree of local characteristics using empirical mode decomposition (EMD). Pitch based features are then computed from the signal at suitable intermediate frequency range. Experimental results and comparison with other works on benchmark dataset indicate the effectiveness of the methodology.

Singer identification or singer based classification of song data is very important in the context of music retrieval. Normally a song is accompanied by background music that may cause hindrance. To address this issue we have presented a simple methodology to extract the vocal dominating segments and also to reduce the impact of the instruments. Keeping the physiological aspects of the voice production process and perceptual aspects of human auditory system in mind, features are designed to represent the voice profile of a singer. Spectrogram based vocal-print is proposed to capture the salient timbral characteristics. Mel-frequency cepstral coefficients (MFCC) based features are used as supplement.

Emotion being a perceptual and subjective concept, classification based on emotion of music is quite challenging. It is very difficult to design the low level descriptors to represent the emotion. Two methodologies are detailed in this work. In the first approach, a large feature set is considered. The set includes time domain features, spectral features, linear predictive coding and MFCC based features. Different classifiers like, neural network, support vector machine and random forest are tried. In general the performance of such approaches is limited. It is difficult to obtain a consistent feature set that works across the classifier and datasets. To get rid of these issues, deep learning based approach is tried. A conventional neural network built around VGGNet is proposed. It provides substantial improvement of performance.

For Indian classical music, *raga* is the basic melodic framework. Manual identification of raga demands high expertise which is not available easily. Thus an automated system for *raga* identification is of great importance. In this work, we have studied the basic properties of the ragas in North Indian (*Hindusthani*) classical music and designed the features to capture the same. Pitch based Swara (note) profile is formed. Occurrence and energy distribution of notes generated from the profile are used as features. Note sequence plays an important role in the raga composition. Proposed note co-occurrence matrix summarizes this aspect.

Finally, an experiment has been done for multi-aspect classification. It simply combines the methodologies developed for individual attribute based classification. Previous methodologies have been tried on datasets which are benchmarked for individual aspect. A database has been created to study the performance of multi-aspect classification strategy. Experiment on it shows that proposed methodology performs satisfactorily.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Music is one of the natural form of arts that spreads its essence over our mind. It has substantial social and physiological impact. With the advent of technology there has been enormous growth in music industry. Distribution and capturing of music data have become easier. All these lead to the huge repository of music data. There is also a high volume of consumer of music data. Hence organized storage and retrieval of music data demands attention. An automatic music information processing system for efficient classification and retrieval of music data is an important issue. Major users intend to search music of interest based on certain metadata or characteristics. Common metadata includes *genre* (music style), *singer*, *emotion*, *raga* (specially for *Hindusthani* classical music). Manual annotation is prohibitive because of large data volume. Moreover, characteristics like emotion is subjective and identification of *raga* requires expertise that is not readily available. Hence automatic identification of the characteristics can act as the fundamental step for organizing the collection and thereafter retrieving the desired data.

In the context of a music retrieval system proper organization of the large collection of music data is very important. Music data can be archived in a structured manner based on various metadata like *genre*, *singer*. One approach for extracting such metadata may be manual where domain expert annotates the piece of music. As discussed earlier manual approach has certain limitations. However, it enables text based retrieval for a metadata oriented query. The problem of annotation may be less severe as nowadays there exists different music formats with metadata embedded in it [1]. But music recorded from other sources lacks this information. A major concern arises when the user does not provide the metadata as music query. On the contrary, user may submit the music clip as the query and expecting the music with similar characteristics from the retrieval. Thus, a content based music retrieval system becomes essential which will automatically extract the properties from the query signal and compare with the same obtained from the music signal in the database. Automatic classification of music signal based on genre, singer etc. [2–5] has gained impetus over the last decade and it can serve crucial role in various applications like music retrieval and recommendation system, archival and indexing of music database, annotating a music database.

1.1 Genre Based Classification

Music genre is a conventional category to identify a pieces of music belonging to a shared tradition. It refers to a variety of facets of music. Music can be divided into different genres in many ways. Such as the period during which a musical composition was written as well as the style of the music. The instruments used in the music and the treatment applied on those instruments. The geographical origin of the music and the cultural and ethnic background plays important role in identifying the genre of a music clip.

Generally, feature based approach is followed in identifying the music genre. At first suitable features are extracted to represent genre characteristics. Subsequently features are fed to a classifier for identification process. The strength of the features in discriminating different genres plays an important role towards performance. The pioneering work of Tzanetakis et al. [6] proposed several texture and pitch feature. Even after a decade of that initial work, genre identification is an active area of research [7].

Researchers have experimented with a variety of features and classifiers. Some popular and widely used features are Mel-frequency cepstral coefficients (MFCCs), pitch, texture, chromagram, tonality, some statistical parameters derived from spectrogram etc. Among the classifiers, multi-class Support vector machine (SVM) and artificial neural network (ANN) with Multi-layer perceptron (MLP) are widely used.

Panagakis et al. [8] have used MFCCs, chromagram and wavelet transform based summarization of spectrogram. Huang et al. [9] considered wide range of features like - intensity, pitch, timbre, tonality, and rhythms. Best combination of the features was selected using self-adaptive harmony search (SAHS) method. Markov and Matsui [10] used MFCCs, line spectral pairs (LSP), timbre, spectral crest factor (SCF), spectral flatness measure (SFM) and chromagram features. For classification, a Bayesian non-parametric model was used. Schindler and Rauber [11] worked with music video and used both audio and visual based features to classify music genre. Acoustic features are statistical spectrum descriptors (SSD), rhythmic patterns, rhythm histograms, MFCCs and chromagram. Visual features include color statistics. Nanni et al. [12] used spectrogram to extract a set of texture descriptors. Three different representations (mel-scale divided, linear divided and whole) of the spectrogram were used to extract features. For further improvement [13], they have combined texture based visual features from the video.

It appears that MFCCs and pitch based features have been considered by number of researchers. But opportunity is there to improve the performance wih the help of suitable descriptors.

1.2 Singer Based Classification

A song is the composition of singing voice and accompanying instrumental music. In order to capture the singer characteristics one major issue is to extract the segments which contain only the singing voice. But it is difficult to have such segments as it is normally accompanied by instruments and the

existence of only voice is quite rare. It is observed that certain efforts are taken to remove the effect of instruments. Separating the singing voice and accompanying instrumental music is not an easy task and it is still an active research area. However, researchers have considered to focus on voice dominating segments for the purpose of singer identification. Designing the suitable descriptors is another major challenge.

Berenzweig et al. [14, 15] made an early attempt towards singer identification. In their work, a speech recognizer was used to detect the vocal segments. The use of only vocal segments instead of whole music improves the identification accuracy. Cai et al. [16] have considered sparse representation based classification to detect the vocal segments. Different auditory features based on MFCCs, linear predictive mel-frequency cepstral coefficients (LPMCCs), gammatone cepstral coefficients (GTCCs) are extracted from the vocal segments. Finally Gaussian mixture model (GMM) is used to model the singers. Su and Yang [17] proposed a system based on the idea of bag-of-frames (BOF). First of all robust principal component analysis (RPCA) is deployed to extract the voice segments. From the extracted voice segments, the log-magnitude spectrograms are encoded by l1-regularized sparse coding to obtain BOF features. Finally, SVM is used for classification.

Kroher et al. [18] considered both low level and high level descriptors. Their effort was focused towards flamenco (traditional music of southern Spain) songs. They have tried to estimate the voiced sections from the polyphonic music following pitch saliency. Along with MFCCs based timbre features, vibrato features are computed using a plug in. A transcription based high level features are also used to cope up with the improvisation. For singer identification SVM classifier is used. In their work, Hu and Liu [19] tried to separate singing voice and musical accompaniment using computational auditory scene analysis (CASA) method. In a subsequent effort [20], they have relied on spectrogram analysis to filter out instrumental accompaniment from a song. Spectrogram is decomposed into two separate matrices by employing NMPCF (non-negative matrix partial co-factorization). But, it requires prior knowledge regarding the spectrograms of pure singing voice and pure musical accompaniment. The resulting spectrogram of singing voice thus obtained still bears the impact of instruments. It is refined and reconstructed using pitch based harmonic mask estimation method. For singer identification, framewise gammatone frequency cepstral coefficients (GFCCs) are computed.

It is observed that like most of the audio applications, here also MFCCs and its variants are heavily used. Several works have relied on spectrogram. All these are indicative about the fact that frequency has a major role in discriminating the singing voices. Few recent works are either oriented for specific type of singing or relying on prior knowledge about the accompanying instrument or additional input in the form of transcript. Hence, proper low level feature based system is still an active area of research.

1.3 Emotion Based Classification

Music is associated with an emotion and accordingly it generates an intuitive feeling to the listener. Emotion conveyed by a music clip depends on the structural properties of the music clip. Tempo, melody, mode, loudness, rhythm are such properties. Tempo indicates the speed or pace of a musical excerpt. A music excerpt with fast tempo normally conveys emotions like happiness, excitement, anger, whereas slow tempo conveys sadness, serenity. Mode indicates the type of scale or tonality of the music excerpt. A music excerpt with major tonality conveys emotions like happiness, joy. Whereas minor tonality conveys sadness. Loudness is defined as the physical strength and amplitude of a music excerpt. A music excerpt with high loudness conveys emotions like anger, excited etc. whereas low loudness conveys emotions like tired, sleepy, relaxed etc. Melody is the linear succession of musical tones that the listener perceives as a single entity. A melody with complementing harmonies conveys emotions like happiness, relaxation, serenity. whereas a melody with clashing harmonies conveys emotions like excitement, anger, unpleasantness. Rhythm indicates the regularly recurring pattern of beat of a music excerpt. Music excerpt with consistent rhythm bears happiness, peace etc. Music excerpt with irregular rhythm bears emotions like amusement, uneasiness.

Identification of inherent emotion present in a music has emerged as a task [21, 22] in recent times. Emotion being a subjective issue multiple views have come up. Some consider a music to be part of one emotion [23] only. Whereas group of researchers like to assign multiple emotion tags [24, 25]. Instead of conventional handcrafted feature based approach, efforts are now directed to build deep learning architecture [26, 27].

Zhang et al. [28] proposed a feature based approach based on root mean square (RMS) energy, MFCCs, zero crossing rate (ZCR), fundamental frequency (f_0), voicing probability and few statistical parameters. Finally Random Forest is used for classification. Koch et al. [29] proposed a method for recommendation of on-line videos. The low level audio features are extracted using MIRtoolbox [23] and support vector machines (SVM) is used for classification. Gomez et al. [25] proposed a feature based multi-label music emotion detection system. The descriptor includes mean and standard deviation of spectral centroid, spectral roll-off and mel-frequency cepstral coefficients (MFCCs). K-Nearest Neighbors (kNN) is used for classification.

Labeling a music excerpt by exactly one emotional tag is not appropriate as it may have multiple emotional aspects. The best way to represent emotional aspect of music excerpt is by plotting them in a two dimensional plane, as proposed by Thayer [30] and Russell [31]. Russell [31] first proposed the two dimensional (2D) emotion model based on human psychology. Later Thayer [30] adopted that concept for acoustic domain. In 2D emotion plain, an emotion is represented by two parameters called Valence (X-axis) and Arousal (Y-axis). Researchers try to match the Valance and Arousal i.e. two axis of the 2D emotion plane by regression. Sometimes they also tagged the audio clip with an emotion, considering the valence and arousal position of the clip in the 2D emotion plane. Chen et al. [24] also considered this approach. Different spectral and MFCC based features are extracted and regressed with the valence, arousal groundtruth. Gaussian mixture model (GMM) has been used for regression.

In recent times deep learning models are being deployed for emotion tagging. So far most of the works are directed towards speech emotion recognition (SER). Architectures like convolutional neural network(CNN), recurrent neural networks (RNN) with long-short term memory (LSTM) etc. are being tried. A series of convolution and pooling operation is applied on raw audio signal or on its spectrogram representation to generate feature vectors. Huang et al. [32] used partially supervised convolutional neural network (semi-CNN) for identification of emotion in speech. Spectrogram of the speech signal is used as input. Albornoz et al. [33] used deep belief network (DBN) and restricted Boltzmann machines (RBM) to produce the essence of deep learning. They used conventional speech features for emotion recognition. The feature vector comprised of first twelve mel-frequency cepstral coefficients(MFCCs), fundamental frequency (f_0), zero crossing rate (ZCR) and energy. Coutinho et al. [34] used recurrent neural networks (RNN) as deep architecture with long-short term memory (LSTM). As feature, they used low-level acoustic descriptors (LLDs), sensory dissonance, roughness, tempo and event density. Liu et al. [27] also used CNN for music emotion recognition (MER). Weninger et al. [26] used the concordance correlation coefficients (CCCs) features. The LSTM based deep RNN with three hidden layer used to predict the valence and arousal of a audio clip.

Emotion being an psychological aspect it is very difficult to describe it in terms of low level features. The emotional impact is the culmination of lots of properties like melody, rhythm, tempo, style etc. All these factors along with the subjectiveness make the problem quite complex. Researches have combined numerous features and finally moved towards deep learning to cope up with the challenges.

1.4 *Raga* Based classification

Indian Classical Music is regarded as one of the most prestigious and the highest class of music. It is broadly divided into two categories- North Indian or Hindustani classical music and South Indian or Canatic classical music. Hindustani classical music is mainly found in Northern part of India, Bangladesh and Pakistan. *Raga* (the composition) and *tala* (rhythmic cycle) remain the central notion in both the systems. Indian Classical Music follows a particular musical style which is known as *Gharana*. Gharanas have their basis in the traditional mode of musical training and education. Each Gharana has its own style.

Raga is the underlying structure of Indian classical music and *Swara* plays the most significant role. A *swara* is similar as Note in Western Music. There are seven main *swara*s (pure notes) and five intermediate *swaras* (altered notes). Some important properties of *raga* are *Vadi* and *Samvadi*, *Arohi* and *Avrohi swara*. Each *raga* has two important kinds of notes. The most significant note (*swara*) is known as the *Vadi* and the note (*swara*) with next significance is known as *Samvadi*. *Vadi* is the *swara* on which the singer can pause for a significant time or stressing it. The sequence of *swara*s of a raga has unique ascending and descending order. The specific ascending and descending order in which the *swara*s within a *Raga* are played is called *Arohi* and *Avarohi* respectively. Another

property called *Laya* indicates the tempo or speed of a *raga*. *Raga* based retrieval is important for Indian Classical music. For such application classification of *raga* is the fundamental step. Detecting the *raga*s manually involves high level of expertise. The availability of such high profile experts is also an issue.

Automatic Raga identification efforts focused on transcript [35–37] oriented descriptors and acoustic feature [38–40] oriented descriptors. In transcript oriented method, *swara*s are segmented and identified to generate raga transcription. The transcripts are then matched with standard raga transcripts (template transcripts) [35, 37]. Raga identification using feature based method focused on develop features that can capture certain property of raga. Several low level features like MFCCs, chromagram, timbre [41], *swara*-histogram [38] and pitch [39, 40, 42–44] have been extracted to capture raga properties. Kumar et al. [39] proposed a non-linear SVM based framework. Pitch-class profile and n-gram distribution of notes are used as two types of features and for each type a kernel is used to represent the similarities between the music signals. Koduri et al. [44] put forward a methodology on first-order pitch-distribution based approach for both the Canatic and Hindustani Music. Various classifiers such as support vector machine (SVM) [39], K-Nearest Neighbors (K-NN) [44] and clustering techniques [40] are also used in recognizing the ragas. *Raga* of Indian classical music has a strong grammatical foundation. Each *raga* has well defined set of *swara* (note) sequence and properties. None of the past efforts exploited such structural or compositional properties in identifying the *raga*. Thus, addressing the task from this perspective still remains open.

1.5 Present Work

Objective:

- The objective of this work is to develop an automated system to classify music data based on *genre*, *singer*, *emotion* and *raga*.
- To build up such system we have investigated existing features and strategies. Based on the understanding we have proposed variants of certain features and developed the methodologies to fulfill the objective.

1.5.1 Datasets Used

GTZAN Genre Dataset: The GTZAN [6] dataset consists of 1000 music excerpts. The duration of each excerpt is 30 seconds. The dataset has ten genres, each represented by 100 music clips. All clips are sampled at 22050 Hz per second. The genres are - *blues, classical, country, disco, pop, jazz, reggae, rock* and *metal*.

artist20 Singer Dataset: The *artist20* dataset [45] is a subset of *uspop2002* dataset. The dataset includes the songs of twenty artists. There are six albums per artist, and 1413 songs in total.

mir-1k Singer Dataset: The *mir-1k* [46] dataset contains 1000 song clips from 110 *karaoke* (Chiness pop) songs. The songs sung by nineteen singers among them eight singers were female and eleven singers were male.

Soundtracks Emotion Dataset: The dataset [47] consists of 360 audio-clips collected from background tracks of movies with duration around 30 seconds. Each clip is annotated with different emotion class like anger, fear, sad, happy and tender. A clip can have multiple tags with confidence value.

Bi-Modal Emotion Dataset: This dataset [48] consists of 162 songs. Each song clip is of 30 seconds duration. Both, audio signal and lyrics (textual) data (hence, Bi-Modal) of the songs are available. In our work, we have considered the audio signal part and ignored the lyrics.

Raga Dataset: The dataset is prepared for *raga* identification by taking music from personal collection. Both of the vocal and instrumental based raga performance are there in the dataset. The detailed description is presented in Table 1.1. It consists of twenty four different Hindustani *ragas*. The selected *ragas* are the masterpieces of more than fifty great instrumental artists and more than twenty five great vocal artists. The dataset consists of digital raga recordings from last century to current year. All the recordings are sampled at 22050Hz and mono channeled. Instead of taking full length recording, an excerpt of 30 seconds duration is used for experiment.

Table 1.1 The *raga* dataset description.

Type	Vocal and Instrumental
Ragas	Bageshri, Bahar, Bhairabi, Bhairav, Bibhas, Bihag, Desh, Durga, Hamer, Jaunpuri, Jog, Kafi, Kalyani(yaman), Kanada, Kedar, Khamaj, Kirwani, Lalit, Malhar, Malkauns, Marwa, Purvi(Purvagauda), Sarang, Todi
Collection	The dataset contains 1648 raga clips. The duration of each clip is 30 seconds. Among them, 1190 raga clips from instrumental and 458 raga clips from vocal performances.

Multi-aspect Dataset: The dataset is prepared for multi- aspect classification by taking music from personal collection. It consists of 202 music clips. The dataset broadly annotated with tags for *Genre, Singer* and *Emotion*. It contains the recordings of six different singers - *Abbasuddin Ahmed, Asha Bhosle, Anita Saha, Pandit Jasraj, Kishore Kumar* and *Anup Jalota* with four different genres - *Folk, Rabindra sangeet, Devotional* and *Classical* over the four different emotions - *Joy, Peaceful, Romantic* and *Sadness*. To carry out the experiment, we have considered the music excerpts with 30 seconds duration, sampling at 22050 Hz, mono channel.

1.5.2 Contribution of The Work

In this work, we have explored various time domain and frequency domain features for classifying the music data collection based on different meta aspects like *genre*, *singer*, *emotion* and *raga*. Mostly, we have proposed variants of different low level features and methodologies for the purpose. We identify the following as major contributions of the present work.

- Different frequency domain features are explored for *genre* identification. Finally, empirical mode decomposition is utilized to obtain smoothed signal for extracting the pitch based music *genre* descriptor [49, 50].
- To capture the singer characteristics, energy based simple methodology is proposed to extract the voice dominating segments. Based on such extracted segments spectrogram based vocal-print is proposed as the descriptor for singer identification [51, 52].
- A simple system for emotion based classification of music signal is proposed using low level features. But to cope up with the complexity of representing emotion, a deep learning based methodology is proposed [53, 54].
- A system to identify the *raga* for Hindusthani classical music is proposed that exploits the compositional properties of such music [55].

1.5.3 Organization of The Dissertation

Automated classification of music signal is an active area of research. It can act as the fundamental step for various applications like archival, indexing and retrieval of music data. In this work we have presented an automated system to classify a music signal based on various aspects like its *genre*, *singer*, *emotion*. For *Hindusthani* classical music, *raga* being a crucial property we have worked on *raga* identification also. It has been discussed in this chapter that such systems will have two major modules like *feature extraction* and *classification*. Current work focuses mostly on the extraction of suitable low level features that can depict the characteristics meaningfully. For classification we have relied on conventional classifiers. Subsequent chapters of the dissertation elaborates the methodology for classifying the music on each individual aspect.

In broad sense *genre* reflects the style. To capture the characteristics of different genres, signal is first decomposed to extract the component reflecting the desired degree of local characteristics using empirical mode decomposition (EMD). Pitch based features are then computed from the signal at suitable intermediate frequency range. The methodology and experimental results are elaborated in Chapter 2.

Singer identification or singer based classification of song data is very important in the context of music retrieval. Normally a song is accompanied by background music that may cause hindrance. To address this issue we have presented a simple methodology to extract the vocal dominating segments

and also to reduce the impact of the instruments. Keeping the physiological aspects of the voice production process and perceptual aspects of human auditory system in mind, features are designed to represent the voice profile of a singer. Spectrogram based vocal-print is proposed to capture the salient timbral characteristics. MFCCs based features are used to supplement. Detailed description of the methodology and results are presented in Chapter 3.

Emotion based classification of music has come up as a challenge in recent times. Emotion being a perceptual and subjective concept, the task is quite challenging. It is very difficult to design the low level descriptors to represent the emotion. Two methodologies are detailed in Chapter 4. In the first approach, a large feature set is considered. The set includes time domain features, spectral features, linear predictive coding and MFCCs based features. Different classifiers like, neural network, support vector machine and random forest are tried. In general the performance of such approaches is limited. It is difficult to obtain a consistent feature set that works across the classifier and datasets. To get rid of these issues, deep learning based approach is tried. A conventional neural network built around VGGNet is proposed. It provides substantial improvement of performance. Chapter 4 details the methodology.

For Indian classical music, *raga* is the basic melodic framework. Manual identification of raga demands high expertise which is not available easily. Thus an automated system for *raga* identification is of great importance. In this work, we have studied the basic properties of the ragas in North Indian (*Hindusthani*) classical music and designed the features to capture the same. Pitch based *swara* (note) profile is formed. Occurrence and energy distribution of notes generated from the profile are used as features. Note sequence plays an important role in the *raga* composition. Proposed note co-occurrence matrix summarizes this aspect. Methodology for *raga* identification is elaborated in Chapter 5.

Chapter 6 presents a multi-aspect classification of music. It simply combines the methodologies developed for individual attribute based classification. Previous methodologies have been tried on datasets which are benchmark for individual aspect. A database has been created to study the performance of multi-aspect classification strategy. The results are summarized in this chapter.

Finally, the work is summarized in Chapter 7 by putting the concluding remarks and scope of further research.

Chapter 2

Genre Based Classification

2.1 Introduction

With the advent of different multimedia tools, it is quite easy to distribute, capture and store audio data. As a result, the size of music database grows very rapidly. To support fast browsing and retrieval of desired piece of music, it is very important to store them in an organized manner. In this context, classification of the music signal is very important. Genre being an important aspect, classification based on genre has gained impetus.

Different categories of music differ considerably in terms of rhythm, style and socio-cultural background. Usually the experts categorize the genres by framing a set of rules. This categorization process (by human experts) does not follow a universal taxonomy. Thus, the judgment being quite subjective, it is very much person specific and error prone. Moreover, the volume of data and its growth rate make the manual classification very labour intensive. Thus, an automated system is very much in demand.

The rest of the chapter is organized as follows. A survey of past work is placed in Section 2.2. In Section 2.3 the proposed methodology is elaborated. Experimental results are presented in Section 2.4. The chapter is summarised in Section 2.5.

2.2 Past Work

Music genre classification has become an interesting and popular research topic since last decade. Tzanetakis et al. [56] proposed the most significant proposal on automatic music genre classification and introduced several music features like texture feature, pitch feature etc. They proposed a way to represent the pitch feature of music signals both in symbolic and audio form. They calculated pitch histogram by applying multiple pitch detection algorithm for polyphonic signals. Xiao et al. [57] summarized the musical pattern by chroma histogram. They measure the similarity between two chroma histogram using transposition-invariant matching method. Genussov et al. [58] proposed

a geometric method to classify musical genres. They extracted timbral texture based feature. A non-linear method diffusion maps used to map the feature into Euclidean space. Tzanetakis et al. [59] used percussive patterns and bass lines patterns for genre classification. They used a clustering method based on one-pass dynamic programming and k-means clustering. Bello [60] presented a novel method for measuring the structural similarity between music. The method used recurrence plot analysis to characterize patterns. A practical approximation of the joint Kolmogorov complexity and the normalized compression distance are used to measure the structural similarity in music. Beside this several techniques have been published for genre classification such as Mayer et al. [61, 62] used Cartesian ensemble scheme based on the principle of late fusion for classification. Then they described that the lyrics domain of music could be combined with the acoustic domain. Low-level features such as spectral centroid, zero crossing rate and mel-frequency cepstral coefficients (MFCCs) are also very important to find the similarity in music [60]. Panagakis et al. [8] have used MFCCs, chromagram and wavelet transform based summarization of spectrogram (auditory cortical representations) features. For classification, they used a classifier based on joint sparse low-rank classification (JSLRR). The JSLRR based classifier is an alternative of sparse and low-rank representation which reduces noise. It also specifies the subspace where data defiled by outliers. Huang et al. [9] have used intensity, pitch, timbre, tonality, and rhythm based features. To select the best combination for the feature set, they proposed a self-adaptive harmony search (SAHS) method. The SVM classifier used for classification purpose. Markov and Matsui [10] have used MFCCs, line spectral pairs (LSP), timbre, spectral crest factor (SCF), spectral flatness measure (SFM) and chromagram features. They proposed a Bayesian non-parametric model called Gaussian Processes (GP) for classification. In their experiment, the performance of the Gaussian Processes (GP) classification model is better compared to SVM. Schindler and Rauber [11] used both audio and visual based features to classify music genre. The acoustic features they have extracted are statistical spectrum descriptors (SSD), rhythm patterns, rhythm histograms, MFCCs and chromagram. Color statistics and emotion related features taken from music videos as visual information. Nanni et al. [12] extracted a set of texture descriptors from the spectrogram. Three different representations (mel-scale divided, linear divided and whole) of the spectrogram is used to extract features. In their later work [13], they have combined visual features with the textures to improve the performance. They have used SVM for classification.

Music also contains highly structured attributes such as pitch, rhythm, tempo etc. The values of the features and their spatio-temporal structure are used to represent the music for classification purpose. For classification also number of techniques like k-nearest neighbors, Gaussian mixture model (GMM) [56], Support Vector Machines (SVM) [63], artificial neural network [64], similarity functions [65] have been tried by the researchers.

It is evident that the researchers have worked with a variety of features and classifiers. In our work, the motivation is to develop of a simple system which instead of combining various types of feature will try to focus on the use of features belonging to specific types. Pitch being one significant perceptual aspect that helps us in discriminating the genres, we have relied on it in this work.

2.3 Proposed Methodology

A set of different features has been used by researchers for music genre classification. Designing or selecting proper features is important in any classification problem. As it is observed in the early works frequency domain features dominates in identifying the genre. In this work, we consider two experiments. In the first effort we follow the conventional approach of working with a variety of frequency domain features (detailed in Section 2.3.1). But the classification accuracy is limited and hence we explore the challenge further. Finally, we propose an improved methodology that performs a pre-processing in the form of empirical mode decomposition (EMD) of the music signal to extract smoothed intermediate signal. Subsequently pitch based features are computed from such intermediate signal(s) (detailed in Section 2.3.2).

2.3.1 Approach Based on Low level Frequency Domain Features

To compute the low level features, audio clip is divided into number of frames of short-term duration. Frame level features are summarized to describe the audio clip in a meaningful manner. In this approach, we have relied on frequency domain features to design the descriptors. The features under consideration have been categorized as timbre, tonality, pitch and statistical features.

Timbre Features

Timbre feature of a music signal indicates the quality and texture of the signal. It describes the spectral shape of a music signal. Timbre is represented using different spectral features and mel frequency cepstral coefficients (MFCCs). To extract the spectral features, audio signal is split into frames consisting of W (here, it is taken as 512) samples with half overlap between successive frames. For each frame spectrogram is first obtained. Let $\mathscr{S}(i,n)$ denotes the i-th spectral component in the spectrogram for n-th frame. In our work, dimension of the spectral component (K) is 2000. Spectral features are then computed as follows.

Spectral Flux (SF): Spectral flux indicates the amount of changes or variations reflected in spectral shape. For n-th frame, the spectral flux is computed as:

$$SF(n) = \frac{\sqrt{\sum_{i=0}^{K-1}(|\mathscr{S}(i,n)| - |\mathscr{S}(i,n-1)|)^2}}{K} \tag{2.1}$$

where, $\mathscr{S}(i,n)$ denotes the i-th spectral component in the spectrogram for n-th frame, $K = 2000$ is the dimension of the spectral component. The result of the spectral flux is a value within the range $0 \leq SF(n) \leq \mathscr{S}_m$ with $\mathscr{S}_m$ representing the maximum possible spectral magnitude. It captures changes in the power of spectral components over the successive frames.

Spectral Rolloff (SR): It is defined as the q^{th} percentile of the power spectral distribution [66]. SR is identified as the frequency bin for which the overall power spectrum of $\mathscr{S}(i,n)$ covers q percent of

the total power spectrum. In our case q is taken as 85. Mean and standard deviation of the SR point over the frames are taken as features.

Spectral Centroid (SC): The Spectral Centroid of an audio signal represents the center of gravity of the spectral power. SC is commonly accepted as a measure for brightness of the music signal. It is the ratio of the frequency weighted magnitude spectrum with unweighted magnitude spectrum.

$$SC(n) = \frac{\sum_{i=0}^{K-1} K \times |\mathscr{S}(i,n)|^2}{\sum_{i=0}^{K-1} |\mathscr{S}(i,n)|^2} \tag{2.2}$$

where, $\mathscr{S}$ is the spectrogram, K is window length.

Spectral Spread (SSP): Spectral spread also known as instantaneous bandwidth. It measures the centralism of the spectral power about the spectral centroid (SC). It is calculated as

$$SSP(n) = \sqrt{\frac{\sum_{i=0}^{K-1} (i - SC(n))^2 \times |\mathscr{S}(i,n)|^2}{\sum_{i=0}^{K-1} |\mathscr{S}(i,n)|^2}} \tag{2.3}$$

Spectral Slope (SSL): It is the measurement of slope of a spectral shape. SSL is measured by taking linear approximation of magnitude spectrum. It is calculated as

$$SSL(n) = \frac{\sum_{i=0}^{K-1} (i - \mu_i)(|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})}{\sum_{i=0}^{K-1} (i - \mu_i)^2} \tag{2.4}$$

where, $\mathscr{S}$ is the spectrogram, K is window length, $\mu_{\mathscr{S}}$ is the overall mean of spectral magnitude of the spectrogram and μ_i is the spectral component.

Once the frame level spectral features are computed, those are summarized to obtain the clip level descriptors. For each feature, its mean and standard deviation over the frames are considered.

Mel Frequency Cepstral Coefficients (MFCCs): The extraction procedure of the MFCCs [67] is based on hearing perceptions of human cochlea. To calculate MFCC features, the music signal is divided into number of frames using windowing function. Fast Fourier transform (FFT) [68] algorithm used to transform these time domain music frame into frequency domain. The frequency domain form of all frames of the music signal represents frequency spectrum. A series of triangular filter bank is designed according to mel scale that matches human auditory system. The equation to convert a frequency f in hertz to m in mel scale is

$$m = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$

Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency. The mel filter bank is applied on the frequency spectrum. The output of each filter is the sum of the filtered spectral components of that filter. The output of all filter is called mel spectrum. Finally, discrete cosine transform (DCT) [69] is applied on mel spectrum which gives the mel frequency cepstral coefficients. For each frame, thirteen coefficients are considered. Finally, mean value of each co-efficient over the frames are used as clip level descriptor.

Tonality Features

The measure of tonality estimates the harmonic context or periodicity of a music signal. A music signal with good tonality expected to have low noisy and high periodicity elements. To represent tonality, we rely on the spectrogram based spectral features.

Spectral Flatness Measure (SFM): It is the proportion of geometric mean and arithmetic mean of a magnitude spectrum [70, 71], as shown below,

$$SFM(n) = \frac{K \times \sqrt[K]{\prod_{i=0}^{K-1} \mathscr{S}(i,n)}}{\sum_{i=0}^{K-1} \mathscr{S}(i,n)} \tag{2.6}$$

where, $\mathscr{S}$ is the spectrogram, K is window length. Mean and standard deviation of SFMs over the frames are considered as the features. For uniform (flat) distribution of power spectral component it provides higher value.

Spectral Crest Factor (SCF): It is the measurement of the quality of a acoustic signal [71]. It is computed as the proportion of highest of the power spectrum with total power spectrum.

$$SCF(n) = \frac{\max\limits_{0 \le i \le K-1} |\mathscr{S}(i,n)|}{\sum_{i=0}^{K-1} |\mathscr{S}(i,n)|} \tag{2.7}$$

where, $\mathscr{S}$ is the spectrogram, K is window length.

Tonal Power Ratio: The tonal power ratio calculated by taking the ratio of the power of tonal components with the total power of the magnitude spectrum.

Individual frame level tonality features are also summarized as their mean and standard deviation over the frames.

Pitch Based Features

Each musical note is recognized with its corresponding musical instrument digital interface (MIDI) pitch. To generate pitch profile, a music clip is decomposed into number of frequency bands. The frequency bands are corresponding to the center frequencies of the MIDI pitches. A note's associated frequency is called center frequency. For each pitch, a appropriate multi-rate filter bank comprising of elliptic filters is designed. These bandpass filter passes all frequencies around a note's respective center frequency. Finally, for each band, local energy content is measured by computing short-time mean-square power (STMSP). Here only 88 MIDI pitches ($A0$ to $C8$) considered as others are not perceivable to human auditory system. From this pitch representation, pitch histogram features are calculated as follows:

- for each of 88 pitch band, the STMSP Values are quantized. The process is
  - calculate mean (μ_p) and standard deviation (σ_p) of STMSP values in each band.
  - Quantize each STMSP values as $\mu_p + k \times \sigma_p$, where $-2 \le k \le 2$, with step size 0.5.

- for each pitch band, normalized histogram prepared from the quantized STMSP values.
- The histogram for all pitch bands are concatenated. From this, 15-dimensional features are taken after PCA operation.

Statistical Features

Spectral Kurtosis (SK): The spectral kurtosis summarizes the existence of series of momentary variation in frequency and their locations in a spectrogram. The spectral kurtosis is the normalized fourth-order moment of the spectrogram. It indicates how Gaussian the magnitude spectrum distribution looks like. It is calculated as

$$SK(n) = \frac{\sum_{i=0}^{K-1}(|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})^4}{K \times \sigma_{\mathscr{S}}^4} \tag{2.8}$$

where, $\mu_{\mathscr{S}}$ is the mean of spectral magnitude and $\sigma_{\mathscr{S}}$ is the standard deviation of the spectrogram.
Spectral Skewness: It is the ratio of third central moment of the spectral components and the cube of its standard deviation. It is calculated as

$$SSK(n) = \frac{\sum_{i=0}^{K-1}(|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})^3}{K \times \sigma_{\mathscr{S}}^3} \tag{2.9}$$

where, $\mathscr{S}$ is the spectrogram, K is window length, $\mu_{\mathscr{S}}$ is the mean of spectral magnitude and $\sigma_{\mathscr{S}}$ is the standard deviation of the spectrogram. Here also, mean and standard deviation of individual frame level features are considered at the clip level.

Finally neural network with single hidden layer is considered as the classifier. As the classification accuracy is moderate (see Table 2.1 in Section 2.4), we have studied the problem further and propose a solution elaborated in the following section.

2.3.2 Approach Based on Empirical Mode Decomposition (EMD)

The proposed methodology consists of two major modules namely, *feature extraction* and *classification*. Feature extraction module describes the signal content in the form of a feature descriptor. The descriptors should be capable of discriminating the genres. Again,it must not be sensitive to the variations present with in a genre. Classification module uses the descriptors as input and identifies the genre. The feature extraction module consists of two parts:

1. *Empirical Mode Decomposition (EMD)* - This may be considered as a pre-processing step. Each music signal is decomposed by applying EMD. It generates number of intrinsic mode function (IMF) which contains local characteristics of the music. The k^{th} IMF of the music signal is used for further processing. So that, the details are suppressed considerably and the core aspect of the genre category is revealed.

2. *Pitch representation* - A feature vector is defined for each music clip. It captures the pitch based feature corresponding to decomposed signal. The steps are as follows.

   (a) Generate pitch signal for decomposed music.

   (b) Quantize the local energy values of the pitch signal into bins.

   (c) A normalized histogram of quantized energy values is prepared for each pitch band. Finally the feature vector is obtained by concatenating the histograms of all the pitch bands.

Feature Extraction

Different genres differ strongly in their frequency content. Pitch being the perceptual form of the frequency we have focused on pitch based features. A time varying music signal consists of different frequency components. Noise or abrupt variations are captured in the high frequency part whereas the low frequency components lacks the detail characteristics of the genres. Inclusion of very fine details in feature computation may bear the impact of noise and it may be sensitive towards variations present in the same genre. On the other hand, the low frequency components may lose the potential to distinguish the genres. Hence, it is important to decompose the signal and work with component(s) suitable for the purpose. Hence, the computation of pitch based feature is preceded by the signal decomposition phase.

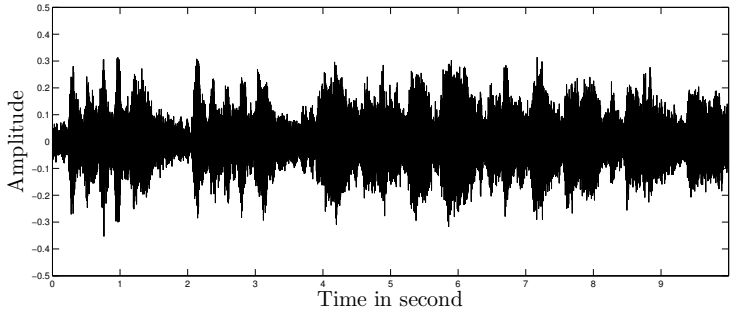
Empirical Mode Decomposition(EMD): Fourier transform and wavelet transform are widely used for decomposing the signal. But, none of them use an adaptive transform basis. Moreover, the signals may not be stationery. In this context, empirical mode decomposition (EMD) is a better alternative. It is a decomposition technique suitable for non-linear, non-stationery data analysis.The major advantage of EMD is that the basis function is derived dynamically from the signal itself.

EMD generates number of intrinsic mode functions (IMF) and a residual (remainder). IMFs represent the details or local characteristics at various scales whereas residual denotes the approximation (low frequency components). Figure 2.1 shows a sample music signal and few IMFs obtained from it by applying EMD. An algorithm for EMD is presented in [72]. The steps are summarized as follows.

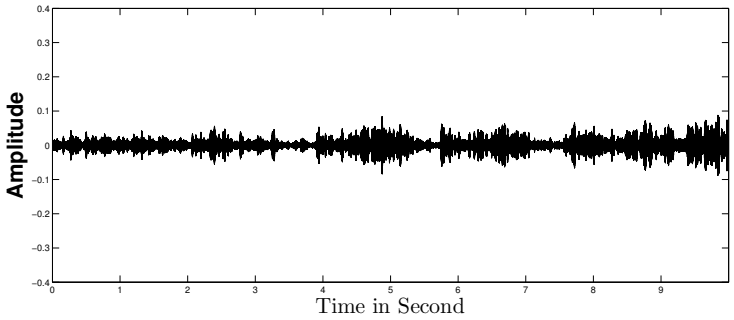
- Two smooth envelopes are created from the given sequence (signal). By joining local maxima (minima) of the sequence, upper (lower) envelope is formed. This requires the identification of all local extrema that are further connected by cubic lines to produce the upper and the lower envelopes.

- The mean of these envelopes are subtracted from the initial sequence to obtain a resultant sequence. It results in to the extraction of required empirical function in the first approximation.

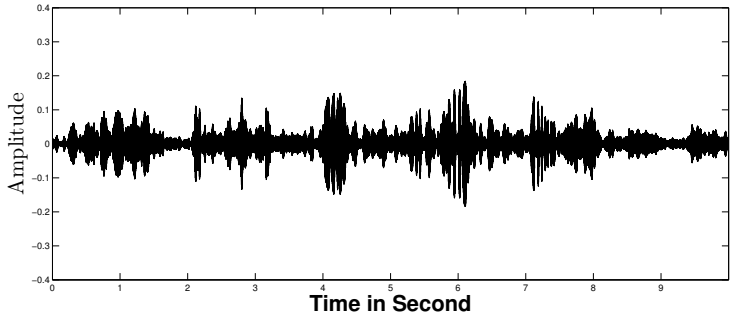
- The previous steps are repeated on this resultant sequence. This is known as shifting process and it is repeated multiple times (may be predefined).



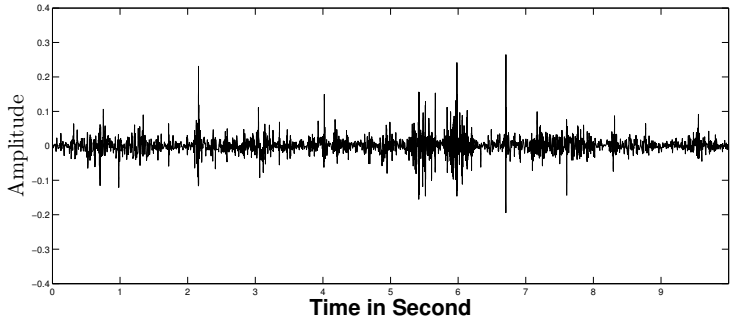
(a) music signal



(b) IMF 4



(c) IMF 7



(d) IMF 10

Fig. 2.1 A time-amplitude representation of a Classical music clip and its IMFs.

- Once multiple shifting stage is over the resultant sequence corresponds to first IMF.
- Residue is obtained by subtracting the IMF from the original signal.

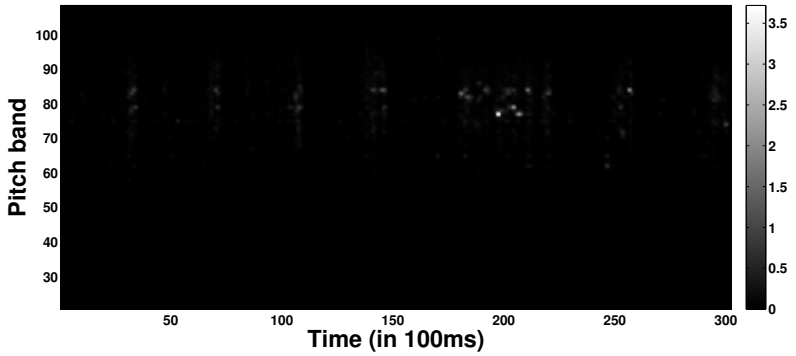
The first IMF corresponds to the shortest period component of the data and the residue contains the information of longer period duration. The process may be continued by considering the residue signal as the original one for next iteration to obtain the next IMF. Subsequent IMFs will represent the characteristics of relatively lower frequency components. Thus, for a moderate value of k, the k^{th} IMF will reflect the signature of intermediate frequency component. It will be free from both, the initial high frequency components arising out of noise or additional ornamentation of the artist and also the non discriminating low frequency components. In our experiment, k is empirically chosen as 10.

Pitch Representation: A signal is decomposed into IMFs. We consider k-th IMF to extract the pitch information. Chroma Toolbox [73] is used to decompose the chosen IMF into pitch bands, where each band corresponds to a pitch of the equal-tempered scale.

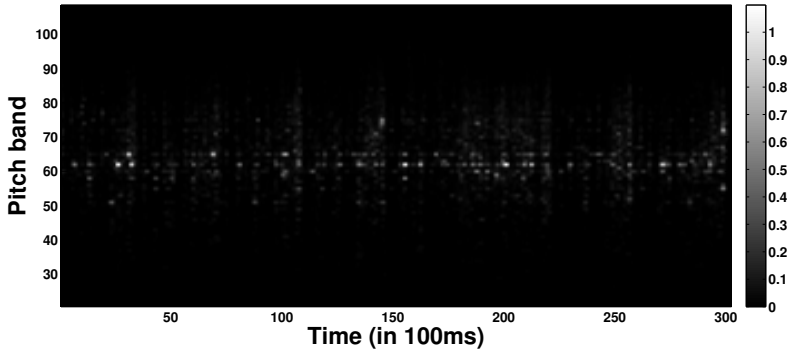
A musical note is identified with its corresponding MIDI pitch p. For example, the note A0 is defined with MIDI pitch $p = 21$, A4 is defined with $p = 69$ *etc*. The associated frequency of a note is referred as center frequency of the note. We restricts ourselves within the MIDI pitch ranges, that is perceivable to human auditory system *i.e.* from $p = 21$ (A0) to $p = 108$ (C8). For each pitch, a bandpass filter is designed which passes all frequencies around its respective center frequency. To distinguish adjacent notes, the width of the bandpass filters are kept narrow. The elliptic filters, described in [74], are used for their excellent cutoff properties. An array of bandpass filters for each pitch is defined to decompose the input signal into several pitch bands. These filters form *pitch filter bank*.

To obtain the pitch representation, *pitch filter bank* is applied to the IMF signal. Then for each pitch band, local energy or short-time mean-square power (STMSP) is computed. A compact representation of the local energy distribution in the sub-bands is shown by a time-pitch plot in Figure 2.2 where y-axis corresponds to different pitch bands and x-axis denotes the time. The colour denotes value of STMSP. The plots corresponding to different IMFs for a music signal are shown. It may be noted that plot for 10th IMF provides more information in comparison to previous IMFs. Figure 2.3 shows the similar plots for different genres. It may noted that distribution of STMSP varies for different genres. The genre *blues* in the figure has a smooth distribution over pitch bands 30 to 40, *pop* has the distribution over pitch bands 30 to 60, whereas *classical* has discontinuous distribution over pitch bands 21 to 60. Thus, the distribution of STMSP over various bands is utilized in obtaining the descriptor.

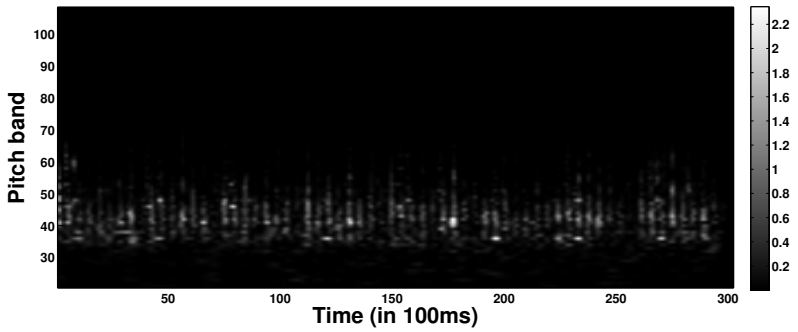
The steps for deriving the features are similar to those described in Section 2.3.1. For each MIDI band, STMSP values are quantized into number of bins with $\mu + k\sigma$ as the quantization levels and k varies from -2 to $+2$ with step size 0.5. The μ and σ correspond to mean and standard deviation of STMSP values in the band. For each band the normalized histogram of quantized STMSP values is formed. Histograms of all the bands are concatenated to form the descriptor. For each pitch-band, STMSP values are actually quantized into ten bins. It has been observed that the contribution of



(a) Pitch representation of 5th IMF

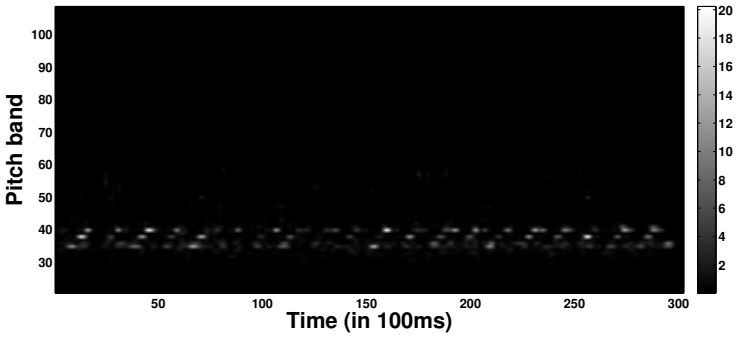


(b) Pitch representation of 7th IMF

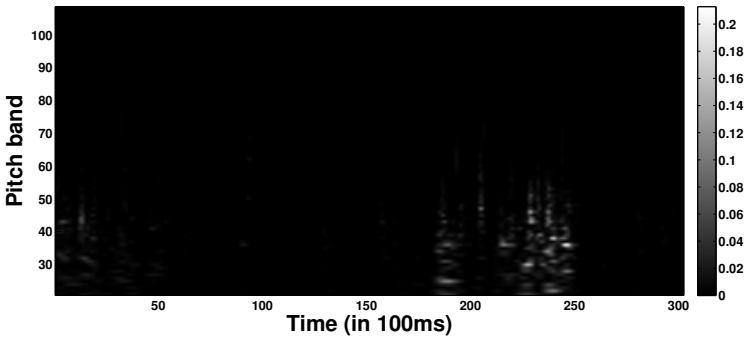


(c) Pitch representation of 10th IMF

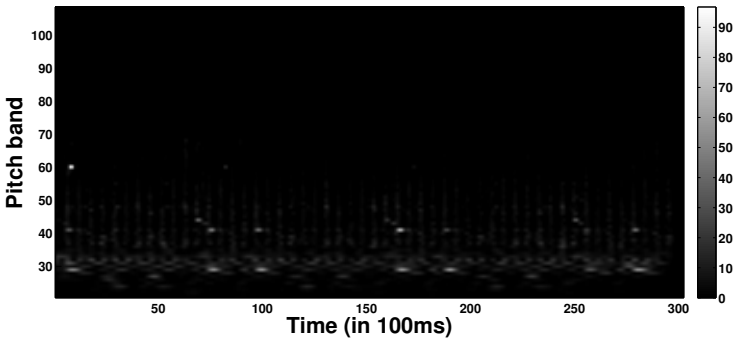
Fig. 2.2 Pitch representation of different IMFs of Disco music clip. Color bar denotes STMSP.



(a) Sample Pitch representation of Blues Genre



(b) Sample Pitch representation of Classical Genre



(c) Sample Pitch representation of Pop Genre

Fig. 2.3 Sample Pitch representation of different genre. Color bar denotes STMSP.

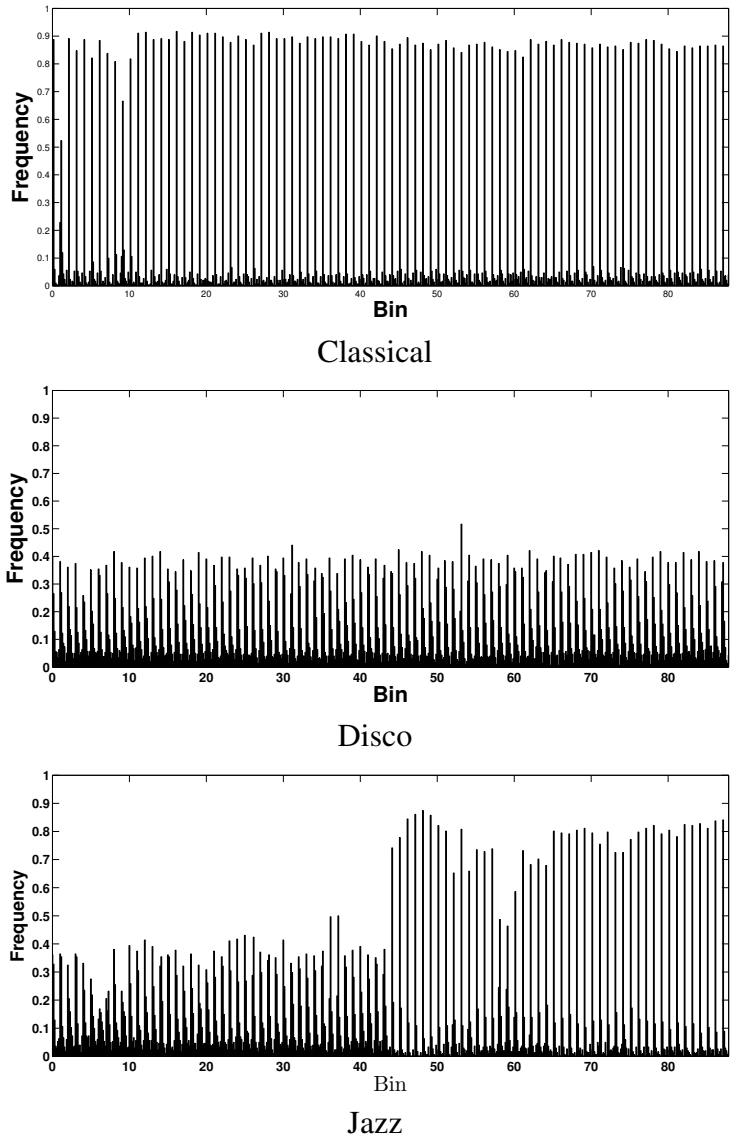


Fig. 2.4 Sample normalized pitch histograms in feature vector of different genre.

the first three bins are marginal. Hence those are ignored and finally, 616 dimensional histogram is obtained and used as the descriptor. Figure 2.4 shows the histograms corresponding to sample signals of different genre. It is clear that the distribution are visibly different.

Classification

In our work, we have used an artificial neural network with multi layer perceptron for classification. In an artificial neural network, nodes are connected together to form a network which mimics a biological neural network. A class of statistical models is called neural if they possess the following characteristics:

- It consist of sets of adaptive weights, *i.e.* numerical parameters that are tuned by a learning algorithm.
- They are capable of approximating non-linear functions of their inputs.

The adaptive weights (connection strength between nodes) are activated during training and prediction. Multi-layer perceptron (MLP) algorithm created by Rosenblatt [75], has multi-layer learning network. The neural network with back propagation [76] can effectively applicable for pattern recognition and classification problem.

The first layer is the input layer consisting of the number of nodes same as the dimension of the feature vector. Last layer is the output layer having number of nodes same as the number of labels into which the input has to be categorized. In between there may be number of hidden layers with suitable number of nodes. Nodes in a layer send data via synapses to the nodes in the next layer. Starting from input layer it reaches to the output layer. The synapses store parameters called adaptive weights which manipulates the data in calculation. A neural network is typically defined by three types of parameters:

- The interconnection pattern between the different layers of nodes.
- The learning process for updating the adaptive weights of the interconnections (training), and
- The activation function that converts a node's weighted input to its output activation (testing).

Determining the number of hidden layers and number of nodes in those are important issue. A rough estimation can be made as follows. In a network with single hidden layer, number of nodes in it (N_{hid}) can be approximated as [77]:

$$N_{hid} = \sqrt{N_{in} * N_{out}} \tag{2.10}$$

N_{in} and N_{out} denote number of nodes in the input and output layer respectively. In case of two hidden layers, number of nodes in first and second hidden layer (N_{hid1} and N_{hid2} respectively) can be taken as [77]:

$$N_{hid_1} = N_{out}\left(\sqrt[3]{\frac{N_{in}}{N_{out}}}\right)^2;\ N_{hid_2} = N_{out}\left(\sqrt[3]{\frac{N_{in}}{N_{out}}}\right) \tag{2.11}$$

As the feature vector is of 616 dimension, we have considered 616 nodes in the input layer and the output layer contains 10 nodes as we have dealt with ten different genres. For single hudden layer number of hidden layer nodes is taken as 78. For two such layers number of nodes are 156 and 36.

Table 2.1 Performance of the proposed methodologies on GTZAN dataset.

Features	**Classification Accuracy (in %)**
Approach I: based on Timbral, tonal, pitch and statistical features	79.30
Approach II: Pitch based features of intermediate IMF	97.70

2.4 Experimental Results

Experiment has been conducted on GTZAN dataset [56]. The dataset has 10 genres: *blues*, *classical*, *country*, *disco*, *hiphop*, *jazz*, *metal*, *pop*, *reggae* and *rock*. Each genre has 100 clips of 30 seconds duration. All signals are sampled at 22050 Hz and are of type mono. For each genre 90% clip is randomly chosen as training data and the remaining clips are considered as test data. The process is repeated by selecting the training data randomly for 10 times. The experiment is conducted for both the approaches. First one is based on low level frequency domain features representing timbre, tonal, pitch and statistical features. Second one deals with pitch based descriptor computed from tenth IMF extracted through EMD. Classification accuracy for the two approaches are shown in Table 2.1. It is clear that first approach performs moderately whereas the other one improves the performance significantly. For second approach, the average classification accuracy obtained with single hidden layer is 95.40%. The performance improves with two hidden layers and average accuracy becomes 97.70%. The corresponding confusion matrix is shown in Table 2.2.

Number of researchers have reported their results on GTZAN dataset. We have compared the performance of proposed methodologies with those. The classification accuracy along with the features and classifier used by the systems are summarized in Table 2.3. It is observed that the proposed approach I is better than some of those work and performance of approach II is substantially better than most of the works. It may also be noted that proposed work relies only on pitch based feature and use of EMD helps in extracting the desired signal component which is rich in pitch content.

Table 2.2 Confusion matrix for approach II on GTZAN dataset.

	1	2	3	4	5	6	7	8	9	10
1	99	0	0	0	1	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0
3	0	3	95	0	0	1	0	0	0	1
4	1	0	0	98	0	0	0	1	0	0
5	0	1	2	0	95	1	0	0	0	1
6	0	0	0	0	1	97	2	0	0	0
7	0	0	0	0	0	0	100	0	0	0
8	1	0	0	0	0	0	0	99	0	0
9	0	0	0	0	2	0	0	0	98	0
10	0	2	1	0	0	1	0	0	0	96

(1-Blues, 2-Classical, 3-Country, 4-Disco, 5-Jazz, 6-Hiphop, 7-Metal, 8-Pop, 9-Reggae, 10-Rock)

Table 2.3 Comparison of performance.

Methodology	**Feature**	**Classifier**	**Accuracy (in %)**
Tzanekis and Cook[78]	Timbre, Baseline, Rhythm	SVM	76.10
Silva et al.[79]	MFCC	LPBHN	62.10
Lykartis and Lerch[80]	Pitch, Chroma, Spectral Spread, Peak Amplitude Value, ZCR	SVM	72.80
Panagakis et al. [81]	Timbre, pitch, temporal	SVM	84.30
Kotropoulos et al. [82]	Timbre, pitch, temporal	LDA	84.96
Lee et al. [83]	Timbre	LDA	90.60
Arabi et al. [84]	Timbre, beat, chord	SVM	90.79
Panagakis et al. [85]	Timbre, pitch, temporal	SRC	93.70
Sigtia et al. [86]	FFTs on frames	Deep NN	83.00
Huang et al. [9]	Intensity, pitch, timbre, tonality, rhythm	SVM	97.20
Proposed Approach	Pitch of IMF	NN	97.70

2.5 Summary

In this work, we have proposed a simple methodology for genre based classification. First of all we have experimented with number of frequency based low level features that performs moderately. Finally in the second approach empirical mode decomposition is utilized in extracting desired signal component by ignoring the extreme (high and low frequency) characteristics. From that extracted signal, pitch based feature vector is formed considering the local energy distribution over various pitch bands. Multi-layer perceptron network is used for classification. Experiment is carried out with the GTZAN dataset. Comparison of performance with other contemporary works indicates the effectiveness of the proposed methodology.

Chapter 3

Singer Based Classification

3.1 Introduction

Singer is a dominating metadata in vocal music. In the context of music information retrieval, singer based search is quite frequent. Very often listeners are interesting in finding the songs of the artist of his/her choice. Moreover, instead of the metadata (singer in this case) user may submit a query music clip and then may like to retrieve the pieces of same singer from the collection. So direct metadata based search can not help. From the content of the music signal itself, the signature for singer identification has to be extracted. Thus, an automated system for singer identification becomes essential to organize the music collection accordingly and to cater singer based music retrieval. Normally, there are accompanying instruments with the vocal. In an automated system, it may be essential to minimize the impact of such instruments for capturing the vocal characteristics.

In this chapter, we present a novel scheme for automatic classification of song based on singer and the methodology can also be used to identify the singer of a song among a given set of candidate singers. The rest of the chapter is organized as follows. Section 3.2 presents a brief survey on singer based classification system. Proposed methodology is elaborated in Section 3.3. Experimental results are discussed in Section 3.4 and finally the summaries are put in Section 3.5.

3.2 Past Work

Singer based classification of music data or identification of singer is an active area of research. Broadly the approaches may be categorized as textual metadata based or signal characteristics based system. We focus our attention on the signal characteristics based approach. A content based system mainly consists of two modules namely *feature extraction module* and *classification (identification) module*. Feature extraction module focuses on automatic signature generation for the singer from the signal content and classification (identification) module uses the signature to classify the song data based on singer or to identify a singer from a known set of singers. It is quite rare to get the data of a

solo performance of a singer. Normally accompanying instruments are also present. Thus, a song signal consists of vocal dominating components and components with emphasized background music. As the acoustic features of a singer are to be extracted from the vocal section, a pre-processing step is required to segment the vocal and non-vocal components. It may also be noted that it is quite difficult to extract the pure voice signal from a signal mixed with voice and instruments. Hence the use of traditional speech recognition technique is not fruitful. Moreover speech and singing voice are quite different. Characterizing the voice of a singer involves a collection of complex features.

One of the early attempts towards singer identification was presented by Berenzweig et al. [14, 15]. In [14], speech recognizer is used to detect the vocal segments. Subsequently it is shown [15] that the use of vocal segments instead of whole music improves the identification accuracy. Kim et al. [87] presented a singer identification system based on voice coding features. Vocal regions are first detected using a bandpass filter and subsequently thresholding is applied based on a harmonics measure. Variants of linear predictive coding (LPC) is used to derive the features from vocal segments. Gaussian mixture model (GMM) and support vector machine (SVM) are used for identification. Zhang [4] also worked with the vocal part. Mel-frequency cepstral coefficients (MFCC) and LPC based features are computed from the vocal segments. Feature vector sequence are used to train GMM for each singer. One major limitation of the work is that the segmentation of vocal part is done manually.

Khine et al. [88] used hidden Markov model (HMM) at the first stage to detect vocal and non vocal segments. Sub-band based log frequency power spectrum is considered as the feature for this purpose. Identified vocal segments are further verified using hypothesis test. Then, song segments are represented by perceptual features like harmonic, vibrato and timbre. For singer identification, the approach of verification is deployed and here also hypothesis test is applied to verify the input feature vectors with the singer models.

Shen et al. [1] proposed hybrid singer identification (HSI) model. Music clips are classified as vocal and non vocal part using SVM. For this features like MFCC, energy, zero crossing rate (ZCR), spectral features like spectral centroid and spectral flux are used as input vector to SVM. For singer identification features are computed from both vocal and non vocal part. To characterize the vocal segments, linear predictive cepstral coefficients (LPCC) based vocal timbre feature and pitch histogram are used. It is assumed that an artist performs with in a limited set of genres. Hence moments are computed from Daubechies wavelet coefficient histogram (DWCHs) to capture the genre information. It is further assumed that a singer sings with a more or less fixed set of instruments. Hence from non vocal segment, MFCC are computed to represent instrument information. Finally, GMM based profile is created for each singer. But the assumptions regarding genre and use of instruments are restrictive.

Cai et al. [16] have considered sparse representation based classification to detect the vocal segments. Different auditory features based on MFCC, linear predictive mel-frequency cepstral coefficients (LPMCC), Gammatone cepstral coefficient (GTCC) are extracted from the vocal segments. Finally GMM is used to model the singers. Lehner et al. [89] did not opt for singing voice separation.

In their work MFCC is used as descriptor and classification is made based on random forest classifier. Su and Yang [17] proposed a system based on the idea of bag-of-frames (BOF). First of all robust principal component analysis (RPCA) is deployed to extract the voice segments. From the extracted voice segments, the log-magnitude spectrograms are encoded by l1-regularized sparse coding to obtain BOF features. Finally, SVM is used for classification. Kroher et al. [18] presented a system based on both low level and high level descriptors. The system is focused towards Flamenco (traditional music of southern Spain) singing. Voiced sections are estimated from the polyphonic music following pitch saliency. Along with MFCC based timbre features, vibrato features are computed using a plug in. A transcription based high level features are also used to cope up with the improvisation. For singer identification SVM is used.

In their work, Hu and Liu [19] separated singing voice and musical accompaniment using computational auditory scene analysis (CASA) method. In a subsequent effort in [20], they have relied on spectrogram analysis to filter out instrumental accompaniment from a song. Spectrogram is decomposed into two separate matrices by employing NMPCF (non-negative matrix partial co-factorization). But, it requires prior knowledge regarding the spectrograms of pure singing voice and pure musical accompaniment. The resulting spectrogram of singing voice thus obtained still bears the impact of instruments. It is refined and reconstructed using pitch based harmonic mask estimation method. For singer identification, framewise gammatone frequency cepstral coefficients (GFCC) are computed.

Sarkar and Saha [50] have applied a pre-processing technique to extract the vocal component. Features based on the variation pattern of zero crossing rate and short term energy are used for singer based classification. Ratanpara and Patel [90] have worked with popular Indian video songs using the features like timbre, chromagram, loudness, MFCCs, LPCCs and Adaboost as classifier. Tsai et al. [91] have dealt with compressed (mp3) file. Such files are first decompressed and then MFCCs are extracted. GMM is used to determine the distribution of the coefficients. Maximum likelihood classification is used for singer identification.

Past study reflects that most of the works have followed a pre-processing steps to extract the segments where the vocal component dominates. With those segments, a variety of features and classification techniques are tried. It is observed that MFCC, LPC and their variants are widely used and to an extent satisfactory outcome is achieved. For classification, various learning techniques are used. Although certain amount of work have been reported, still it is an active area of research and scope of improvement is still there. In our effort, we have utilized the experience of reported efforts and have developed a simple but novel system.

3.3 Proposed Methodology

The proposed methodology consists of three major steps namely *extraction of vocal Component*, *feature extraction* and *singer based classification*. It is observed that depending on the situation a singer

may have different accompanying arrangement, and songs of different genres may be performed by a singer. Hence, unlike the work in [1], proposed methodology does not make any specific assumption regarding the use of a fixed set of instruments by a singer or a singer belongs to limited set of genres. In Section 3.2, it has been noted that in order to characterize a singer it is worth to focus on the segments of the music signal where voice dominates over accompanying instruments. Hence as the first step we extract the vocal component and minimize the effect of other musical accompaniment. Singer characteristics are obtained by computing features from the extracted vocal components and utilized in classifying the singer based classification of music data. The individual steps are elaborated in the following subsections.

3.3.1 Extraction of Vocal Component

A song is the composition of singing voice and instrumental music. As per composition some segments may contain voice with or without accompanying background music and some segments may have only the background music. We refer such segments as vocal and non-vocal segments respectively. Extracting the segments which contain only the singing voice will be the best scenario for subsequent use in characterizing the singer's voice. But it is difficult to attain as the existence of only voice is quite rare. Mostly there exists segments with singing voice along with background music and segments with only the music. Figure 3.1 shows a sample song clip where the deep blue colored segments are vocal (contains voice) parts and the light blue colored segments are non-vocal parts without any singing voice.

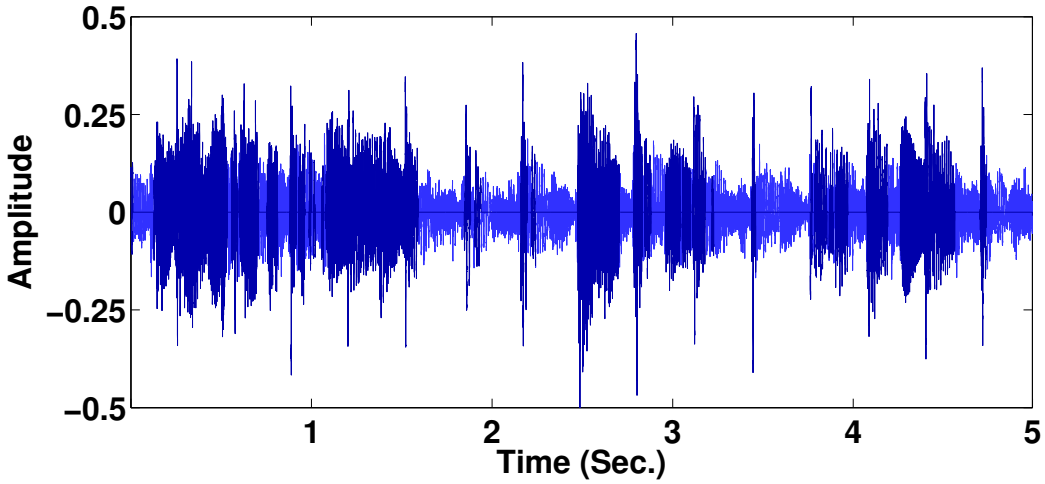


Fig. 3.1 A sample song clip showing vocal (in deep blue) and non-vocal (in light blue) segments.

Few works have been reported which are focused towards the separation of vocal component from a polyphonic music signal. Kopparapu et al. [92] have proposed a method of separation but it is limited

to Indian classical music. Cho et al. [93] proposed a method based on β-order minimum mean-square error spectral amplitude estimation (bSA) algorithm along with iterative back-fitting scheme. Such a complex method may not be essential for the specific application like singer identification where a gross segmentation may serve the purpose. As discussed in Section 3.2 different systems have followed their own scheme to satisfy the need. Number of systems [1, 17, 88] have adopted low level feature based classification of vocal and non-vocal segments using the classifiers like SVM, HMM or RPCA. Tuning of parameters for such approaches is non trivial. Spectrogram decomposition is also used for the same purpose [20]. But it requires prior knowledge about the vocal and non vocal spectrograms.

Proposed way of extracting the vocal dominating component is quite simple. Our goal is to remove the purely musical segments at first step and then to minimize the impact of the music from the remaining segments with voice. The steps are outlined in the following subsections.

Algorithm 1 Removal of non-vocal segments

% Removes the segments which do not contain singing voice. The segments with singing voice (with or without accompanied by background music) is the output %

procedure VOCALSEGMENTS(x) ▷ x is a song clip
  $V_{segment} \leftarrow []$
  $N \leftarrow size(x)$
  $W_L \leftarrow 256$ ▷ window length
  $n_f \leftarrow \frac{N}{W_L}$ ▷ number of frames
  for $i \leftarrow 1$ **to** n_f **do**
    $frame_i \leftarrow Frame(x, W_L, i)$ ▷ selects i^{th} frame
    $\mu_{e_i} \leftarrow \sum_{j=1}^{W_L} |frame_i(j)|$ ▷ energy of i^{th} frame; $frame_i(j)$ is the j^{th} sample of i^{th} frame
  end for
  $\mu_E \leftarrow \frac{1}{n_f} \sum_{i=1}^{n_f} \mu_{e_i}$ ▷ average frame level energy of the clip
  $\sigma \leftarrow \sqrt{\frac{1}{n_f - 1} \sum_{i=1}^{n_f} \left(\mu_{e_i} - \mu_E\right)^2}$ ▷ standard deviation of frame level energy of the clip
  for $i \leftarrow 1$ **to** n_f **do**
    if $\mu_{e_i} > \mu_E - 0.25 \times \sigma$ **then**
      $V_{segment} \leftarrow \{V_{segment}, x_i(j)\}$ ▷ appends vocal segment
    end if
  end for
  return $V_{segment}$
end procedure

Removal of non-vocal segments

It is observed that a vocal segment has more energy compared to a non-vocal segment as singing voice always emphasized over accompanying instrumentals in a song. The same is also reflected in

Figure 3.1. Thus, the non-vocal segments can be removed based on the energy distribution following Algorithm 1. The song clip is divided into number of frames. Each frame contains W_L samples. In our work, it is taken as 256. For each frame, its energy (μ_{e_i}) is computed. It may be noted that without losing the generality, sum of the absolute magnitude of the amplitude is used instead of their squared sum in approximating the equation. Mean (μ_E) and standard deviation (σ) of the frame level energy of the clip are computed. Frames with energy more than $\mu_E - k * \sigma$ are considered as vocal frame and samples of such frames are append in $V_{segment}$. In our work, k is empirically set to 0.25. Thus the non-vocal segments are removed and $v_{segment}$ is processed further to minimize the impact of background music, if any. Figure 3.2(a) shows the output after the removal of non-vocal segments corresponding to the signal shown in Figure 3.1.

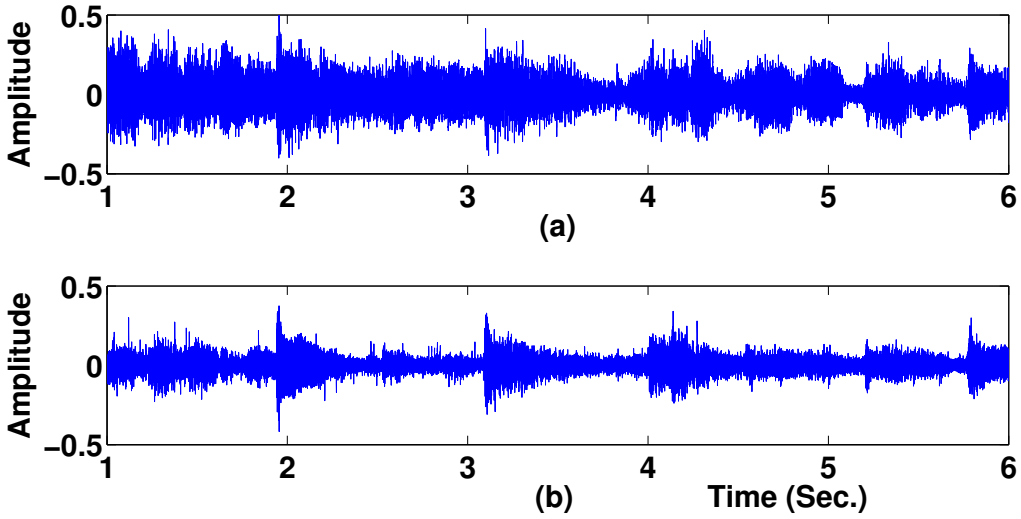


Fig. 3.2 Extraction of vocal components:(a) Output after the removal of non-vocal segments and (b) output after minimizing the impact of musical accompaniment.

Diminishing the impact of musical accompaniment

The vocal segments extracted so far also bear the impact of the accompanying instrument. At this stage we try to curtail their impact to make the vocal component becomes more dominant. It is noted that the fundamental frequency of speech normally belongs to the range of 100Hz to 300Hz. But for singing voice, the range is more wide, depending on the voice profile and style adhered by the singers. The wavelength of singing voice is less compared to wavelength of musical instruments. This is very useful property for distinguishing sound of singing voice and sound produced by musical instrument [94]. Despite of all these, it is possible to have the partial overlap between frequency range of the singing voice and that of various instruments used. Our goal is not to extract precisely the signal components corresponding to the singer voice. Rather, it is intended to formulate a simple

mechanism for reducing the impact of background music as far as possible. Rhythmic instruments like drums are most frequent and give rise to low frequency components. Other instruments may operate at frequencies higher than that of the singing voice. Based on these observations we apply a bandpass filter to retain the frequency components only within the range of 150Hz to 1500Hz. It preserves the voice part and removes the contribution of the instruments significantly. The steps are very simple as shown in Algorithm 2. $V_{segment}$, the output of the first step containing the vocal segments is the input. On each frame of this signal, bandpass filter is applied and finally the refined signal append in V_{refine}. It is used for extracting the features. Figure 3.2(b) shows the output corresponding to the signal in Figure 3.2(a).

Algorithm 2 Reducing the impact of musical accompaniments

% Reduces the impact of accompanying instruments from the vocal segments. Refined signal is the output. %

procedure VOCALREFINE($V_{segment}$) ▷ $V_{segment}$ contains vocal segments obtained as the output of Algorithm 1
  $V_{refine} \leftarrow []$
  $N \leftarrow size(V_{segment})$
  $W_L \leftarrow 256$ ▷ window length
  $n_f \leftarrow \frac{N}{W_L}$ ▷ number of frames
  $bpf \leftarrow [150Hz \leftrightarrow 1500Hz]$ ▷ bandpass filter for frequency band 150Hz to 1500Hz
  for $i \leftarrow 1$ **to** n_f **do**
    $frame_i \leftarrow Frame(V_{segment}, W_L, i)$ ▷ selects i^{th} frame
    $V_{spec_i} \leftarrow fft(frame_i) * bpf$
    $V_{refine} \leftarrow \{V_{refine}, ifft(V_{spec_i})\}$ ▷ appends refined segment
  end for
  return V_{refine}
end procedure

3.3.2 Feature Extraction

Song data is very complex in nature. Different physiological aspects are elaborated in the work of Kob et al. [95]. Singing voice generation system consists of number of subsystems: sub-glottal and glottal part (lungs, bronchi, trachea, muscles), vocal folds within larynx and the vocal tract (upper part of larynx, pharynx, oral and nasal cavities). Lungs are the pumping unit which delivers the airflow and air pressure to vibrate the vocal fold. Vocal folds convert the airflow into acoustic waves. This source spectrum is usually of monotonically decreasing nature with frequency. Vocal tract filters the acoustic waves emanating from the vocal fold to produce a wide range of sound. Resonances of the tract enriches the higher frequency components. To summarize the process, we hear the sound radiated from the mouth and the spectral envelope of this final signal is the outcome of the interactions

between the spectrum of voice source at larynx, the gain at vocal tract and the impedance at the lip and mouth.

There are several properties like loudness, pitch, voice range profile, vibrato, falsetto that can be used to characterize the singing voice [95]. Variation in the sub-glottal pressure and the muscular adjustment of the glottis result into the variation in loudness [96]. It is reflected as the variation of amplitude in the source spectrum. Perceived pitch is mostly determined by the fundamental frequency of the vibration of vocal folds. Furthermore, different laryngeal mechanisms adds a wide range of frequencies. The frequency range and corresponding sound pressure level depict the voice range profile [95]. Pitch is an intrinsic property of the voice. But one can enhance the voice range profile by proper voice training. Singers may adopt two characteristics namely vibrato and falsetto [96], [97], [98]. Vibrato is used to reflect the emotional aspect of the song more accurately. It is the modulation of fundamental frequency accompanied by amplitude variation. Falsetto is a special way of singing where singer sings one octave higher than his/her normal range to imitate a different tone. It reflects less dynamic variation in comparison to that in the normal mode of the singer. It is understood that the characteristics of the singing voice are caused by the physiological structure of the voice generation system of the singer and the training exercised by the singer to control the mechanism.

The task of singer identification can be accomplished by modeling either physical singing voice generation process or musical signal that we perceive. In our work, we try to model the signal by extracting suitable features. Studying the voice generation system and the voice characterising properties, we have inferred that the distribution of energy over the frequency range can well summarize the singing voice. Furthermore, timbral features (capable of discriminating the singers even if they are similar in terms of pitch and energy) can also be captured based on the same. Features are designed keeping all these observations in mind.

In our work, features are extracted from the refined vocal component obtained after the pre-processing task. Features are considered from two different perspective. To capture the perceptual aspect at the listener end, MFCC (mel frequency cepstral coefficients) based features are utilized. On the other hand spectrogram based features are designed that will act as the vocal-print of the singing voice along with the underlying physical process.

MFCC based feature

The mel frequency cepstral coefficients (MFCCs) is considered as listener end feature as it takes the functionality of cochlea in human auditory system into consideration. The mel scale is related to perceived frequency of a pure tone to its actual measured frequency. Human ear can detect small changes at low frequencies very efficiently. But can not detect small changes at high frequency. Human cochlea vibrates at different locations depending on the frequencies of the audio signal the ear receives. Accordingly different nerves of the brain are fired to provide the perception of the frequency. In audio signal processing, the frequency perception technique of human ear is performed by mel filterbank. The shape of the filterbanks is triangular. The initial filters are very narrow as the human

ear can sense the small differences. Higher the frequencies, corresponding Mel filters get wider, to become less concerned about small variations. In short, MFCC is a compact description of the shape of the spectral envelope of an audio signal from perceptual perspective.

Algorithm 3 Spectrogram based vocal-print

procedure VOCALFEATUE(V_{refine})▷ V_{refine} contains refined vocal segments obtained as the output of Algorithm 2. $Vocal_{print}$ is the output providing the feature vector.
  $Vocal_{print} \leftarrow []$
  $Spec \leftarrow []$ ▷ to store the components of filtered spectrum
  $N \leftarrow size(V_{refine})$
  $W_L \leftarrow 256$ ▷ window length
  $n_f \leftarrow \frac{N}{W_L}$ ▷ number of frames
  for $i \leftarrow 1$ **to** n_f **do**
    $frame_i \leftarrow Frame(V_{refine}, W_L, i)$ ▷ selects i^{th} frame
    $V_{spec} \leftarrow fft(frame_i)$
    $\mu_{spec\_90} \leftarrow 0.90 \times \sum V_{spec}$ ▷ 90% of total power
    while $\mu_{spec\_90} > 0$ **do**
      $[max_{ampl}, f_{maxAmpl}] \leftarrow getMax(V_{spec})$ ▷ max_{ampl} and $f_{maxAmpl}$ store the amplitude and frequency of strongest component in V_{spec} respectively
      $Spec(f_{maxAmpl}, i) \leftarrow max_{ampl}$
      $V_{spec}(f_{maxAmpl}) \leftarrow 0$
      $\mu_{spec\_90} \leftarrow \mu_{spec\_90} - max_{ampl}$
    end while
  end for
  $k \leftarrow 1$
  for $sf \leftarrow 150$ **to** 1500 ; $sf = sf + 10$ **do** ▷ Feature vector generated by taking frequency band wise mean.
    $sum \leftarrow \frac{1}{n_f \times 10} \sum_{frq=sf}^{sf+9} \sum_{frm=1}^{n_f} Spec(frq, frm)$
    $Vocal_{print}(k) \leftarrow sum$
    $k \leftarrow k + 1$
  end for
  return $Vocal_{print}$
end procedure

The steps for computing MFCC are elaborated in [67]. First of all the signal is divided into frames. Corresponding to each frame, log of amplitude spectrum is computed. The spectrum is then transformed into mel scale. Mel frequency $m(f)$ corresponding to the signal frequency f is computed as:

$$m(f) = 1125 * log_e(1 + \frac{f}{700}) \tag{3.1}$$

It may be noted that there is a nonlinear relationship between the actual frequency scale and the mel scale to incorporate the perception model. Finally, discrete cosine transform (DCT) is applied on the mel spectrum to obtain the coefficients. In our work, frame size is taken as 256. First thirteen coefficients of all the frames are considered. The frame level coefficients may be concatenated to represent the signal characteristics in detail. But the dimension becomes prohibitive. On the other hand, the common practice of considering the average value for each coefficients over the frames is too general. Hence we have followed an intermediate approach. Consecutive frames form a group and average of the coefficients are taken at group level. In our experiment, eight group of frames are considered and 104 dimensional vector is obtained.

Spectrogram based vocal-print

Timbre is an important perceptual aspect that has immense potential in discriminating the singers. Even if the singers possess similarity in basic characteristics like pitch, energy they differ in terms of timbre. It has motivated us to explore on this. Timbre includes the spectra and the spectral envelope of the audio signal. In this context, spectrogram can be thought of as a representation capturing both the aspects in a compact form. It visually depicts the distribution of energy at different frequency level over time. Thus, it provides the detailed frequency profile of a singer.

From the spectrogram we have extracted the vocal-print of a singer. To construct the spectrogram pre-processed vocal component of the signal is broken into frames consisting of 256 samples. FFT is applied on each frame to obtain the spectrum at a point of time. The same is carried on the subsequent frames to obtain the distribution over the time scale. Being a detailed description, it not only shows the dominating frequencies but also their variations in terms of strength over time. Dominating frequency always may not be the fundamental frequency. It is the frequency that is most heard. It can be stated that a picture of voice range profile is obtained. Along with the fundamental characteristics of the singing voice, impression of vibrato and falsetto applied by a singer is also present in a spectrogram.

In order to generate the vocal-print, we have applied an energy based filtering on each frame level spectrum. Stronger components in the spectrum constituting the 90% of the total power in the frame are only retained. It helps in minimizing the effect of falsetto. It is intended as falsetto is an artifacts and not the natural quality of a singer. By training also one can adopt the similar falsetto. Moreover, voice components are supposed to be of higher energy. Thus, the filtering is likely to reduce further the effect of accompanying instruments present in the pre-processed signal. So, the resulting spectrogram mostly corresponds to the vocal components of the singer of a song. In order to form the feature vector, frequency scale (150 Hz to 1500 Hz)is divided into number of bands of width 10Hz. Average spectral power in the bands are concatenated to form the 135-dimensional feature vector representing the vocal-print. Because of this band formation and averaging, the effect of vibrato is minimized and the dimension of feature vector is reduced. Width of the band is intentionally kept small. So that, wide apart frequency components do not get merged. The implementation details of the computation of vocal-print is presented in Algorithm 3.

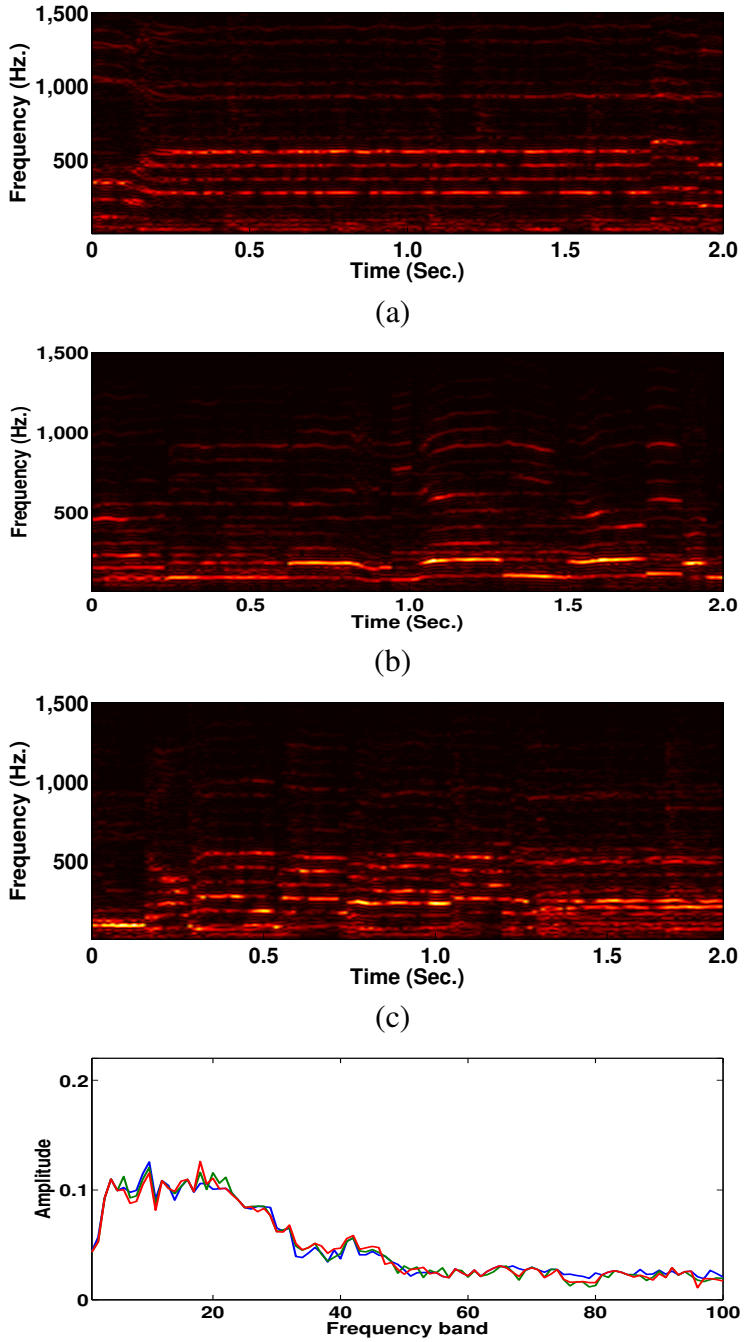
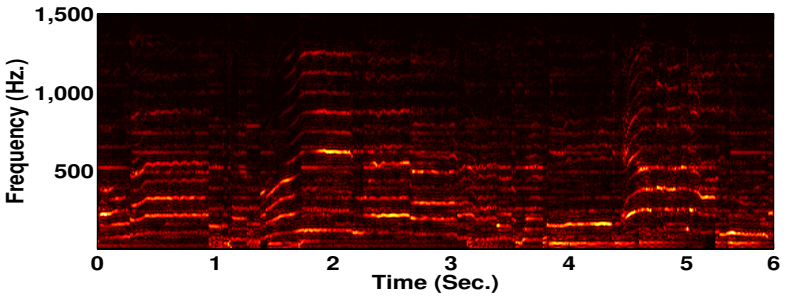
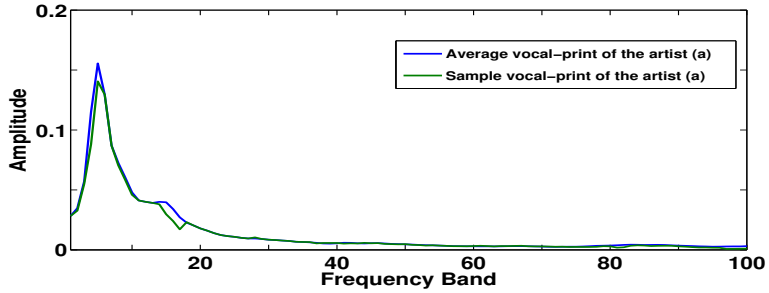


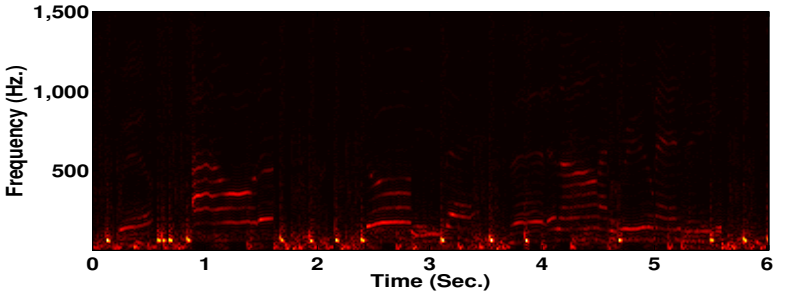
Fig. 3.3 Vocal-print of a singer. Filtered spectrograms of three different songs of same singer from *artist20* database are shown in (a), (b) and (c). Vocal-prints for (a) in blue, (b) in green and (c) in red respectively are plotted in (d).



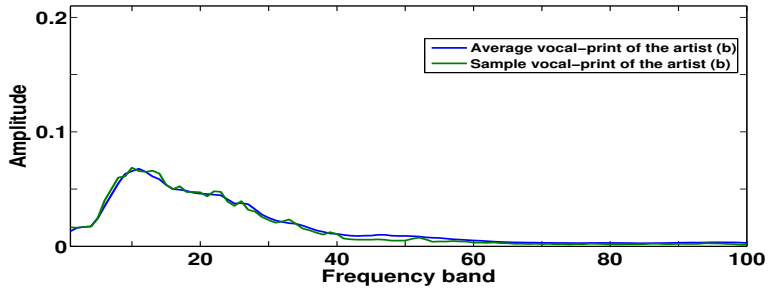
(a)



(b)



(c)



(a)

Fig. 3.4 Vocal-print of two different singers from *artist*20 dataset. Filtered spectrogram for two different singers are shown in (a) and (c). Feature vector (in green) corresponding to the spectrograms and average vocal-print (in blue) of the two singers are shown in (b) and (d).

Figure 3.3 shows the filtered spectrograms of three different songs of same singer from *artist20* database [45]. Corresponding feature vectors are also shown in different colours. Similarity of the pattern in the feature vectors is noticeable. On the other hand, Figure 3.4 shows the filtered spectrograms of two different singers from *artist20* database. Average vectors for the singers are obtained by taking average of the individual feature vectors of all the songs of the particular singer. Individual feature vectors corresponding to the spectrograms and the average feature vectors are shown. It is found that the individual vector for both the singers are closely following the corresponding average vectors and the vectors of different singers are considerably different. Thus the proposed way of vocal-print representation maintains the similarity for the songs of same singer and also capable of discriminating different singers.

3.3.3 Classification Technique

In this work, a multi layer perceptron (MLP) based neural network with back propagation is used for classification. It is a supervised classifier. The first layer of a neural network is the input layer. Number of nodes in first layer is equal to the dimension of the feature vector. Last layer is called output layer. Number of nodes in output layer is equal to the number of output classes for the specified inputs. A neural network may have one or more hidden layers. Layers are made up of a number of interconnected nodes which contain an activation function. Feature vectors are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to the output layer, where the output denoting the class is generated. MLP [75] and back propagation [76] algorithms are effective tool for pattern recognition and classification.

For a neural network, we need to determine the number of hidden layers and number of nodes in each hidden layer. In our work, we have considered two hidden layers with number of nodes hid_1 and hid_2 for the first and second hidden layer respectively. The values are estimated as follows [77].

$$hid_1 = N_{out}\left(\frac{N_{in}}{N_{out}}\right)^{\frac{2}{3}}; \tag{3.2}$$

$$hid_2 = N_{out}\left(\frac{N_{in}}{N_{out}}\right)^{\frac{1}{3}} \tag{3.3}$$

Where N_{in} denotes number of nodes in the input layer *i.e.* dimension of feature vector, and N_{out} denotes number of nodes in the output layer *i.e.* the number of output class labels.

3.4 Experiment and Discussion

We have worked with two datasets namely *artist20* [45] and *mir-1k* [46]. The *artist20* dataset is a subset of *uspop2002* dataset. It includes the songs of twenty artists. There are six albums per artist,

Table 3.1 Classification accuracy of the proposed system on *artist20* dataset.

Feature	**Accuracy (in %)**
MFCC based feature	50.25
Spectrogram based vocal-print	53.50
Spectrogram based vocal-print and MFCC based feature	75.50

Table 3.2 Classification accuracy of the proposed system on *mir-1k* dataset.

Feature	**Accuracy (in %)**
MFCC based feature	85.00
Spectrogram based vocal-print	88.50
Spectrogram based vocal-print and MFCC based feature	97.50

and 1413 songs in total. The songs are sampled at 16 KHz. For each song, clips are of 30 seconds duration. *mir-1k* dataset contains 1000 song clips from 110 *karaoke* (Chinese pop) songs. The length of the clips ranges from 4 to 13 seconds. The songs sung by nineteen singers among them eight are female and eleven are male.

Neural network is used as the classifier. Performance of the proposed methodology is measured for each dataset. The experiment is carried out separately for MFCC based features and spectrogram based vocal-print. Finally, it is repeated for both the features combined together. In each case, network is trained with 70% data, tested on 20% data and 10% data is used for validation. It is done iteratively so that each of the song in the dataset becomes part of test data.

Table 3.1 and 3.2 show the classification accuracy for *artist20* and *mir-1k* dataset respectively. It is observed that the accuracy is better for spectrogram based vocal-print on both of the datasets. Moreover, performance drastically improves when they are combined. Thus, the effectiveness of proposed vocal-print is well established. The idea of combining the perceptual aspect in the form of MFCC based features is also well justified. The confusion matrices for the proposed methodology

Table 3.3 Comparison of performance on *artist*20 dataset.

Methodology	**Feature**	**Classifier**	**Accuracy (in %)**
Langlois et al. [99]	MFCC based features	HMM	59.14
Su and Yang [17]	codeword based bag of features (BOF)	SVM	66.00
Shahreza et al. [100]	MFCC based features	GMM	71.5
Proposed system	Spectrogram based vocal-print and MFCC based features	NN	75.50

with combined features are shown in Table 3.4 and 3.5. Several researches [17, 99, 100] have worked on *artist20* datasets and reported their results. The features and classification techniques used by them are shown in Table 3.3. It is also clear that proposed methodology outperforms them.

3.5 Summary

In this work we have presented a novel methodology for singer based classification of song data. First of all a simple pre-processing is applied on the music clips to extract the vocal segments and also to minimize the effect of background music on those. Features are computed from such segments to capture the characteristics of the singing voice. Spectrogram based vocal-print can well represent the voice profile. It takes care of voice production system and timbral aspects. MFCC based features are also combined to reflect the perceptual aspect. Multi-layer perceptron based neural network with back propagation is used as the classifier. Experiments with number of databases reflect the usefulness of the vocal-print and also justifies the combining of two types of features. Comparison of performance with number of systems establishes the effectiveness of the proposed methodology.

Table 3.4 Confusion matrix for proposed system on *artist*20 dataset.

Actual \ Classified	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	44	4	0	1	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0
2	0	62	2	0	1	1	1	5	0	4	0	1	0	2	0	2	2	0	1	0
3	0	1	44	2	0	2	0	4	0	0	0	0	0	0	0	1	1	0	0	0
4	0	1	1	48	4	2	0	2	0	1	1	0	0	1	7	0	0	1	0	4
5	1	0	0	3	56	3	0	0	0	1	1	0	1	0	1	0	1	2	0	2
6	1	0	1	3	3	50	2	0	1	0	0	1	0	3	0	1	0	0	1	4
7	0	0	0	0	0	1	56	0	0	0	3	0	0	0	0	9	0	10	0	0
8	0	0	0	0	0	0	0	64	1	1	0	0	0	0	0	0	0	0	0	0
9	1	3	0	0	0	0	0	0	79	0	0	1	0	0	3	0	0	0	0	0
10	0	0	10	0	0	0	0	0	0	30	0	1	0	0	0	0	0	0	0	8
11	0	0	0	0	0	1	2	0	0	0	46	0	5	0	1	7	0	0	8	0
12	2	0	0	0	1	2	0	0	0	0	0	59	0	0	0	0	0	0	1	0
13	1	0	0	0	2	0	4	0	1	0	3	0	46	0	0	7	0	0	3	0
14	5	5	2	3	2	0	2	0	0	0	0	2	0	53	0	0	1	0	0	0
15	0	0	0	8	0	0	5	0	7	0	0	0	0	2	44	0	0	0	3	4
16	0	0	0	0	0	0	7	0	0	0	3	0	5	0	0	67	0	6	0	0
17	0	0	0	0	0	1	2	0	0	0	0	0	1	0	0	0	48	0	0	0
18	0	1	0	0	2	0	3	0	1	0	4	0	1	0	1	3	1	50	2	0
19	0	1	1	1	0	0	5	0	0	0	6	0	2	0	0	3	0	2	62	0
20	0	2	0	0	0	0	0	0	0	2	3	0	0	3	9	0	0	0	0	45

Table 3.5 Confusion matrix for proposed system on $mir-1k$ dataset.

Actual \ Classified	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
1	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	48	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
3	0	0	38	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	1
4	0	0	0	130	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
5	0	2	0	1	37	0	0	1	0	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0
10	0	0	0	2	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	2	0	0	49	0	1	0	0	0	0
14	0	0	0	0	0	0	1	0	0	0	0	0	0	43	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	96	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65

Chapter 4

Emotion Based Classification

4.1 Introduction

Every piece of music is associated with an emotion and accordingly it generates an intuitive feeling to the listener. Identification of inherent emotion present in a music is an active area of research [21, 22, 101]. Despite the use of sophisticated techniques, identification of emotional category of musical excerpts is quite challenging. This is mainly due to the subjectiveness of emotion. The perception of emotion may vary from person to person. Moreover, the conveyed emotion depends not only on the structural features of music but also on the state (gender, age, personality etc.) and contextual aspects (like occasion and place) of the listener. This makes the task of emotion based classification of music further difficult.

Automatic classification of music according to their emotional content is an imperative task. More so, with the rapid growth in the size of digital music libraries. A good music emotion recognition (MER) system helps to group the songs according to their emotion. Such categorization can act as a fundamental step for developing a music recommendation system that enables user to retrieve the music according to his/her choice.

One basic approach for classification is to compute the descriptors from the music signals and then feed them to certain classifier [102–104]. But, success is limited for such systems as it is difficult to represent emotion by means of low level of features. In this context, deep learning has drawn attention. It has already achieved significant outcome in various tasks of computer vision [105–109] and natural language processing [110]. In recent times deep learning approaches are being tried for speech emotion recognition [32, 111–113]. A very few attempts [27, 34] are reported for music emotion recognition.

In this work, a convolutional deep learning network is proposed that helps us to extract the meaningful features. Moreover, the burden of designing the low level descriptors is removed. Performance of the proposed system is evaluated on two popular music emotion datasets. With this brief introduction, rest of the chapter is organized as follows. Survey of past work is presented in

Section 4.2. Section 4.3 elaborates the proposed methodology. Experimental results and summaries are put in Section 4.4 and 4.5 respectively.

4.2 Past Work

Music emotion recognition (MER) has drawn the attention of the researchers over a decade. Still it remains as an active area of research [27, 28, 34]. It is observed that two major steps are involved in the process: designing the suitable features to describe the music signal and thereafter identifying the emotion. Features may be conventional hand crafted ones as considered by most of the works or learnt features which has become the trend with the advent of deep learning. Using the features regression based approach can be followed to map the music into emotion plane suggested by the model of Thayer [30] and Russell [31]. The alternative approach is to rely on the classifier. In this section, we present brief survey on the features and emotion identification approaches.

A wide variety of hand crafted features have been used by the researchers. The patterns inherent in a music signal provide the perception of emotion [114]. Features are used to summarize the patterns. Energy or the power of a music clip is frequently used [28, 102, 103, 115, 116] as it has very correlation with arousal [117]. A music clip with fast tempo is often correlated with positive valence and slow tempo is correlated with negative valence [117]. Hence, use of tempo is also very common [48, 116, 118–120]. Timbral features, captured in different forms are also utilized by the researchers. Such features include mel-frequency cepstral coefficients (MFCC) [104, 111, 121, 122], daubechies wavelets coefficient histogram (DWCH) [48, 119, 120]. Zero crossing rate (ZCR) [28, 121, 123] and pitch [28, 33, 104] are also useful. Variants of spectral features [28, 122, 123] like spectral rolloff, spectral flux as well as tonality [119, 118] are also considered in various works.

As it is not an easy task to design hand crafted features for a given goal, in recent time considerable efforts have been put to learn the features using deep network. Although most of the works are on speech emotion recognition [33, 111, 112, 124]. It is still worth to follow those to understand the applicability of deep learning in the context of audio signal. Few efforts [27, 34] are directed towards music also. Convolutional neural network (CNN) has been tried by number of researchers [27, 112, 113]. Most commonly, a CNN is fed spectrograms generated from audio signals. A series of convolution and pooling operation is performed on it to build the feature vector. Sometimes recurrent neural networks (RNN) with long short-term memory (LSTM) [34, 112] has been considered. For RNN, input is the raw audio signal and LSTM divides into number of frames.

To recognize the emotion, regression based approach has also been followed. Emotion in music can be represented as two orthogonal components- *Arousal* and *Valence*. *Arousal* of a music represents energy, activation or intensity whereas *valence* denotes how pleasant a music is. Several two-dimensional models have been proposed, of which Russell's [31] and Thayer's [30] model are widely used. Figure-4.1 is a simple representation of circumplex model proposed by Russell where X and Y axis denote *valence* and *arousal* respectively and it shows the position of different emotional classes

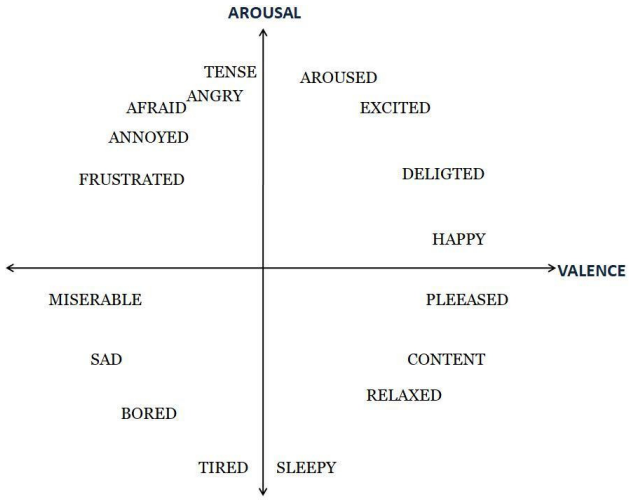


Fig. 4.1 Two dimensional emotion plane: Valence vs. Arousal.

in the two dimensional (2D) plane. In this approach, music clips in the training set are annotated with *valence* and *arousal* values and it is used to prepare the 2D emotion plane. Regression model is formed to predict *arousal* and *valence* by considering the low level features as the observed values. Separate regression models can be trained for *arousal* and *valence* [120]. The regressed value for both are used to find the position of the song in the 2D emotion plane describing the emotion. Researchers have worked with different regression models for predicting *valence* and *arousal* values. Yang et al. [119] experimented with three different regression algorithms namely, multiple linear regression (MLR), support vector regression (SVR) and AdaBoost.RT (BoostR) with a feature set consisting of spectral contrast, DWCH (daubechies wavelets coefficient histogram) and features obtained from PsySound [125] and MARSYAS [6]. SVR is also used by Han et al. [118]. They have used scale, average energy, harmonics and rhythm as musical features. Gaussian process regression (GPR) [123] is also applied on multiple sets of features extracted by MARSYAS. As the annotations are collected through surveys, the issue of inconsistency remains while training the models.

In classifier based approach music clips are first represented by a set of features. Thereafter, feature vector is fed as input to the classifier for emotion recognition. Commonly used classifiers include support vector machine (SVM) [102, 103, 121], artificial neural network (ANN), radial basis function ANN (RBF-ANN) [104], Gaussian mixture model (GMM) [115, 126], random forest [28] etc. Researchers have experimented with different parameter and kernel setups for the classifiers. In some cases, principal component analysis (PCA) [104] and linear discriminant analysis (LDA) have been used for reduction of feature dimension.

It is observed that variety of features and classifiers/regression models have been considered by the researchers. But success of all such systems are quite limited. Hence, emotion based categorization still remains an active area of research.

4.3 Proposed Methodology

In general, for classification problem, designing a uniform set of features that works across various datasets and classifiers is very critical. Emotion being very much subjective and psychological issue, it is further challenging. Limitations of hand picked low level features (mostly designed intuitively) affects the performance of a classifier. It has motivated us to apply deep learning network to design a set of features that will work more consistently for different datasets. The audio signal, may be pre-processed is fed to the deep network to learn the complex structural factors of music contributing to emotion.

Proposed methodology consists of three stages. At first, the audio signal is pre-processed to represent it into a concise but meaningful form which is fed to our convolutional neural network. A post processing is applied on the prediction output of the network. Pre-processing steps, proposed network architecture and the post-processing steps are elaborated in the following sections.

4.3.1 Pre-processing

The raw audio signal goes through a sequence of steps before being fed to the network. Each clip is normalized so that sample amplitudes are restricted within $[-1, 1]$. The music clip is divided into number of segments – each of 5 seconds duration [22, 127, 128]. These small segments are used as the unit to perceive the emotion. It makes the task more challenging. Consecutive segments have a considerable overlap with its neighbor that helps to increase the data volume also.

A two dimensional spectrogram [129] is computed for the segment and used as the input for the proposed network. Past study indicates that spectral features play important role in identifying the emotion. Spectrogram is our choice for input as it summarizes spectral information in a concise form. Moreover, convolutional neural networks (CNN) have shown promising performance on image data. Spectrogram being a pictorial representation, it can be utilized as the input for the networks similar to those used in image and vision problem.

To obtain the spectrogram, the audio segment is divided into number of frames with half overlap among the consecutive frames. In our work, a frame consists of 1024 samples. The spectrogram is obtained by taking short-time Fourier transform on the frames. Thus, it reflects time-frequency spectrum of the signal. The horizontal and vertical axes denote time (frame number) and frequency respectively. An element of the spectrogram shows the energy of a frequency component at an instance. Thus, energy variation of various frequency components over time is captured in the spectrogram. The frequency scale is converted from linear scale to mel-scale as it resembles human auditory system. To reduce the dimension, the mel-scale is divided into 128 bins. The logarithm of the values are considered to dampen the effect of large magnitude. Log values are scaled by using standardization procedure *i.e.* mean subtraction and division by the standard deviation. In our work, the spectrogram is formed using 196 frames. Thus, the dimension of spectrogram becomes 196×128 and it is fed to the network.

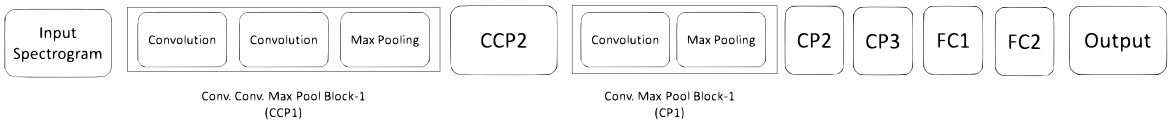


Fig. 4.2 Block diagram of the proposed convolutional neural network. CCP, CP and FC stand for *Convoluton-Convolution-Pooling*, *Convolution-Pooling* and *Fully Connected* respectively.

4.3.2 Proposed Network Architecture

Convolutional neural network (CNN) is biologically inspired architectures characterized by their local receptive structures, sparse connectivity and shared weights. It has been successfully applied in image processing [130] tasks and also in speech recognition [131]. Two dimensional convolution has been applied along dimensions of time and frequency on the input spectrogram. Every layer of convolution has a fixed number of filters which convolve with the inputs to the corresponding layer and produces feature maps. We denote the m-th feature map of the k-th layer as h_m^k. Corresponding input and bias for the k-th layer are x^k and b^k respectively. For the m-th feature map of k-th layer weight is W_m^k. Elements of h^k is obtained as follows:

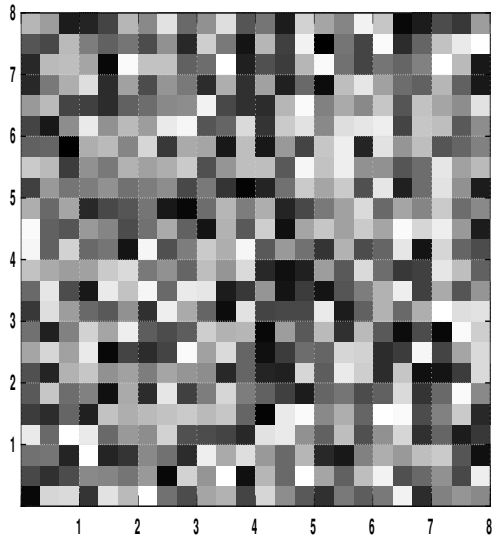
$$h_{ijm}^k = \sigma((W_{ijm}^k * x^k) + b^k) \tag{4.1}$$

where σ is some non-linearity function, $*$ denotes convolution operation and (i, j) denotes pixel location on surface image.

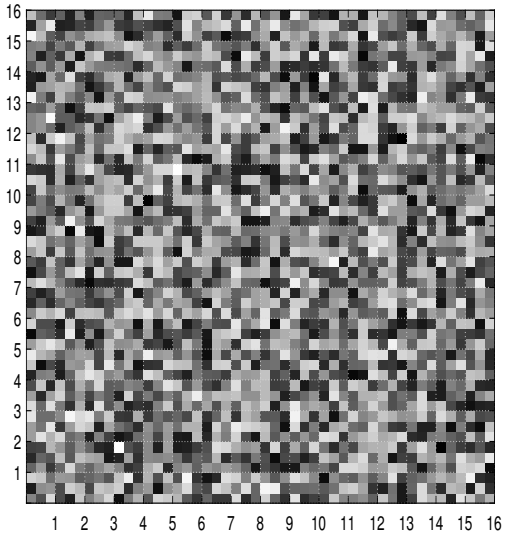
Unlike conventional ANNs, not all neurons in a layer are connected to all neurons in the next layer. The neurons in a layer respond to activation that falls within its own receptive area. As layers are stacked, the receptive areas of the neurons become increasingly global. This helps capture both short and long term dependencies which are extremely significant in case of audio. Again, in CNN, the filters are replicated across a layer enabling the sharing of parameters. This ensures that same features are detected regardless of their position contributing to translational invariance.

Convolution layers are typically followed by pooling layers where the convolved data is down sampled usually by considering the maximum or average values for every small subsection of the matrix. Pooling helps to reduce the number of parameters in the model, thereby reducing over fitting concerns. It also helps to achieve translational invariance.

The proposed architecture is built around VGGNet [130]. In the proposed model, we have considered fewer layers. It alleviates the problem of over-fitting on the small sized training datasets. Figure 4.2 shows the schematic diagram of the proposed network. The first two blocks of the network are referred as CCP blocks. One such block consists of two convolution layers followed by a max pooling layer. It is then followed by three blocks (referred as CP), each consisting of alternating layers of convolution and max pooling. Finally, there are three fully connected (FC) layers. The last one is with the same dimension as the number of output classes. The detailed architecture is given in Table 4.1.



(a)



(b)

Fig. 4.3 Visualization of filters. (a) first convolution layer (64 filters of kernel size 3X3). (b) seventh convolution layer (256 filters of kernel size 3X3).

Table 4.1 Architecture of the proposed convolutional neural network (CNN).

Data Shape	Layer Type	Description
$196 \times 128 \times 1$	Input	Log mel spectrogram
$196 \times 128 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$196 \times 128 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$98 \times 64 \times 64$	Max Pool	Kernel: 2×2 Stride:2×2
$98 \times 64 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$98 \times 64 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$49 \times 32 \times 64$	Max Pool	Kernel: 2×2 Stride: 2×2
$49 \times 32 \times 128$	Conv	Kernel: 3×3 Stride: 1×1
$16 \times 10 \times 128$	Max Pool	Kernel: 3×3 Stride: 3×3
$16 \times 10 \times 128$	Dropout	Keep prob. = 0.75
$16 \times 10 \times 256$	Conv	Kernel: 3×3 Stride: 1×1
$5 \times 3 \times 256$	Max Pool	Kernel: 3×3 Stride: 3×3
$5 \times 3 \times 256$	Dropout	Keep prob. = 0.75
$5 \times 3 \times 256$	Conv	Kernel: 3×3 Stride: 1×1
$1 \times 1 \times 256$	Max Pool	Kernel: 3×3 Stride: 3×3
$1 \times 1 \times 256$	Dropout	Keep prob. = 0.75
256	Fully Connected	Flattened to 1D tensor with 256 neurons
256	Dropout	Keep prob. = 0.5
256	Fully Connected	256 neurons
256	Dropout	Keep prob. = 0.5
4	Softmax	4 output classes

Convolution is performed along both time and frequency axes using small square filters of size 3×3. A fixed stride length of 1 is used for all the convolution layers. The number of filters is progressively increased in the later blocks of the network. We are motivated to use small kernel dimensions for convolution to reduce the number of trainable parameters. Instead of alternating convolution and max pooling layers, in the CCP blocks we have used two convolution layers one after the other. This is to realize larger sized filters at a lower cost. Max pooling layer is used to down sample the data. The kernel size for pooling is 2×2 in the CCP blocks and 3×3 for the CP layers. Once again the increase in size for the later blocks reduces the number of weights in the flattening layer. It is observed that further increase in the depth or width of the layers did not improve the performance of the proposed model. The output of the last pooling layer is flattened and then fed to the fully connected layer. L2 regularization is applied to the weights of the fully connected layers. The last three CP blocks and the FC layers are followed by a dropout layer to further prevent over-fitting on the training data. ReLu activation [132] is non-saturating and allows faster training. Hence it is used for convolution and first FC layer instead of sigmoid or tanh activation. The activation function

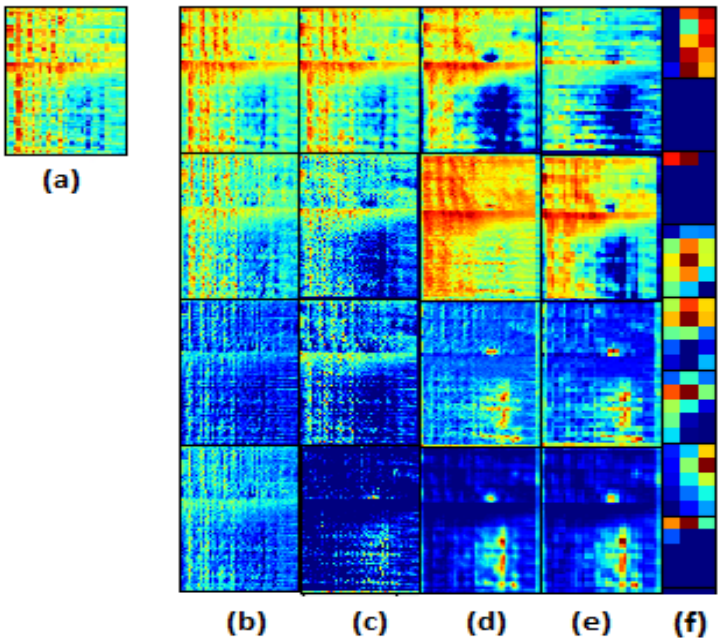


Fig. 4.4 An input spectrogram and output after different convolution layer: (a) input spectrogram and few filtered output after (b) first convolution layer, (c) second convolution layer, (d) fifth convolution layer, (e) sixth convolution layer, and (f) seventh convolution layer.

is defined as $ReLu(x) = max(0,x)$ where x is an input to a neuron. For the last FC layer, Softmax activation is used and it is represented as

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{4.2}$$

where z is a multidimensional vector having as many dimensions as the number of output classes.

Figure 4.3 shows the filters corresponding to the first and last convolution layer. There are 64 and 256 filters of size 3×3 in those layers respectively. Initially, from the input spectrogram the filters capture the variation of signal energy with frequency or time or both. It can be thought of as equivalent to determining the edges of different orientations in case of image. Dominating frequency components at different instances are highlighted by max pooling. At later stages, convolution leads to further abstract representation. Figure 4.4 shows one input spectrogram and output after various convolution layers. In the spectrograms blue denotes minimum energy and red corresponds to maximum. For each layer few sample spectrograms have been shown. It is observed that in the initial stages, local details are observed and gradually those are summarized. Figure 4.5 shows the output after last convolution layer corresponding to music clips of four different emotions. For better visualization, instead of 256 spectrograms of size 5×3, first 24 have been shown for each clip. It is also observed that they differ considerably for different emotions.

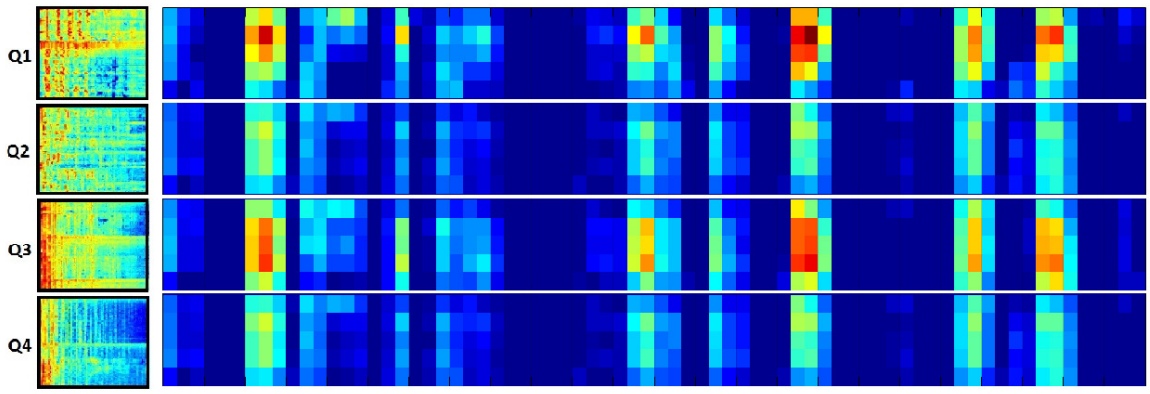


Fig. 4.5 Input spectrogram for four emotions and the output after seventh convolution layer where Q1, Q2, Q3 and Q4 denote happy, anger, sad and tender respectively.

4.3.3 Post-processing

The network estimates a class label for every segment of each test clip. Different segments may be identified as different emotional class. We used a combination of voting and run-length based technique to estimate the label for the whole clip. Suppose, number of segments in the clip be N_s and N_{c_i} be the number of segments predicted as i-th class. Then the voting strength for i-th class (VS_i) is computed as $\frac{N_{c_i}}{N_s}$. To measure the run-length strength, segments are arranged chronologically along time scale. Sequence of segments with same label forms a run. Run-length strength for i-th class (RLS_i) is defined as $\frac{L_i}{N_s}$. Where, $L_i = max\{RL_{ij}\}$. RL_{ij}s denote the length of the runs for i-th class. Finally, the score (S_i) that the clip belongs to i-th class is obtained as:

$$S_i = w_1 \times VS_i + w_2 \times RLS_i \tag{4.3}$$

where w_1 and w_2 are weights given to the strengths and $w_1 + w_2 = 1$. If $S_k = max\{S_i\}$ then the clip is labeled as class k.

Associating run-length information helps to capture the effect of persistence of an emotion. A listener is able to conceive emotions having prolonged spans (even if they are fewer) better than those having multiple short spans [128, 133]. Hence, w_2 should be more than w_1 to emphasize run-length. Over emphasis on w_2 can have detrimental effect in case the segments are categorized randomly giving rise to low run. We have experimented with a set of values ranging from 0.4 to 0.8. Best result was obtained for $w_2 = 0.7$. However, the performance was also very close for $w_2 = 0.6$.

4.4 Experimental Results

Experiments are performed on two different benchmark datasets following both the approaches namely *feature based* and *deep learning based approach*. The model in Figure 4.1 shows the position of

different emotions in a two dimensional plane. In our work, instead of dealing with so many classes, we have considered four broad categories. These are *happy*, *anger*, *sad* and *tender/neutral*. It may be noted that the four categories correspond to the first, second, third and fourth quadrant of the Russell's model.

4.4.1 Datasets

Soundtrack [47] and Bi-Modal [48] are the two datasets used in our work. In both the datasets, sampling rate for audio clips are 22.05 KHz. The details of the datasets are as follows.

Soundtracks: The dataset [47] consists of 360 audio-clips collected from background tracks of movies with duration around 30 seconds. Each clip is annotated with different emotion class like anger, sad, happy and tender. A clip can have multiple tags with confidence value. In our experiment label with maximum confidence is considered for matching. Number of audio clips in anger, happy, sad and tender emotion categories are 156, 58, 68 and 78 respectively.

Bi-Modal: This dataset [48] consists of 162 songs. Each song clip is of 30 seconds duration. Both, audio signal and lyrics (textual) data (hence, Bi-Modal) of the songs are available. In our work, we have considered the audio signal part and ignored the lyrics. The clips are annotated with the four quadrants as shown in Figure 4.1. Number of songs in quadrant 1, 2, 3 and 4 are 52, 45, 31 and 34 respectively. It may be noted that the quadrants correspond to *happy*, *anger*, *sad* and *tender* respectively.

4.4.2 Hand crafted Feature Based Approach

We have first worked with hand crafted features to study their performance. In order to compute the low level features, audio signal is divided into number of frames each consisting of n (taken as 512 in our work) samples and there is half overlap between the successive frames. Various time domain features like *short term energy (STE)* and *zero crossing rate (ZCR)* [71] which can reflect the arousal and frequency content respectively. Spectral features [71] like *spectral flatness*, *spectral crest factor*, *spectral centroid*, *spectral rolloff* and *spectral flux* have been considered. Thirteen *linear prediction cepstral coefficients (LPCC)* [134] have been included in the feature set to represent the production model of the vocal tract.First thirteen *mel frequency cepstral coefficients (MFCC)* [67] have been considered to take care of hearing perception of the listener.

All the features (*time domain, spectral, LPCC and MFCC*) are computed over the frames. Finally, mean and standard deviation of the individual features over all the frames in the music clip are taken as the clip level features. Thus, 66-dimensional feature vector is formed to represent a clip. All the features are extracted from the audio signal using the toolbox MARSYAS [6].

Music clips are represented by the extracted feature set and then supervised approach is followed for classification. The classifier is trained with a training dataset and thereafter the trained model is used for test data. We have used three different types of classifiers in this regard – a large-

Table 4.2 Classification performance for different combinations of hand crafted feature sets and classifiers.

	Bimodal (F1-score in %)			SoundTracks (Acc. in %)		
Features	SVM	RF	NN	SVM	RF	NN
A	44.78	42.02	46.48	43.61	41.46	41.51
B	47.12	50.77	54.66	50.00	48.06	52.68
A + B	47.71	50.88	56.62	51.38	48.27	51.80
A + B + C	53.23	52.94	62.74	53.61	49.10	54.31
A + B + C + D	54.26	52.54	63.45	53.77	49.71	55.41

Feature Sets: A = time domain features; B = spectral features; C = MFCC; D = LPCC

margin classifier (support vector machine(SVM) [135]), a decision tree based classifier (random forest(RF) [136]) and a perceptron based classifier (neural network(NN) [137]). All are implemented using Scikit-learn library [138].

We performed experiments with different combination of time-domain, spectral features, MFCC and LPCC based features. Random Forest is trained with a total number of ten trees in the forest and for splitting a node Gini-index [139] is used as impurity measure. In SVM, radial basis function (RBF) kernel is used and the value of regularization parameter is taken as one. For Neural networks, LBFGS weight-optimization method along with adaptive learning rate and Sigmoid activation function is used. For every experiment (combination of dataset, feature set and classifier), five fold cross validation is applied and average accuracy is reported. Table 4.2 summarizes the result. It is very difficult to obtain a feature combination that works best across the classifier and dataset. However, in most of the cases when all the features are combined provides better result. In general, the success of various feature-classifier combination is quite limited. It may be noted that for BiModal and Soundtracks dataset, F1-score and classification accuracy are used as performance metric respectively as the same have been used by other researchers working with those datasets.

4.4.3 Deep Learning Based Approach

As discussed in Section 4.3, music clip is divided into number of segments of five seconds duration. A song/music has a dominating emotional category. But, it may not remain constant over the whole clip. To address the issue, segments are overlapped heavily (four seconds in our case). Thus, the presence of segments with dominating emotion will be emphasized in comparison to deviated ones. Moreover, the substantial overlap will increase the number of segments that helps the network and makes the post-processing meaningful even for clips of small duration. It may be noted that a clip (*i.e.* all the segments) as a whole is either used for training data or as test data.

The proposed network has a total of $1,203,140$ trainable parameters. Training data is split into mini batches with batch size of 64 and training is done by minimizing categorical cross entropy loss [140, 141] between predictions and targets. Adam optimizer [142] is used with a learning rate

of 0.001. Dropout technique is implemented where a random fraction of neurons are switched off to prevent overfitting of the training set. L2 regularization is also applied to the weights of the fully connected layers. Weight initialization is done using truncated normal initializer. All the codes for training the model were written in Python using Keras [143] library. Experiment is carried out on Nvidia Quadro M5000 GPU with 8 GB of memory.

Table 4.3 Precision, recall and f-1 score (in %) for *Soundtracks* dataset.

Class	**Precision**	**Recall**	**F-1 score**
quadrant 1	58.20	71.23	63.61
quadrant 2	54.37	50.68	51.46
quadrant 3	82.25	82.65	82.28
quadrant 4	60.02	42.91	49.32

Table 4.4 Precision, recall and f-1 score (in %) for *Bi-Modal* dataset.

Class	**Precision**	**Recall**	**F1-score**
quadrant 1	80.46	81.59	80.98
quadrant 2	92.07	74.04	81.79
quadrant 3	72.97	68.52	68.82
quadrant 4	74.20	86.85	79.70

Class wise precision, recall and f-1 score for soundtracks dataset are shown in Table 4.3. Table 4.4 shows the performance data for Bi-modal dataset. For both the datasets, it has been observed that significant confusion arises between sad and tender (quadrant 3 and 4 of 2-D Russell's plane [31]), happy and anger (quadrant 1 and 2) classes. This may be accredited to the fact that both sad and tender classes belong to the low arousal category of 2D Russell's plane. Happy and anger belong to the high arousal category. It indicates that proposed model is stronger in discriminating emotions based on arousal and relatively poor for valence.

Performance of the proposed deep learning based methodology is compared with the work of Saari et al. [116]. They have worked with Soundtracks dataset. A set of 66 frame level audio features (extracted with MIRtoolbox) has been considered. Wrapper-selection has been employed and best performance is achieved with 4 randomly selected features. As they have worked with four fold cross validation, for comparison we have also followed the same. Table 4.5 shows that proposed methodology provides better result. It may be noted that, hand crafted feature based experiment with all the features combined together and neural network as the classifier (as shown in Table 4.2) provides better accuracy. Proposed deep learning based approach improves the result further.

Malherio et al. [48] have worked with Bi-modal dataset. As we have ignored textual data, performance is compared with audio based work of Malherio et al. [48]. They have used loudness, pitch, timbral, rhythmic, spectral contrast, and daubechies wavelets coefficient histogram (DWCH) as

Table 4.5 Comparison of performance for *Soundtracks* dataset.

Methodology	**Accuracy (in %)**
k-NN BE of Saari et al. [116]	56.5 ± 2.8
SVM BE of Saari et al. [116]	54.3 ± 1.9
Proposed deep learning based approach	67.71 ± 3.63

Table 4.6 Comparison of performance for *Bi-Modal* dataset.

Methodology	**F1-score (in %)**
Malherio et al. [48] (using only audio features)	72.60
Proposed deep learning based approach	77.82 ± 4.06

acoustic domain features and SVM classifier. Table 4.6 shows the comparative results. As Malhero et al. [48] has followed ten fold cross validation, for comparison we have also followed the same. It may be noted that the performance of hand crafted feature based experiment of ours (as shown in Table 4.2) is inferior to the work of Malherio et al. [48]. But, deep learning based approach performs better.

By observing the experimental outcomes, it is well understood that designing the features to represent the emotion is quite difficult. More difficult is to obtain a consistent set of features that provides optimal result for any classifier and for different datasets. Classification accuracy is also limited for the conventional approach based on hand picked features and classifier. In this context, proposed convolutional neural network improves the performance substantially for both the datasets.

4.5 Summary

In this work, we have followed deep learning based approach for music emotion recognition and experiment is carried out on two benchmark dataset. Experiment has also been done with handcrafted features. Different time domain and spectral features are chosen based on the past efforts of the researchers. LPCC and MFCC based features are also included as those correspond to the aspects of vocal production and human perception respectively. Although the combined feature set provides a moderate result for both the datasets, but the performance varies for different classifier. To avoid the difficulty of designing proper features, deep learning based approach is considered. Proposed convolutional neural network (CNN) is the modified version of VGGNet with comparatively less number of layers. It works with the audio segment of very small duration, even of five seconds to recognize the emotion. Experimental result shows that proposed network improves the recognition accuracy considerably. In future, a combined CNN-LSTM network may be considered for music emotion recognition.

Chapter 5

Raga Based Classification

5.1 Introduction

Indian Classical Music is regarded as one of the most prestigious and the highest class of music. *Raga* is the basic melodic framework of such music and also the most frequently used metadata for its description. Manual detection of *raga* involves expertise of high degree and their availability is also an issue. Hence, an automated system for *raga* identification is of immense significance for classical music collection.

Hindustani classical music is mainly found in Northern part of India, Bangladesh and Pakistan. *Raga* and *tala* (rhythmic cycle) remain the central notion in both the systems. *Khayal* [144] and *Dhrupad* [145] are the two main forms of Hindustani classical music. *Dhrupad* is primarily of devotional type and performed by male singers. *Khyal* is a relatively newer Hindustani vocal music. It is very popular because of its romanticism and emotional influence. The vocal performance is accompanied by string instruments like *tanpura*, *veena*.

In this work, we deal with Hindusthani classical music. As the task of *raga* identification requires domain knowledge, fundamental concepts regarding *raga* and its properties are discussed in Section 5.2. Section 5.3 presents the review of past work. Proposed methodology is outlined in Section 5.4. Section 5.5 and 5.6 provide the experimental results and summary respectively.

5.2 Basic Properties of *Raga*

Raga is the discernible melodic form underlying all Hindusthani classical music. It acts as a communication medium for two or more music lover's mind. A composition attracts the listeners due to its emotional content. Experts define the term raga in various ways. *Bharat Muni* in his *Natya Shastra* used the term raga to indicate aesthetic enjoyment or pleasure. According to Pandit Jasraj, the meaning of raga is *love* [146]. Matanga Muni coined the first technical definition of raga as *"In the opinion of the wise, the particularity of notes and melodic movements, or that distinction of melodic*

Table 5.1 *Swaras* of Hindusthani music and Western chromatic scale.

Position	Swara	Symbol	Western Note
1	Shadja	*sa*	C
2	Rishabha (Komal)	*re*	C#
3	Rishabha (Suddha)	*Re*	D
4	Gandhara (Komal)	*ga*	D#
5	Gandhara (Suddha)	*Ga*	E
6	Madhyama (Suddha)	*ma*	F
7	Madhyama (Tivra)	*Ma*	F#
8	Panchama	*pa*	G
9	Dhaivata (Komal)	*dha*	G#
10	Dhaivata (Suddha)	*Dha*	A
11	Nishada (Komal)	*ni*	A#
12	Nishada (Suddha)	*Ni*	B

sound by which one is delighted, is Raga." [39]. As described in [42, 147], raga can be thought of as melodic atoms where atoms are the sequences of *swara*s (notes). Thus at the lowest level, a raga is composed of a sequence of swaras [148]. According to Pandit Jasraj, six primary *raga*s of Hindustani classical music are *Bhairav*, *Malkauns*, *Deepak*, *Shri*, *Megh* and *Hindol*. Each of them has five *ragini* (feminine counterpart of a raga) [146]. All other ragas are derived from these primary ragas and raginis.

Swara(note): Indian Classical Music is characterized by seven main swaras (pure notes) and together they are referred as *saptak* or *sargam* (gamut). The *swara*s are *Shadja (sa), Rishab (re), Gandhar (ga), Madhyam (ma), Pancham (pa), Dhaivat (dha)* and *Nishada (ni)*. The five intermediate swaras *re, ga, ma, pa, dha* are called *vikrit swaras* (altered notes). The swara *sa* and *pa* are called *Achal swaras* (unmovable notes) and other five swaras can have two or more variants. Re, Ga, Dha and Ni can be *Suddha* (pure) or *Komal* (soft) and Ma can be *Suddha* or *Tivra* (sharp). Thus, twelve swaras are there as shown in Table 5.1. *Sa* is accepted as the first or fundamental swara and the others appear consecutively in the frquency scale. As discussed in [149], swaras may be further categorized as *Dirgha* (prolonged), *Amsa* (frequent), *Alpa* (rarely used) etc. Thus, not only the sequence but also the roles played by the swaras is significant in characterizing a raga. In Table 5.1, correspondence between swaras and Western notes has been shown assuming Sa corresponds to C. But, in Hindusthani music, swara frequency is not fixed. A performer may choose different tonic frequency leading to a linear shift of the swaras. Even then the raga remains same.

Vadi and Samvadi Swara: Every raga has two important kinds of swaras, the *Vadi* (most significant) and the *Samvadi* (next in significance). These are important in the construction of the Raga. *Vadi* is usually the swara which is most frequent, and often it is the swara on which the singer can pause for a significant time or stressing it. The note that is prohibited from being used in a raga is called *Vivadi*. The rest are referred to as *Anuvadi* (residual) notes. The concept of *Vadi* and *Samvadi*

Table 5.2 *Vadi* and *Samvadi* swaras of the ragas.

Raga	**Vadi**	**Samvadi**
Bahar	ma	sa
Bhairav	dha	re
Bhairabi	ma	sa
Bageshri	ma	sa
Bibhas	dha re	pa sa
Bihag	ga	ni
Desh	re	pa
Durgaa	dha	re
Hamer	dha	ga
Juanpuri	sa	pa
Jog	ma	sa
Kafi	pa	sa
Kalyani(Yaman)	ga	ni
Purvi	ga	ni
Sarang	re	pa
Todi	dha	ga

swara is an important characteristics of raga. Table 5.2 provides a list of *Vadi* and *Samvadi* swara for different ragas.

Arohi and Avrohi: The selection of swaras of a raga has unique ascending and descending progression. In an octave, the specific ascending and descending order in which swaras within a raga are played is called the *Arohi* and *Avarohi*. Note sequence in the two progressions plays crucial role in raga composition. *Arohi* and *Avrohi* swara sequence for different ragas are shown in Table 5.3.

Combination of notes can be played to make the raga wonderful and sweet to listen. Such tunes match different sentiments and mood evoked in human mind and heart. Sentiments of ragas may match with the different feelings of natural phenomena like rain, storm, thunder, murmuring of river, undulation of waves, sense of infinity of universe and many others. Other feelings like loneliness, love, joy, detachment, rage, and sorrow may be realized if the ragas are played masterly. Role of the swaras and their combinations give rise to the flavors of raga.

5.3 Past work

Approaches for developing the automated systems to identify the raga in Indian classical music have strong resemblance with the raga recognition approaches of a human being. Human being follows either an intuitive or an analytic approach [44]. In intuitive approach, a listener relies on his vast experience of classical music. For a new piece of music, he identifies the raga by matching it with the already known tunes. This approach is focused on learn by example. On the other hand, the

Table 5.3 *Arohi* and *Avrohi* sequence of the ragas.

Raga	**Arohi**	**Avrohi**
Bahar	sa ma, pa Ga ma, dha, ni sa	re ni sa dha Ni pa, ma pa Ga ma, re sa
Bhairav	sa Re ga, ma, pa Dha, ni sa	sa ni Dha, pa ma ga, Re, sa
Bhairabi	sa Re Ga ma pa Dha Ni sa'	sa' Ni Dha pa ma Ga Re sa
Bageshri	sa Ga ma dha Ni sa'	sa' Ni dha ma pa dha ma Ga re sa
Bibhas	sa Re ga pa Dha sa'	sa' Dha pa ga Re sa
Bihag	'ni sa ga ma pa ni sa'	sa' ni dha pa Ma ga ma ga re sa
Desh	sa re ma pa ni sa'	sa' Ni dha pa ma ga re sa re' Ni dha pa, dha ma ga re, ga 'ni sa
Durga	sa re ma pa dha sa'	sa' dha pa ma re sa
Hamer	sa re ga ma dha ni sa'	sa' ni dha pa ma pa ga ma re sa
Juanpuri	sa re ma pa dha ma pa dha ni sa'	sa' ni dha pa dha ma pa ga re sa
Jog	sa ga ma pa ni sa'	sa' ni pa ma ga ma Ga sa
Kafi	sa re Ga ma pa dha Ni sa'	sa' Ni dha pa dha ma Ga re sa
Kalyani	'ni re ga Ma dha ni sa	sa' ni dha pa Ma ga re sa
Purvi	'ni Re ga Ma Dha ni sa'	sa' ni Dha pa Ma ga ma Re ga, Ma ga Re sa
Sarang	sa re ma pa ni sa'	sa' Ni pa ma re sa
Todi	sa re ga ma pa dha ni sa'	sa' ni dha pa ma ga re sa

analytic approach is more knowledge oriented. A listener is well equipped with the knowledge of structure and grammar of classical music and recognizes a raga in a systematic manner by identifying the swaras (notes) and analyzing the note sequence. The properties like *arohi, avrohi, vadi, samvadi* etc. (discussed in Section 5.2) play important role in identifying a raga. In intuitive approach, a listener characterize a raga by its overall acoustic impact which is the combined outcome of different properties of a raga. Thus, the properties are implicitly utilized in recognition. On the other hand, in analytic approach a listener consciously tries to identify the properties and thereafter utilizes the same to recognize a raga.

In an automated system it is important to design the descriptors from a music signal. In this process very often the knowledge of classical music properties are taken into consideration to ensure that in a way the descriptors can reflect the properties. Subsequently the descriptors are used to recognize the ragas based on machine learning techniques or by matching analytically. Thus, the automated raga identification systems are mostly hybrid in nature as it combines both the intuitive and analytic approach.

Considerable amount of past efforts focused on transcript oriented descriptors. Several researchers [150, 36, 151] have worked with hidden Markov model (HMM). Sequence of notes contributes in estimating the underlying structure of raga and it is exploited in modeling the ragas. It may be noted that such schemes require the segmentation of notes from a music signal and subsequent

identification. This can be thought of as generation of transcript. Pandey et al. [150] automated the process of extracting the notes. But, it is heuristic and moreover applied on solo vocal. However, the performance of HMM based raga identification for Hindustani music [151] is not good enough. Sridhar et al. [35] also followed a similar methodology based on note sequence. Measure was taken to identify the fundamental frequency of the singer and used it to minimize the effect of instruments. In these transcript based schemes, corresponding to each raga template of note sequence is stored. For an unknown music clip, extracted note sequence is compared with those stored templates using string matching techniques or using classifiers like K-NN, SVM etc. Shrey et al. [37] have presented a raga verification technique using longest common segment set. The success of all these schemes depends heavily on the performance of note extraction and identification (*i.e.* transcript generation).

Apart from transcript like descriptors, researchers also tried to develop features that can directly capture certain property of raga. An initial work was proposed by Chakravorty et al. [152]. In this work, instead of the music signal notation is used as input. It follows an analytic approach. At the first stage forbidden notes are identified from the notation and based on that a set of possible ragas are taken as the candidates for match. At the next level, *arohi-avrohi* section is extracted from the notation and lexically matched with those of the candidate ragas for final identification. Shetty et al. [153] worked with *vadi* notes along with *arohi-avrohi* properties. In Indian classical music, melody is an important aspect. A raga can have number of frequently occurring melodic phrases where a phrase stands for a sequence of notes. Normally, these phrases are sung at the initial stage and acts as a clue to the listener. Several works [37, 154, 155] are reported on melodic phrase based raga identification. But, proper extraction of melodic phrase is a big challenge.

A wide variety of low level features computed from the signal are also used as descriptors. Dighe et al. [41] relied on mel -frequency cepstral co-efficient (MFCC), chroma and timbre features. Swara(Note) histogram based structural analysis [38] is utilized in raga identification. Pitch based features [39, 40, 42–44] are in wide use. Based on pitch histogram variants of pitch-class profile are presented in [42, 43]. Pitch-profiles are obtained by weighting the quantified bins of the pitch histogram. Quantification is done in a manner that bins correspond to the notes. Bins are weighted once with the number of occurrences of corresponding note and once with the total duration of all the instance of that note to obtain the pitch profile. In [44], steady regions in the pitch contour are used as a pre-processing step to minimize the effect of instruments. In another variant [44], resolution is varied in forming the quantified histogram. Such pitch-profile lacks temporal information. Hence, along with pitch-profile, n-gram histograms are also considered in the work of Kumar et al. [39].

From the survey it is observed that most of the works are guided by the properties of ragas and accordingly focuses on the methodology. Transcript based approaches are more analytic but faces the challenge of automatic generation of transcript. For others, the descriptors reflect the characteristics of raga and at the next stage they rely on classifiers for the recognition. A non-linear support vector machine (SVM) that combines two kernels is used in [39] whereas K-NN classifier with KL divergence is tried in [44]. Clustering techniques are also used in recognizing the ragas [40]. Although different approaches and methodologies have been tried, it is still an active area of research.

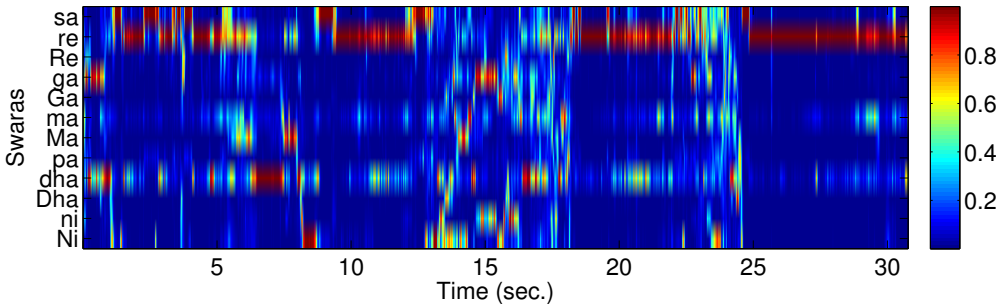


Fig. 5.1 Swara profile (chromatic scale profile) for raga *Malhar* (vocal).

5.4 Proposed Methodology

Proposed methodology can be thought of as the combination of analytic and intuitive approach. It consists of two major modules namely *feature extraction* and *raga identification*. Feature extraction module follows an analytic approach as it tries to capture the compositional properties of a raga. On the other hand, raga identification module is intuitive. It relies on a supervised classifier. Like an experienced listener, it also gathers experience during the learning phase and utilizes the same in identifying a raga during test phase.

5.4.1 Feature Extraction

Properties of a raga are described in Section 5.2. Proposed methodology focuses on the properties like *Vadi* and *Samvadi* swara, sequence of swara. Features are extracted from the music clip to capture these characteristics. In Indian classical music swara means a note in the octave. There are seven basic swaras of the scale are Sa, Re, Ga, Ma, Pa, Dha, and Ni. Among them, Re, Ga, Dha and Ni can be either *Suddha* (pure) or *Komal* (soft). Ma can also be either Suddha or Tivra (sharp). Thus, like Western chromatic scale, Indian classical music also has twelve swara. Features are computed based on the distribution of the strength (energy) of these swaras. We refer to this as swara profile and it serves as the foundation for designing the descriptors.

Swara Profile: There are twelve swaras or notes in Western music and Indian classical music. Depending on the scale their frequency may vary. A swara/note in successive scales are one actave apart. Taking scales/octaves into consideration, there are 88 music notes (A0 to C8) spreading over different scales or octaves. The swara profile stands for the signal energy distribution across a predefined set of tweleve swaras independent of their scale. It is also known as chroma scale profile. It is computed as follows [73, 156].

- The audio signal is decomposed into 88 frequency bands using filter banks. The sub-bands correspond to MIDI pitches or the piano notes (A0 to C8).

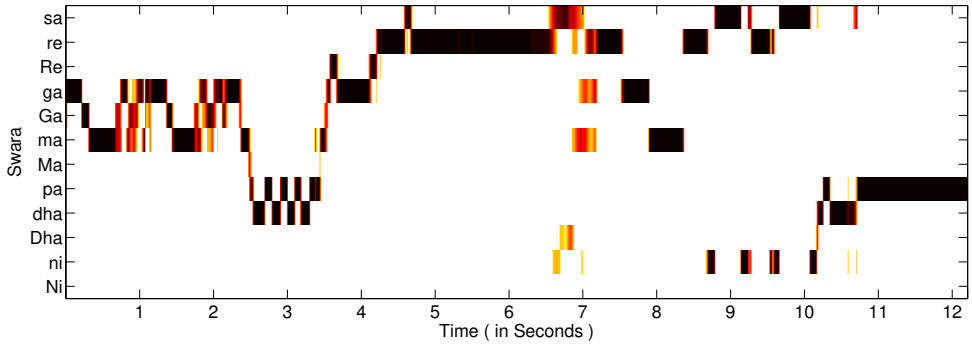


Fig. 5.2 Dominant swara for raga *Kaunshi Kanada (Flute).*

- Each sub-band signal is divided into number of frames with half overlapping. The hamming window function is applied on the frames and windowed output is used in subsequent step.
- For each sub-band, the magnitude spectrogram is obtained by applying fast Fourier transform (FFT) on windowed frames. The horizontal axis of the spectrogram corresponds to the time (or frame) and vertical axis corresponds to the energy.
- Magnitude spectrogram for each swara/note is obtained by adding up the spectrograms of the corresponding sub-bands. Twelve spectrograms thus obtained are normalized to form the swara profile or chroma scale profile. The profile is finally captured into a chromagram where the vertical axis represents twelve swaras and time scale (frame number) is along the horizontal axis. A particular element in the chromagram stands for energy of the note or swara at a particular instance.

Figure 5.1 shows a sample swara profile of a vocal clip of raga *Malhar* performed by Pandit Jasraj. The red color represents strong (high energy) swaras whereas blue color represents weak ones. The swara profile or chromagram of a raga reflects the strength of each swara in the raga on a temporal scale. A swara with very low strength (intensity value) indicates its absence in the profile. As a global visualization of the contribution of each swara in the raga is captured in the chromagram, it helps us in deriving the features to characterize a raga.

The different swaras in a raga have different levels of significance and that too varies with time. All these contribute to important property in characterizing a raga. As discussed in Section 5.2 *vadi* is the most frequent swara. The *samvadi* is the second-most prominent swara after *vadi*. A swara which is neither emphasized nor de-emphasized is called *anuvadi*. Swaras which are de-emphasized are referred to as being *durbal* or weak, while swaras which are excluded are called *vivadi*. *Vadi* swara, along with the *Samvadi* swara of a raga, usually brings out the uniqueness of the raga. In general, the dominance based categorization of the swaras in a raga is an important property. In order to represent this aspect we formulate the descriptors from the swara profile of a raga.

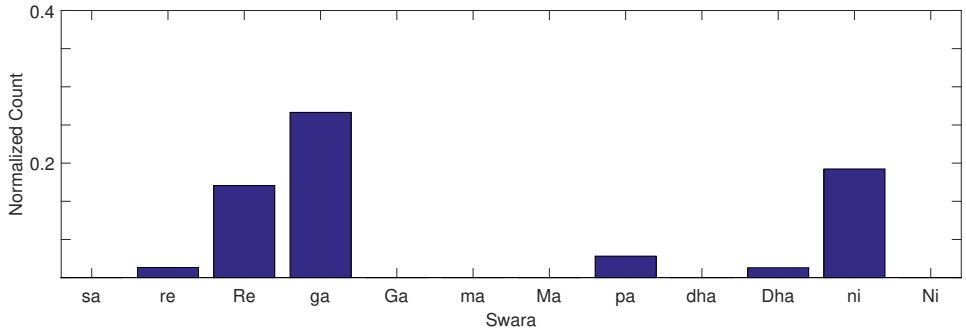


Fig. 5.3 Occurrence-histogram for raga *Maru Bihag* (vocal).

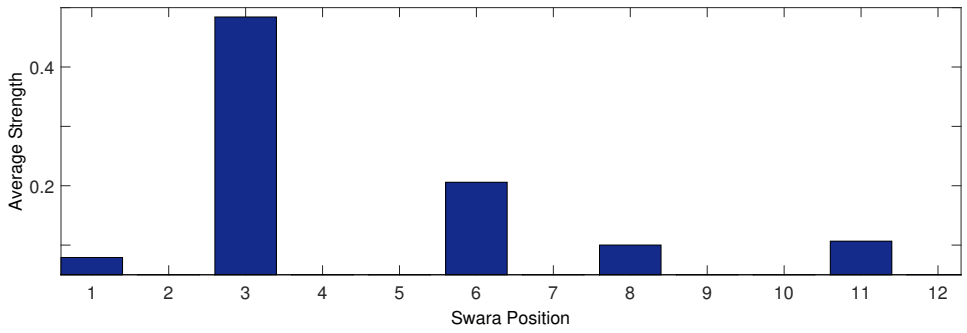
Occurrence Histogram of Dominant Swaras: It is 12-dimensional descriptor that shows the frequency for each swara appearing as the most dominating one. It is computed as follows.

- Corresponding to each frame in the chromagram (swara profile), the swara with maximum energy is determined. It is the most dominating swara of the frame.
- A 12-dimensional occurrence-histogram showing the count for each swara is obtained. Count denotes number of times the particular swara is most dominating.
- Finally, the occurrence-histogram is normalized to form the features based on dominating swara.

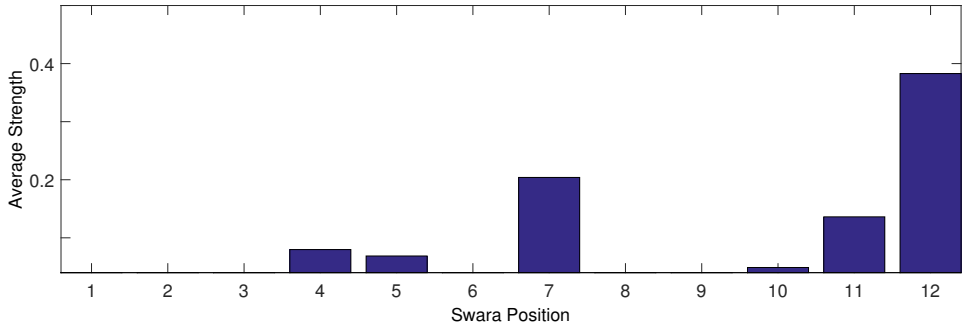
Ideally, top two peaks in the occurrence-histogram correspond to *vadi* and *samvadi* note. The minima correspond to *vivadi* swaras. The histogram thus summarizes the relative significance of the twelve swaras. It may be noted that, in our work, individual swaras are not extracted from the signal. A frame may contain multiple swaras. As a result, the swara profile presents the relative energy of all the swaras in a frame. Hence only the most dominating swara from each frame is considered to form the occurrence-histogram. The time complexity for computing the feature from the chromagram is of $O(n)$ where n is the number of frames.

Figure 5.2 shows the dominant swaras at different time instances (frames) obtained by processing the audio signal of raga *Kaunshi Kanada*, a flute based performance of Pandit Hari Prasad Chaurasia. The darker color in that figure indicates strongly emphasized swaras. Figure 5.3 shows the occurrence-histogram for vocal clip on *Maru Bihag* raga. It also shows that *Gandhara* (ga) swara as *vadi* and *Nishada* (ni) as *Samvadi* and they correspond to top two peaks in the occurrence-histogram.

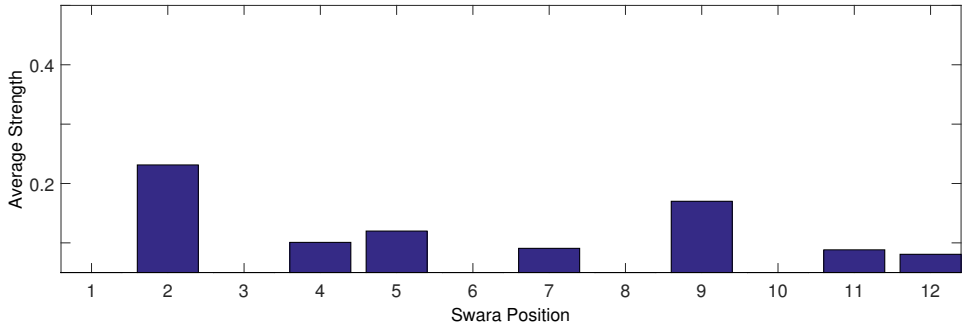
Strength Distribution of Swaras: It is also a 12-dimensional feature vector that captures the average strength of the swaras in the music signal. In each frame of the chromagram (swara profile), only the swaras with normalized energy higher than a threshold are considered. These are taken as the significant swaras in the frames. In our experiment a very low threshold is chosen. For the insignificant swaras in the frames, strength is taken as zero. Average strength of each swara is computed by considering all the frames. 12-dimensional feature vector thus obtained is normalized



(a)



(b)



(c)

Fig. 5.4 Strength distribution of swaras for different ragas: (a) *Raga Malkauns Jor (Sitar) – Audhav jaati* (five swaras), (b) *Raga Bahar – Shadav jaati* (six swaras) and (c) *Raga Yaman – Sampurna jaati* (seven swaras).

and taken as the strength distribution vectors. It incurs a time complexity of $O(n)$, n being the number of frames in the chromagram.

Figure 5.4 shows the strength distribution for different ragas. It is worth noting that ragas are also categorized into *jaati* based on the number of significant swaras present. *Audhav, Sadhav* and *Sampurna jaati* are examples of *jaati* with five, six and seven significant swaras respectively. It is observed in Figure 5.4 that proposed strength distribution is able to detect the number of significant swaras present and thus it is atleast capable of classifying the ragas based on *jaati*. Two or more ragas may belong to same *jaati*. The relative strength distribution can be utilized in discriminating the ragas belonging to same *jaati*.

Features based on sequence of swaras: The main essence of a raga depends on the sequence of notes/swaras in the composition. Ragas are also correlated with emotion. The sequence of swaras results in to the arousal of different emotions like joy, sorrow, excitement. It has motivated us to capture the note sequence in the form of a co-occurrence matrix. It is a 12×12 matrix where (i, j)-th element denotes count of the co-occurrence of notes i and j. The steps for computing the matrix is as follows.

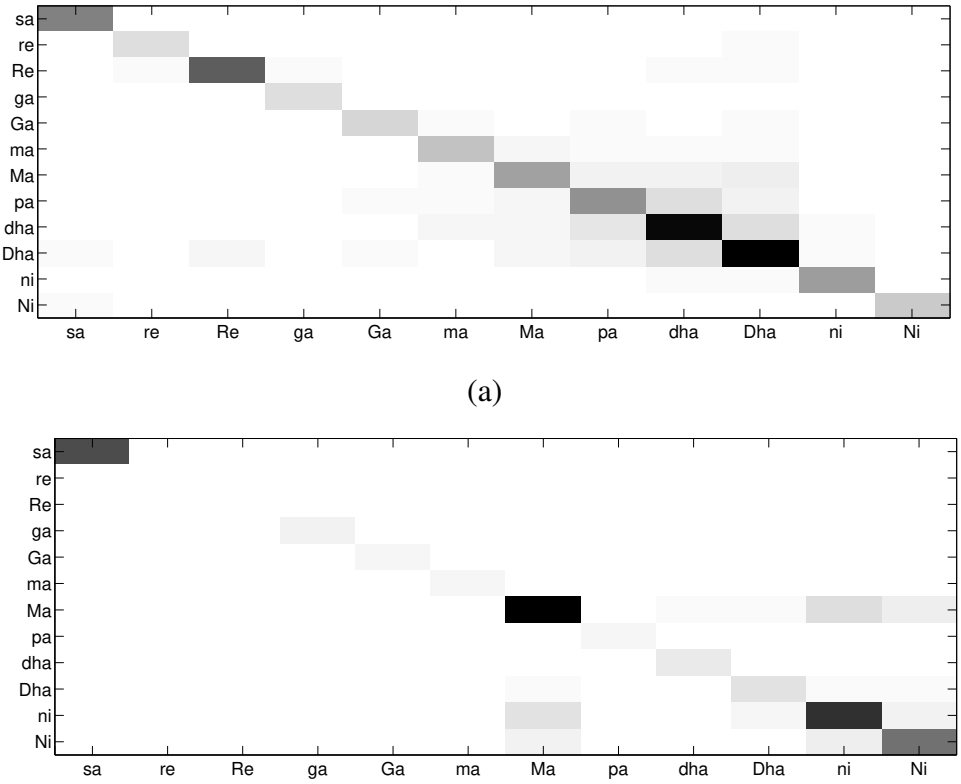
- Initialize each element of the matrix, $mat[i][j]$ with zero.
- At each time instance (frame) of the chromagram (swara profile), consider the swara with maximum energy as the dominating one.
- For each pair of consecutive time instance t and $t+1$
  - Let d_1 and d_2 are the dominating swara at time instance t and $t+1$.
  - $mat[d_1][d_2] = mat[d_1][d_2] + 1$.
- Normalize the matrix by dividing each element by the sum of all elements in the matrix.

It may be noted that only dominant swara of the frame is considered. As the matrix is prepared for the clip consisting of number of frames it can capture the distribution of the note sequence of a raga. The normalized matrix provides a sort of probability distribution of the occurrence of note pairs. The time complexity for computing the co-occurrence matrix is of $O(n)$ where n is the number of frames in the chromagram.

Figure 5.5 shows a sample co-occurrence matrices for raga *Marwa* and raga *Bageshree*. Darker a matrix element, more frequent is the corresponding note pair. Moreover, the effect of *Arohi* and *Avrohi* will also be embedded in the upper and lower triangle of the matrix. Thus, the matrix summarizes the note sequence of the raga along with its ascent and descent pattern.

5.4.2 Classification

In this work, we have used multi class support vector machine (SVM) as classifier [135]. Essentially it is binary classifier that handles multi classes through a series of one versus rest classification. In



(a)

(b)

Fig. 5.5 Co-occurrence matrix of swaras for different ragas: (a) *Raga Marwa* and (b) *Raga Bageshree*.

SVM, the goal is to learn the mapping: $X \rightarrow Y$ where $x \in X$ is descriptor or feature vector and $y \in Y$ is a class label (*i.e.* raga in our case). It tries to find the maximum-margin hyperplane separating two different class y_i and y_j. It is a complex optimization problem. Sequential minimal optimization (SMO) [157] is an iterative algorithm that solves such optimization problem. We have used SMO for training the SVM and it makes the process faster.

5.5 Experimental Results

In order to carry out the experiment, we have prepared a dataset that reflects wide variety. Both, the vocal and instrumental based raga performances are included in the collection. The audio clips correspond to twenty four different ragas of Hindusthani classical music covering the performance of more than fifty renowned instrumental artists and more than twenty five great vocal artists. Table 5.4 provides a brief description of the dataset. All the recordings are sampled at 22050 Hz and mono channeled.

Table 5.4 Description of the *raga* dataset.

Type	The dataset contains both Vocal and Instrumental based ragas
Collection	The dataset contains 1648 raga clips. The duration of each clip is 45 seconds. Among them, 1190 raga clips from instrumental and 458 raga clips from vocal performances.
Ragas	Bageshri, Bahar, Bhairabi, Bhairav, Bibhas, Bihag, Desh, Durga, Hamer, Jaunpuri, Jog, Kafi, Kalyani(yaman), Kanada, Kedar, Khamaj, Kirwani, Lalit, Malhar, Malkauns, Marwa, Purvi(Purvagauda), Sarang, Todi

Instead of working with the complete recording, we have tried to identify the raga based on a small part of it. We have considered clips of 45 seconds duration. As discussed in Section 5.4, each clip is represented by a 168-dimensional feature vector. It includes occurrence histogram of dominant swaras, strength distribution of swaras and co-occurrence of note pairs. First two categories of the features are global and they do not capture temporal information. Co-occurrence of note pairs incorporates temporal aspects to some extent. Finally, SVM classifier is used for classification. The proposed methodology is tested separately on vocal and instrumental dataset. Confusion matrices for the vocal and instrumental collections are shown in Table 5.5 and 5.6 respectively. Classification accuracy of the proposed system based on ten fold cross validation is summarized in Table 5.7.

Table 5.5 Confusion matrix for vocal collection.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Raga
1	75.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	3.7	0	0	0	5.6	1.8	Kanada
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Bibhas
3	0	0	78.1	0	0	0	0	0	0	0	0	0	0	0	0	12.5	0	0	0	0	0	9.4	0	Bihag
4	11.1	0	5.6	72.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11.1	0	Sarang
5	0	0	0	0	83.3	0	0	0	0	0	0	0	0	0	0	16.7	0	0	0	0	0	0	0	Desh
6	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	Durga
7	0	0	20	0	0	0	60	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	Hamer
8	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	Kafi
9	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaunpuri
10	25	0	0	0	0	0	0	0	0	50	0	0	0	0	0	12.5	0	0	0	0	0	12.5	0	Jog
11	2.9	0	0	0	0	0	0	0	0	0	82.4	0	0	0	0	14.7	0	0	0	0	0	0	0	Bageshri
12	27.7	0	5.6	0	0	0	0	0	0	0	0	61.1	0	0	0	5.6	0	0	0	0	0	0	0	Kalyani
13	10	0	0	15	0	0	0	0	0	0	0	0	35	0	0	30	0	0	0	0	0	10	0	Kedar
14	12.5	0	0	0	0	0	0	0	0	0	0	0	0	50	0	25	0	0	0	0	0	0	12.5	Khamaj
15	14.3	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	14.3	0	21.4	0	Lalit
16	22	0	0	0	0	0	0	0	0	0	0	0	2	0	0	66	0	0	0	0	0	10	0	Malhar
17	10	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0	0	10	0	Malkauns
18	9.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	81.8	0	0	0	9.1	0	Marwa
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62.5	37.5	0	0	0	Purvi
20	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	5	0	0	0	70	0	15	0	Todi
21	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	10	50	0	0	Bahar
22	6	0	3	0	0	0	0	0	0	0	0	0	0	0	0	6	6	0	0	0	0	77.3	1.6	Bhairav
23	9.1	0	0	9.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.1	72.7	Bhairabi

We have compared the performance of our system with two other systems [44, 39] on the same dataset. Koduri et al. [44] relied on pitch based features. First of all pitch contour is determined following the methodology proposed by Salamon et al. [158]. Stable regions in the pitch contours are then extracted and based on the same twelve dimensional (each corresponds to a swara) octave folded pitch histogram is formed. The $p_{instance}$ and $p_{duration}$ are each 12-dimensional feature vectors. These are weighted pitch histogram. Each bin in the histogram is weighted by the number of instances of the swara in the clip to obtain $p_{instance}$. For $p_{duration}$ weight is the duration of corresponding swara in the clip. For recognition, modified KL divergence is used as the distance measure and k-NN is used as classifier. Vijay et al. [39] in their work have considered pitch based $p_{instance}$ and $p_{duration}$ as features and KL-divergence as corresponding distance measure. Additionally n-gram distribution of notes is considered. To compare such distributions radial basis function kernel is used. Finally, SVM framework that combines kernel over two sets of features and corresponding distance measures has been used for raga identification. We have implemented both the systems and carried out the

Table 5.6 Confusion matrix for instrumental collection.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Raga
1	94.7	0	0	0	0	0	0	0	0	0	0	0	0	2.1	0	0	1.1	0	0	2.1	0	0	0	Kanada
2	0	89.6	0	0	2.6	2.6	0	0	0	0	2.6	0	0	0	0	2.6	0	0	0	0	0	0	0	Bihag
3	0	0	90.4	0	0	0	0	0	0	0	3.9	0	3.9	0	0	1.8	0	0	0	0	0	0	0	Sarang
4	0	0	0	79.2	0	0	0	0	0	0	0	0	16.6	0	0	4.2	0	0	0	0	0	0	0	Desh
5	0	0	0	0	93.8	0	0	0	0	0	3.1	0	0	0	0	0	0	0	0	3.1	0	0	0	Durga
6	0	0	0	0	0	71.9	0	0	0	0	0	0	6.3	0	0	0	0	0	0	0	18.8	3	0	Hamer
7	0	0	0	0	0	0	84.9	0	0	4.3	0	0	4.3	0	0	0	0	0	0	2.2	0	0	4.3	Kafi
8	0	0	0	0	75	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaunpu.
9	0	0	1.9	0	0	0	1.9	0	86.5	0	0	0	3.9	0	0	3.9	0	0	0	1.9	0	0	0	Jog
10	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	4.8	0	0	0	1.2	0	0	0	Bageshr
11	2.9	0	0	0	0	0	1.5	0	0	2.9	75	0	1.5	0	0	2.9	0	0	0	1.5	0	0	11.8	Kalyani
12	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	7	0	0	0	0	0	7	0	Kedar
13	5	0	1.3	0	0	1.3	0	0	0	0	2.5	0	82.5	0	0	3.7	0	0	0	3.7	0	0	0	Khamaj
14	0	0	0	0	0	0	4.3	0	0	0	2.2	0	0	91	0	2.2	0	0	0	0	0	0	0	Kirwani
15	6.3	0	0	0	0	0	0	0	0	6.3	0	0	0	0	75	0	0	9.3	0	3.1	0	0	0	Lalit
16	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	91	0	1	0	0	1	3	0	Malhar
17	0	0	0	0	0	0	0	0	2.2	0	0	0	2.2	0	0	0	91	0	0	4.3	0	0	0	Malka.
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	Marwa
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	67	0	0	0	Purvi
20	5.8	0	0	0	0	0	0	0	0	0	0	0	1.2	0	0	1.2	0	0	0	89.5	0	0	2.3	Todi
21	0	0	0	0	0	18.8	0	0	0	0	9.4	0	0	0	0	6.2	0	3.1	0	0	62.5	0	0	Bahar
22	6.8	0	2.7	0	0	0	0	0	2.7	0	0	0	4.1	0	0	2.7	0	3	0	0	0	77	0	Bhairav
23	2.6	0	0	0	0	0	0	3.8	0	0	20.5	0	6.4	0	0	10.3	0	0	0	0	0	0	56.4	Bhaira.

Table 5.7 Classification accuracy (in%) of proposed system.

Raga	**Instrumental Dataset**	**Vocal Dataset**
Bahar	62.50	50.00
Bhairav	77.03	77.27
Bhairabi	56.41	72.73
Bageshri	94.05	82.35
Bibhas	–	100.00
Bihag	89.47	78.13
Desh	79.17	83.33
Durgaa	93.75	75.00
Hamer	71.86	60.00
Juanpuri	25.00	100.00
Jog	86.54	50.00
Kafi	84.78	50.00
Kanada	94.68	75.92
Kedar	85.71	35.00
Khamaj	82.50	50.00
Kirwani	91.30	–
Kalyani(Yaman)	75.00	61.11
Lalit	75.00	50.00
Malhar	91.03	66.00
Malkaunsh	91.31	75.00
Marwa	100.00	81.82
Purvi	33.33	62.5
Sarang	90.38	72.22
Todi	89.54	70.00
Overall	84.29	70.52

Table 5.8 Performance comparison: classification accuracy (in %) of different systems.

System	Instrumental Dataset	Vocal Dataset
Koduri et al. [44] using $p_{instance}$	54.96	49.78
Koduri et al. [44] using $p_{duration}$	58.91	50.22
Vijay et al. [39]	62.99	55.20
Proposed System	84.29	70.52

experiments on our vocal and instrumental dataset. The performance of all the systems is presented in Table 5.8. It is observed that performance of the proposed methodology is superior.

5.6 Summary

We have presented a simple but novel scheme to identify the *raga* in *Hindusthani* classical music. Features are designed to reflect the compositional properties of a *raga*. First of all pitch based *swara* (note) profile is formed. It is utilized to generate occurrence histogram of dominant *swara*s and their strength distribution. Such descriptors have strong correlation with the properties that categorizes the role of *swara*s in the *raga*s. Note sequence is an important property and note co-occurrence matrix captures this temporal aspect to an extent. Thus, features are devised following an analytic approach adopted in manual recognition. Training of the SVM classifier emulates the intuitive approach of a human being where identification relies on prior knowledge gathered by listening the music. Proposed methodology thus combines analytic and intuitive approach. Experiment has been carried out with a diversified dataset and compared the performance with other systems. It is observed that proposed system works better. In future, domain knowledge may be utilized to analyze the classification errors and to take measures for further improvement.

Chapter 6

Multi-Aspect Classification

6.1 Introduction

Automatic classification of music signal is important for organized storage of large collection of music data and also for easy retrieval. The commonly used search criteria includes *genre*, *singer*, *emotion*. For classical music *raga* is very important criteria. Methodologies to classify music based on individual metadata have been developed and described in the previous chapters. Performances have been verified by experimenting mostly with the benchmark dataset. In this chapter we combine the features which have been designed for describing the individual metadata. Finally, multi-aspect classification is done that identifies attributes like *genre*, *singer* and *emotion*. For this purose we have also prepared a dataset and groundthuthed with metadata.

6.2 Methodology

Like the previous cases, the methodology consists of feature extraction and then classification based on those features. The features to capture *genre*, *singer* and *emotion* are mostly as discussed in earlier chapters. For the sake of readability and ready reference we mention the same as follows.

Descriptor for Genre: In Chapter 2, two methodologies have been presented. In one case empirical mode decomposition (EMD) was deployed and subsequently features were computed. In the other approach a set low level features have been computed from the signal without any preprocessing in the form of decomposition. EMD being a very slow process, we have followed the alternate approach based on the low level features as follows.

- Timbral Features
  - Mel Frequency Cepstral Coefficients (MFCCs)
  - Spectral Flux (SF)

  - Spectral Rolloff (SR)
  - Spectral Centroid (SC)
  - Spectral Spread (SSP)
  - Spectral Slope (SSL)
- Tonality Features
  - Spectral Flatness Measure (SFM)
  - Spectral Crest Factor (SCF)
  - Tonal Power Ratio (TPR)
- Statistical Features
  - Spectral Kurtosis (SK)
  - Spectral Skewness
- Pitch based features

Descriptor for Singer: As discussed in Chapter 3, a song is the composition of singing voice and instrumental music. As per composition some segments may contain voice with or without accompanying background music and some segments may have only the background music. The music excerpts are segmented and non-vocal segments are removed based on the time domain energy distribution. A simple frequency domain filtering is applied on the vocal segments to minimize the impact of accompanying instruments. Then MFCCs features and spectrogram based vocal-print features are extracted from the filtered vocal segments to represent the singer characteristics.

Descriptor for Emotion: In Chapter 4, two methodologies have been proposed. One based on deep learning and the other one is based on low level features. To keep the things less expensive and the dataset is not so large, we have adopted low level feature based methodology. The feature set includes the following.

- Timbral Features
  - Mel Frequency Cepstral Coefficients (MFCCs)
  - Spectral Flux (SF)
  - Spectral Rolloff (SR)
  - Spectral Centroid (SC)
  - Spectral Spread (SSP)
- Tonality Features

- Spectral Flatness Measure (SFM)
- Spectral Crest Factor (SCF)

It may noted that features are a subset of descriptors for genre. This can be attributed to the notion that genre and emotion have a close association.

Classification: We have used Neural Network with two hidden layers for classification. Three different networks for *genre*, *singer* and *emotion* are considered. Relevant feature vectors are fed to corresponding network. Each network has nodes in input and output layers sames as the dimension of input vector and number of output classes. Number of nodes in the hidden layers are decided based on nodes in the input and output layers as it has been discussed in Section 2.3.2.

6.3 Results

In order to carry out the experiment, we have prepared a dataset consisting of 202 music clips. The dataset has been annotated with three different metadata - *Genre*, *Singer* and *Emotion*. The dataset contains the recordings of six different singers - *Abbasuddin Ahmed*, *Asha Bhosle*, *Anita Saha*, *Pandit Jasraj*, *Kishore Kumar* and *Anup Jalota*. There are four different genres - *Folk*, *Rabindra sangeet* (a subset reflecting a specific genre type), *Devotional* and *Classical*. Songs belong to four different emotions - *Joy*, *Peaceful*, *Romantic* and *Sadness*. To carry out the experiment, we have considered the mono channel music excerpts of 30 seconds duration, sampled at 22050 Hz.

Table 6.1 Accuracy (in %) for *genre* based classification.

Genre	**Accuracy(in %)**
Folk	98.36
Tagore	100.00
Devotional	95.35
Classical	100.00
Overall	*98.51*

Part of the data has been used to train the network and rest are used for testing. Five fold cross validation has been done, average results are being reported. Table 6.1 shows the genre based classification accuracy. It is observed that an overall classification accuracy of 98.51% is achieved. The classification accuracy for Singer identification is 98.02%. The corresponding accuracy for different singers is shown in Table 6.2. On our *Raga* dataset also we have tried to identify the singer and an overall accuracy of 91.27% has been achieved. Table 6.3 shows the classification accuracy for emotion detection. This being the most difficult task success is limited and an overall accuracy of 76.73% has been obtained.

Table 6.2 Accuracy (in %) for *singer* based classification.

Singer	**Accuracy (in %)**
Abbasuddin	95.75
Asha	100.00
Anita	92.86
Jasraj	100.00
Kishore	100.00
Anup	97.67
Overall	*98.02*

Table 6.3 Accuracy (in %) for *emotion* based classification.

Emotion	**accuracy (in %)**
Joy	63.85
Peaceful	75.00
Romantic	62.63
Sadness	85.71
Overall	*76.53*

Accuracy for classification on multiple aspects is reported in Table 6.4. Different combination of the aspects (*i.e. emotion*, *genre* and *singer*) have been considered. It is observed as the accuracy for detecting *emotion* limits the performance. A detailed result for all the aspects taken together is shown in Table 6.5 and an overall accuracy of 72.77% has been obtained.

6.4 Summary

This work combines the low level descriptors to represent the metadata of music like, *genre*, *singer* and *emotion*. A dataset consisting music clips of different genres, singers and emotional category has been prepared and annotated. With this the methodology for multi-aspect classification of music data has been tested. Even without costly descriptors based on EMD (for genre) and deep learning based approach, proposed methodology performs satisfactorily. It is observed that for *singer* and *genre* identification performance is quite high and it suffers in identifying the emotion. It also affects the multi-aspect classification. In future efforts may be directed to improve on this respect.

Table 6.4 Overall classification accuracy (in %) for combined aspects.

Combination	**Accuracy (in %)**
genre	98.51
singer	98.02
emotion	76.53
genre + singer	97.03
emotion + genre	73.27
emotion + singer	74.75
emotion + genre + singer	*72.77*

Table 6.5 Accuracy (in %) for classification based on genre- singer-emotion taken together.

Genre + Singer + Emotion	**Accuracy (in %)**
Folk-Abbasuddin-Joy	50.00
Folk-Abbasuddin-Peaceful	40.00
Folk-Abbasuddin-Romantic	33.33
Folk-Abbasuddin-Sadness	81.82
Tagore-Asha-Joy	100.00
Tagore-Asha-Peaceful	50.00
Tagore-Asha-Romantic	50.00
Tagore-Asha-Sadness	80.00
Folk-Anita-Sadness	100.00
Classical-Jasraj-Joy	50.00
Classical-Jasraj-Peaceful	63.16
Classical-Jasraj-Romantic	42.86
Classical-Jasraj-Sadness	73.08
Tagore-Kishore-Joy	50.00
Tagore-Kishore-Peaceful	66.67
Tagore-Kishore-Romantic	60.00
Tagore-Kishore-Sadness	92.86
Devotional-Anup-Peaceful	93.02

Chapter 7

Conclusion

In the context of a music retrieval system proper organization of the large collection of music data is very important. Data can be archived in a structured manner based on various metadata like *genre*, *singer*, *emotion*. For Indian classical music *raga* is one of the significant characteristics. One approach for extracting such metadata may be manual where domain expert annotates the piece of music. It enables text based retrieval for a metadata oriented music query. As the volume of data is quite large, manual process of annotation is time consuming and laborious also. Moreover, certain characteristics like *genre*, *emotion* are subjective and expert opinion may vary. *Raga* requires high level of domain expertise. The problem of annotation may be less severe as nowadays there exists different music formats with metadata embedded in it. But music recorded from other sources lacks this information. A major concern arises when the user does not provide the metadata as music query. On the contrary, user may submit the music clip as the query and expecting the music with similar characteristics from the retrieval. It necessitates for an automatic classification of music signal based on *genre*, *singer*, *emotion* etc. Present work has made an effort towards that objective.

We have proposed a simple methodology for genre based classification which has been detailed in Chapter 2. Empirical mode decomposition is utilized in extracting desired signal component by ignoring the extreme (high and low frequency) characteristics. Finally, from the extracted signal feature vector has been formed based on the local energy distribution over various pitch bands. Multi-layer perceptron network is used for classification. Experimental result shows that proposed descriptor performs better than the conventional features.

For singer based classification as detailed in Chapter 3, simple preprocessing has been proposed to extract the vocal segments and also to minimize the effect of background music on those. It enabled us to capture the characteristics of the singing voice emphatically. Spectrogram based vocal-print have been proposed that takes care of voice production system and timbral aspects and reflects the characteristics of a singer. MFCC based features are also combined to reflect the perceptual aspect.

Emotion is perceptual and subjective. Hence it is quite a complex task to accomplish emotion based classification. Chapter 4 has detailed two different approaches. In one approach, wide range

of time domain and spectral features are chosen based on the past efforts of the researchers. Experimentally it has been observed that a moderate performance can be achieved. It may be attributed to the difficulty in designing the low level features to represent emotion. To get rid of this issue deep learning based approach has been considered. A modified version of VGGNet with comparatively less number of layers has been proposed. It works with the audio segment of very small duration, even of five seconds to recognize the emotion. Experimental result shows that proposed network improves the recognition accuracy considerably.

In Chapter 5 we have elaborated the methodology to identify the *raga* in *Hindusthani* classical music. Features have been designed to reflect the compositional properties of a *raga*. Pitch based *swara* (note) profile has been proposed and utilized to generate occurrence and strength distribution of the *swaras*. Such descriptors have strong correlation with the properties that categorizes the role of *swara*s in the *raga*s. Proposed methodology combines analytic and intuitive approach.

Chapter 6 summarizes the whole work by combining the descriptors for multi-aspect classification. A dataset has been prepared for this purpose and annotated with *genre*, *singer* and *emotion*. The performance on the same has been found satisfactory.

Proposed methodologies have been tested on benchmark datasets and performance is compared with number of existing works. It has been observed that proposed schemes work better. However, there are certain areas that may be explored in future. As music is very often multi-instrumentals or voice with instrumentals, a robust filtering technique to separate the components can be a big boost for designing the features for any such applications. In case of *genre* identification, proposed EMD based methodology is time consuming. It is mainly due to the decomposition step. An alternate faster technique with comparable capability can be a future objective. For emotion recognition a CNN based technique has been proposed. In future, CNN-LSTM network can be considered. In general, it is observed that emotion detection is still very open as we have achieved limited success. In case of *raga* based classification of Indian classical music, there is scope to understand and utilize domain knowledge further. In this work, we have relied on deep learning only in case of emotion based classification to surmount the problem of designing suitable features. However, there is enormous opportunity in exploring the design and use of deep learning network for classifying the music data on each aspect under consideration.

References

[1] Jialie Shen, John Shepherd, Bin Cui, and Kian-Lee Tan. A Novel Framework for Efficient Automated Singer Identification in Large Music Databases. *ACM Trans. on Information Systems*, 27(3), 2009.

[2] C C Liu and C S Huang. A singer identification technique for content-based classification of MP3 music objects. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 506–511, 2004.

[3] T Li, M Ogihara, and Q Li. A comparative study on content-based music genre classification. In *Proceedings of the Annual ACM SIGIR International Research and Development in Information Retrieval*, pages 282–289, 2003.

[4] T Zhang. Automatic singer identification. In *Proceedings of the International Conference on Multimedia and Expo*, pages 33–36, 2003.

[5] T Li and M Ogihara. Music artist style identification by semisupervised learning from both lyrics and content. In *Proceedings of the Annual ACM International Conference on Multimedia*, pages 364–367, 2004.

[6] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organised sound*, 4(3):169–175, 2000.

[7] Bob L Sturm. A survey of evaluation in music genre recognition. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval*, pages 29–66. Springer, 2012.

[8] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce. Music Genre Classification via Joint Sparse Low-rank Representation of Audio Features. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(12):1905–1917, dec 2014.

[9] Yin-Fu Huang, Sheng-Min Lin, Huan-Yu Wu, and Yu-Siou Li. Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data & Knowledge Engineering*, 92:60–76, 2014.

[10] Konstantin Markov and Tomoko Matsui. Music genre and emotion recognition using Gaussian processes. *IEEE access*, 2:688–697, 2014.

[11] Alexander Schindler and Andreas Rauber. An audio-visual approach to music genre classification through affective color features. In *Proceedings of the European Conference on Information Retrieval*, pages 61–67. Springer, 2015.

[12] Loris Nanni, Yandre Costa, and Sheryl Brahnam. Set of texture descriptors for music genre classification. In *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2014.

[13] Loris Nanni, Yandre M G Costa, Alessandra Lumini, Moo Young Kim, and Seung Ryul Baek. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, 45(Supplement C):108–117, 2016.

[14] A L Berenzweig and D P W Ellis. Locating singing voice segments within music signal. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[15] A L Berenzweig, D P W Ellis, and S Lawrence. Using voice segments to improve artist classification of music. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[16] W Cai, Q Li, and X Guan. Automatic singer identification based on auditory features. In *Proceedings of the International Conference on Natural Computation*, pages 1624–1628, 2003.

[17] Li Su and Yi-Hsuan Yang. Sparse modeling for artist identification: exploiting phase information and vocal separation. In *Proceedings of the International Society for Music Information Retrieval Conference*, nov 2013.

[18] Nadine Kroher and E Gómez. Automatic Singer Identification For Improvisational Styles Based On Vibrato, Timbre And Statistical Performance Descriptors. In *Proceedings of the International Computer Music Conference/Sound and Music Computing Conference*, Athens, Greece, 2014.

[19] Ying Hu and Guizhong Liu. Singer identification based on computational auditory scene analysis and missing feature methods. *Journal of Intelligent Information Systems*, 42(3):333–352, 2014.

[20] Ying Hu and Guizhong Liu. Separation of Singing Voice Using Nonnegative Matrix Partial Co-factorization for Singer Identification. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(4):643–653, 2015.

[21] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proceedings of the International Society for Music Information Retrieval*, pages 255–266, 2010.

[22] Yi-Hsuan Yang and Homer H Chen. Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):40:1—-40:30, 2012.

[23] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the International conference on digital audio effects*, pages 237–244, 2007.

[24] Yu-An Chen, Ju-Chiang Wang, Yi-Hsuan Yang, and Homer Chen. Component Tying for Mixture Model Adaptation in Personalization of Music Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(7), 2017.

[25] Lucía Martín Gómez and María Navarro Cáceres. Applying Data Mining for Sentiment Analysis in Music. In *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 198–205, 2017.

[26] Felix Weninger, Fabien Ringeval, Erik Marchi, and Björn Schuller. Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2196–2202, 2016.

[27] Xin Liu, Qingcai Chen, Xiangping Wu, Yan Liu, and Yang Liu. CNN based music emotion classification. *arXiv preprint arXiv:1704.05665*, 2017.

[28] Fan Zhang, Hongying Meng, and Maozhen Li. Emotion extraction and recognition from music. In *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1728–1733, 2016.

[29] Christian Koch, Ganna Krupii, and David Hausheer. Proactive Caching of Music Videos Based on Audio Features, Mood, and Genre. In *Proceedings of the Multimedia Systems Conference*, pages 100–111, 2017.

[30] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[31] J A Russell. A circumspect model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[32] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech Emotion Recognition Using CNN. In *Proceedings of the ACM International Conference on Multimedia*, pages 801–804, 2014.

[33] E M Albornoz, M Sänchez-Gutiërrez, F M Martinez, H L Rufiner, and J Goddard. Spoken emotion recognition using deep learning. In *Proceedings of the Iberoamerican Congress on Pattern Recognition*, pages 104–111, 2014.

[34] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn W Schuller. Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks. In *Proceedings of the MediaEval*, 2015.

[35] Rajeswari Sridhar and T. V. Geetha. Raga identification of carnatic music for music information retrieval. *International Journal of Recent Trends in Engineering*, 1(1), 2009.

[36] M. Sinith and K. Rajeev. Hidden markov model based recognition of musical pattern in south indian classical music. In *Proceedings of the IEEE International Conference on Signal and Image Processing*, 2006.

[37] Shrey Dutta, Krishnaraj Sekhar PV, and Hema A. Murthy. Raga verification in carnatic music using longest common segment set. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 605–611, 2015.

[38] Bhiksha Raj Pranay Dighe, Harish Karnick. Raga verification in carnatic music using longest common segment set. In *Proceedings of International Society for Music Information Retrieval Conference*, 2013.

[39] V. Kumar, H. Pandya, and C. V. Jawahar. Identifying ragas in indian music. In *Proceedings of the International Conference on Pattern Recognition*, pages 767–772, Aug 2014.

[40] K. P. Kumar and M. S. Rao. Raaga identification using clustering algorithm. In *Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques*, pages 2077–2081, 2016.

[41] P. Dighe, P. Agrawal, H. Karnick, S. Thota, and B. Raj. Scale independent raga identification using chromagram patterns and swara based features. In *Proceedings of the International Conference on Multimedia and Expo Workshops*, pages 1–4, 2013.

[42] P. Chordia and A. Rae. Raag recognition using pitch- class and pitchclass dyad distributions. In *Proceedings of International Society for Music Information Retrieval Conference*, 2007.

[43] Gopala Krishna Koduri, Sankalp Gulati, and Preeti Rao. A survey of raaga recognition techniques and improvements to the state-of-the-art. *Sound and Music Computing*, 38:39–41, 2011.

[44] Gopala Krishna Koduri, Sankalp Gulati, Preeti Rao, and Xavier Serra. Rāga recognition based on pitch distribution methods. *Journal of New Music Research*, 41(4):337–350, 2012.

[45] A L Berenzweig, B Logan, D P W Ellis, and B Whitman. A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures. *Journal of Computational Music*, 28(2):63–76, 2004.

[46] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Trans. Audio, Speech and Lang. Proc.*, 18(2):310–319, feb 2010.

[47] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[48] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Paiva. Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset. In *Proceedings of the International Workshop on Music and Machine Learning*, 2016.

[49] R Sarkar, N Biswas, and S Chakraborty. Music genre classification using frequency domain features. In *Proceedings of the International Conference on Emerging Applications of Information Technology*, 2018.

[50] R Sarkar and S K Saha. Music genre classification using EMD and pitch based feature. In *Proceedings of the International Conference on Advances in Pattern Recognition*, pages 1–6, jan 2015.

[51] R. Sarkar and S. K. Saha. Singer based classification of song dataset using vocal signature inherent in signal. In *Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pages 1–4, Dec 2015.

[52] Rajib Sarkar and Sanjoy Kumar Saha. Singer wise classification of song data using mfcc and spectrogram based vocal-print. *Communicated to International Journal of Pattern Recognition and Artificial Intelligence, World Scientific*.

[53] Rajib Sarkar, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. Emotion based categorization of music using low level features and agglomerative clustering. In *Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pages 506–516, 2018.

[54] Rajib Sarkar, Sombuddha Choudhury, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. Recognition of emotion in music based on deep convolutional neural network. *Communicated to Multimedia Tools and Applications, Springer*.

[55] Rajib Sarkar, Soumya Kanti Naskar, and Sanjoy Kumar Saha. Raga identification from hindustani classical music signal using compositional properties. *Computing and Visualization in Science*, pages 1–12, 2017.

[56] P Cook G. Tzanetakis A. Ermolinskyi. Pitch histograms in audio and symbolic music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, pages 31–38, 2002.

[57] J Zhou L. Xiao. Using chroma histogram to measure the perceptual similarity of music. In *Proceedings of the International Conference on Multimedia and Expo*, pages 1317–1320, 2008.

[58] I Cohen M. Genussov. Musical genre classification of audio signals using geometric methods. In *Proceedings of the European Signal Processing Conference*, pages 497–501, 2010.

[59] E Tsunoo, G Tzanetakis, N Ono, and S Sagayama. Beyond Timbral Statistics: Improving Music Classification Using Percussive Patterns and Bass Lines. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):1003–1014, 2011.

[60] J P Bello. Measuring Structural Similarity in Music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2013–2025, sep 2011.

[61] A Rauber R. Mayer. Music Genre Classification by Ensembles of Audio and Lyrics Features. In *Proceedings of the International Conference on Music Information Retrieval*, pages 675–680, 2011.

[62] P J León R. Mayer A. Rauber, C Pérez-Sancho, and J M Iñesta. Feature Selection in a Cartesian Ensemble of Feature Subspace Classifiers for Music Categorisation. In *Proceedings of the ACM Multimedia Workshop on Music and Machine Learning*, pages 53–56, 2010.

[63] C Xu, N C Maddage, X Shao, F Cao, and Q Tian. Musical genre classification using support vector machines. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 429–432, 2003.

[64] C Mckay and I Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the International Conference on Music Information Retrieval*, pages 525–530, 2004.

[65] Y Anan, K Hatano, H Bannai, and M Takeda. Music Genre Classification Using Similarity Functions. In *Proceedings of the International Conference on Music Information Retrieval*, 2011.

[66] Tong Zhang and C-C Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on speech and audio processing*, 9(4):441–457, 2001.

[67] Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Conference on Music Information Retrieval*, 2000.

[68] Peter Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[69] Chi-Wah Kok. Fast algorithm for computing discrete cosine transform. *IEEE Transactions on Signal Processing*, 45(3):757–760, 1997.

[70] A Gray and J Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3):207–217, 1974.

[71] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 1st edition, 2012.

[72] Norden E Huang, Man-Li C Wu, Steven R Long, Samuel SP Shen, Wendong Qu, Per Gloersen, and Kuang L Fan. A confidence limit for the empirical mode decomposition and hilbert spectral analysis. In *Proceedings of the royal society of london a: Mathematical, physical and engineering sciences*, pages 2317–2345, 2003.

[73] Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Conference on Music Information Retrieval*, 2011.

[74] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[75] F Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408, 1958.

[76] P Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

[77] Jan Dean. Artificial Neural Network, MLP, Backpropagation. *https://cw.felk.cvut.cz/wiki/_media/courses/a4m33bia-03backprop-2012.pdf*, Czech Technical University in Prague.

[78] E Tsunoo, G Tzanetakis, N Ono, and S Sagayama. Beyond Timbral Statistics: Improving Music Classification Using Percussive Patterns and Bass Lines. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):1003–1014, may 2011.

[79] Diego Furtado Silva, Rafael Geraldeli Rossi, Solange Oliveira Rezende, and Gustavo Enrique de Almeida Prado Alves Batista. Music classification by transductive learning using bipartite heterogeneous networks. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2014.

[80] Athanasios Lykartsis and Stefan Weinzierl. Using the beat histogram for rhythm description and language identification. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 2015.

[81] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce. Non-Negative Multilinear Principal Component Analysis of Auditory Temporal Modulations for Music Genre Classification. *IEEE Transactions on Audio, Speech & Language Processing*, 18(3), 2010.

[82] Constantine Kotropoulos, Gonzalo R Arce, and Yannis Panagakis. Ensemble Discriminant Sparse Projections Applied to Music Genre Classification. In *Proceedings of the International Conference on Pattern Recognition*, pages 822–825, 2010.

[83] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin. Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features. *IEEE Transactions on Multimedia*, 11(4):670–682, 2009.

[84] A F Arabi and Guojun Lu. Enhanced polyphonic music genre classification using high level features. In *Proceedings of the IEEE International Conference on Signal and Image Processing Applications*, pages 101–106, 2009.

[85] Yannis Panagakis and Constantine Kotropoulos. Music genre classification via Topology Preserving Non-Negative Tensor Factorization and sparse representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 249–252, 2010.

[86] Siddharth Sigtia and Simon Dixon. Improved music feature learning with deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[87] Y E Kim and B Whitman. Singer identification in popular music recordings using voice coding features. In *Proceedings of the International Conference on Music Retrieval*, pages 164–169, 2002.

[88] S Z K Khine, T L Nwe, and H Li. Exploring Perceptual Based Timbre Feature for Singer identification. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, pages 159–171, 2007.

[89] Bernhard Lehner, Reinhard Sonnleitner, and Gerhard Widmer. Towards light-weight, real-time-capable singing voice detection. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2013.

[90] Tushar Ratanpara and Narendra Patel. Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. *EURASIP J. on Audio, Speech and Music Processing*, 2015:16, 2015.

[91] Tsung-Han Tsai, Yu-Siang Huang, Pei-Yun Liu, and De-Ming Chen. Content-based singer classification on compressed domain audio data. *Multimedia Tools and Applications*, 74(4):1489–1509, 2015.

[92] Sunil Kumar Kopparapu, Meghna A Pandharipande, and G Sita. Music and vocal separation using multiband modulation based features. In *Proceedings of the IEEE Symposium on Industrial Electronics & Applications*, pages 733–737, 2010.

[93] Hye-Seung Cho, Jun-Yong Lee, and Hyoung-Gook Kim. Singing Voice Separation from Monaural Music Based on Kernel Back-Fitting Using Beta-Order Spectral Amplitude Estimation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 639–644, 2015.

[94] C Chen Julian. *Elements of Human Voice*. World Scientific Publishing Co., 2016.

[95] M Kob, N Henrich, H Herzel, D Howard, I Tokuda, and J Wolfe. Analysing and Understanding the Singing Voice: Recent Progress and Open Questions. *Current Bioinformatics*, 6:362–374, 2011.

[96] J Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *STL-QPSR*, 35(2-3):45–68, 1994.

[97] J Gauffin and Johan Sundberg. Spectral Correlates of Glottal Voice Source Waveform Characteristics. *J. Speech Hearing Research*, 32:556–565, sep 1989.

[98] Yoshiyuki Horii. Acoustic analysis of vocal vibrato: A theoretical interpretation of data. *Journal of Voice*, 3(1):36–43, 1989.

[99] Thibault Langlois and Gonçalo Marques. A music classification method based on timbral features. In *Proceedings of the International Conference on Music Information Retrieval*, pages 81–86, 2009.

[100] S Shirali-Shahreza, H Abolhassani, and M H Shirali-shahreza. Fast and scalable system for automatic artist identification. *IEEE Transactions on Consumer Electronics*, 55(3):1731–1737, aug 2009.

[101] Arefin Huq, Juan Pablo Bello, and Robert Rowe. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[102] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.

[103] Ali Hassan, Robert Damper, and Mahesan Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.

[104] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014.

[105] Guangwu Qian and Lei Zhang. A simple feedforward convolutional conceptor neural network for classification. *Applied Soft Computing*, 70:1034 – 1041, 2018.

[106] Yandre M.G. Costa, Luiz S. Oliveira, and Carlos N. Silla. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, 52:28 – 38, 2017.

[107] Jônatas Wehrmann and Rodrigo C. Barros. Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing*, 61:973 – 982, 2017.

[108] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[109] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[110] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[111] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools and Applications*, 76(5):6785–6799, 2017.

[112] G Trigeorgis, F Ringeval, R Brueckner, E Marchi, M A Nicolaou, B Schuller, and S Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 5200–5204, 2016.

[113] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, 2014.

[114] Carol L Krumhansl. Music: A link between cognition and emotion. *Current directions in psychological science*, 11(2):45–50, 2002.

[115] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2006.

[116] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *IEEE Trans. Audio, Speech & Language Processing*, 19(6):1802–1812, 2011.

[117] Alf Gabrielsson and Erik Lindström. *The influence of musical structure on emotional expression.* Oxford University Press, 2001.

[118] Byeong Jun Han, Seungmin Rho, Roger B Dannenberg, and Eenjun Hwang. SMERS: Music Emotion Recognition Using Support Vector Regression. In *Proceedings of the International Society for Music Information Retrieval*, pages 651–656, 2009.

[119] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457, 2008.

[120] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. Music emotion classification: A regression approach. In *Proceedings of the International Conference on Multimedia and Expo*, pages 208–211, 2007.

[121] Yu-Ching Lin, Yi-Hsuan Yang, and Homer H Chen. Exploiting Online Music Tags for Music Emotion Classification. *ACM Transactions Multimedia Computing Communications and Applications*, 7S(1):26:1—-26:16, 2011.

[122] Qi Lu, Xiaoou Chen, Deshun Yang, and Jun Wang. Boosting For Multi-Modal Music Emotion. In *Proceedings of the International Society for Music Information and Retrieval Conference*, page 105, 2010.

[123] Konstantin Markov, Motofumi Iwata, and Tomoko Matsui. Music Emotion Recognition using Gaussian Processes. In *Proceedings of the MediaEval*, 2013.

[124] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691, 2013.

[125] Densil Cabrera et al. Psysound: A computer program for psychoacoustical analysis. In *Proceedings of the Australian Acoustical Society Conference*, volume 24, pages 47–54, 1999.

[126] L Zao, D Cavalcante, and R Coelho. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. *IEEE signal processing letters*, 21(5):620–624, 2014.

[127] N. Thammasan, K. Fukui, and M. Numao. Application of deep belief networks in eeg-based dynamic music-emotion recognition. In *Proceedings of the International Joint Conference on Neural Networks*, pages 881–888, July 2016.

[128] Emmanuel Bigand, Sandrine Vieillard, François Madurell, Jeremy Marozeau, and A Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8):1113–1139, 2005.

[129] Lawrence R Rabiner and Ronald W Schafer. Introduction to Digital Speech Processing. *Found. Trends Signal Process.*, 1(1):1–194, 2007.

[130] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[131] Ossama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, oct 2014.

[132] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the International Conference on Machine Learning*, pages 807–814, 2010.

[133] Sylvie Droit-Volet, danilo Ramos, Lino Bueno, and Emmanuel Bigand. Music, emotion, and time perception: the influence of subjective emotional valence and arousal? *Frontiers in Psychology*, 4:417, 2013.

[134] K Sreenivasa Rao, V Ramu Reddy, and Sudhamay Maity. *Language Identification Using Spectral and Prosodic Features*. Springer, 2015.

[135] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[136] L Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[137] Marvin Minsky and Seymour Papert. *Perceptrons*. MIT press, 1969.

[138] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[139] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning*, chapter Random Forests, page 592. Springer, 2 edition, 2008.

[140] Yoav Goldberg. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[141] Peter Sadowski. Notes on backpropagation. *homepage: https://www. ics. uci. edu/˜ pjsadows/notes. pdf (online)*, 2016.

[142] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[143] François Chollet and Others. Keras, 2015.

[144] Bonnie Wade. Chĩz in khyāl: The traditional composition in the improvised performance. *Ethnomusicology*, 17(3):443–459, 1973.

[145] Jonathan Katz. Dhrupad: Tradition and performance in indian music. *Ethnomusicology Forum*, 14(1):113–115, 2005.

[146] Jujhar Singh. Interview with pandit jasraj. https://www.youtube.com/watch?v=VPshheRL69M, 2016. [You Tube; accessed 10 Sept. 2016].

[147] Arvindh Krishnaswamy. Melodic atoms for transcribing carnatic music. In *Proceedings of the International Conference on Music Information Retrieval*, 2004.

[148] M. Sharma. *Tradition of Hindustani Music*. A.P.H. Publishing Corporation, 2006.

[149] T. Viswanathan and M.H. Allen. *Music in South India: The Karnatak Concert Tradition and Beyond : Experiencing Music, Expressing Culture*. Global music series. Oxford University Press, 2004.

[150] Gurav Pandey, Gaurav P, Chaitanya Mishra, and Paul Ipe. Tansen : A system for automatic raga identification. In *In Proceedings of the Indian International Conference on Artificial Intelligence*, pages 1350–1363, 2003.

[151] Dipanjan Das and Monojit Choudhury. Finite state models for generation of hindustani classical music, 2005.

[152] J. Chakravorty and A.K Mukherjee, B.and Datta. Some studies on machine recognition of ragas in indian classical music. *Journal of Acoustic Society of India*, XVII(3 & 4):1–4, 1989.

[153] Surendra Shetty and K. K. Achary. Raga mining of indian music by extracting arohana-avarohana pattern. *International Journal of Recent Trends in Engineering*, 2009.

[154] Joe Cheri Ross and Preeti Rao. Detection of raga-characteristic phrases from hindustani. In *In Proceedings of the CompMusic Workshop*, 2012.

[155] Preeti Rao, Joe Cheri Ross, Kaustuv Kanti Ganguli, Vedhas Pandit, Vignesh Ishwar, Ashwin Bellur, and Hema Murthy. Classification of melodic motifs in raga music with time-series matching, 2014.

[156] Meinard Müller, Frank Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *In Proceedings of the International Conference on Music Information Retrieval*, pages 288–295, 2005.

[157] J PLATT. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, 1998.

[158] J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. on Audio, Speech and Language Processing*, 20(6):1759–1770, 2012.

Comput Visual Sci
DOI 10.1007/s00791-017-0282-x

S.I.: ICACNI 2016

Raga identification from Hindustani classical music signal using compositional properties

Rajib Sarkar[1] · **Soumya Kanti Naskar**[1] · **Sanjoy Kumar Saha**[1]

Received: 1 April 2017 / Accepted: 15 September 2017

Abstract Classification of music signal is a fundamental step for organized archival of music collection and fast retrieval thereafter. For Indian classical music, raga is the basic melodic framework. Manual identification of raga demands high expertise which is not available easily. Thus an automated system for raga identification is of great importance. In this work, we have studied the basic properties of the ragas in North Indian (Hindusthani) classical music and designed the features to capture the same. Pitch based Swara (note) profile is formed. Occurrence and energy distribution of notes generated from the profile are used as features. Note sequence plays an important role in the raga composition. Proposed note co-occurrence matrix summarizes this aspect. An audio clip is represented by these features which have strong correlation with the properties of raga. Support vector machine is used for classification. Experiment is done with a diversified dataset. Performance of the proposed work is compared with two other systems. It is observed that proposed methodology performs better.

Keywords Raga identification · Classical music · Raga properties

Rajib Sarkar
rjbskar@gmail.com

Soumya Kanti Naskar
rijunaskar@gmail.com

Sanjoy Kumar Saha
sks_ju@yahoo.co.in

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

1 Introduction

Music is one of the natural form of art that enters through our ear and spreads its essence over our mind and emotion. Indian Classical Music is regarded as one of the most prestigious and the highest class of music. Raga is the underlying structure of Indian classical music. For efficient storage of such music database, data may be organized based on their raga. Such structured collection makes the raga based retrieval quite efficient. Identification of raga serves as the fundamental step for such application. Manual detection of raga involves expertise of high degree and their availability is also an issue. Hence, an automated system for raga identification is of immense significance.

Hindustani classical music is mainly found in Northern part of India, Bangladesh and Pakistan. *Raga* and *tala* (rhythmic cycle) remain the central notion in both the systems. *Khayal* [27] and *Dhrupad* [7] are the two main forms of Hindustani classical music. *Dhrupad* is primarily of devotional type and performed by male singers. *Khyal* is a relatively newer Hindustani vocal music. It is very popular because of its romanticism and emotional influence. The vocal performance is accompanied by string instruments like *tanpura*, *veena*.

In this work, we deal with Hindusthani classical music. As the task of raga identification requires domain knowledge, fundamental concepts regarding *raga* and its properties are discussed in Sect. 2. Section 3 presents the review of past work. Proposed methodology is outlined in Sect. 4. Sections 5 and 6 provide the experimental results and concluding remarks respectively.

Published online: 23 September 2017

2 Basic properties of Raga

Raga is the discernible melodic form underlying all Hindusthani classical Music. It acts as a communication medium for two or more music lover's mind. A composition attracts the listeners due to its emotional content. Experts define the term raga in various ways. *Bharat Muni* in his *Natya Shastra* used the term raga to indicate aesthetic enjoyment or pleasure. According to Pandit Jasraj, the meaning of raga is *love* [23]. Matanga Muni coined the first technical definition of raga as "In the opinion of the wise, the particularity of notes and melodic movements, or that distinction of melodic sound by which one is delighted, is Raga" [12]. As described in [2,10], raga can be thought of as melodic atoms where atoms are the sequences of *swara*s (notes). Thus at the lowest level, a raga is composed of a sequence of swaras [21]. According to Pandit Jasraj, six primary *raga*s of Hindustani classical music are *Bhairav*, *Malkauns*, *Deepak*, *Shri*, *Megh* and *Hindol*. Each of them has five *ragini* (feminine counterpart of a raga) [23]. All other ragas are derived from these primary ragas and raginis.

Swara (note): Indian Classical Music is characterized by seven main swaras (pure notes) and together they are referred as *saptak* or *sargam* (gamut). The *swara*s are *Shadja (sa), Rishab (re), Gandhar (ga), Madhyam (ma), Pancham (pa), Dhaivat (dha)* and *Nishada (ni)*. The five intermediate swaras *re, ga, ma, pa, dha* are called *vikrit swaras* (altered notes). The swara *sa* and *pa* are called *Achal swaras* (unmovable notes) and other five swaras can have two or more variants. Re, Ga, Dha and Ni can be *Suddha* (pure) or *Komal* (soft) and Ma can be *Suddha* or *Tivra* (sharp). Thus, twelve swaras are there as shown in Table 1. Sa is accepted as the first or fundamental swara and the others appear consecutively in the frquency scale. As discussed in [26], swaras may be further categorized as *Dirgha* (prolonged), *Amsa* (frequent), *Alpa* (rarely used) etc. Thus, not only the sequence but also the roles played by the swaras is significant in characterizing a raga. In Table 1, correspondence between swaras and Western notes has been shown assuming Sa corresponds to C. But, in Hindusthani music, swara frequency is not fixed. A performer may choose different tonic frequency leading to a linear shift of the swaras. Even then the raga remains same.

Table 1 Swaras of Hindusthani music and western chromatic scale

Position	Swara	Symbol	Western note
1	Shadja	sa	C
2	Rishabha (Komal)	re	C#
3	Rishabha (Suddha)	Re	D
4	Gandhara (Komal)	ga	D#
5	Gandhara (Suddha)	Ga	E
6	Madhyama (Suddha)	ma	F
7	Madhyama (Tivra)	Ma	F#
8	Panchama	pa	G
9	Dhaivata (Komal)	dha	G#
10	Dhaivata (Suddha)	Dha	A
11	Nishada (Komal)	ni	A#
12	Nishada (Suddha)	Ni	B

Table 2 Vadi, Samvadi swaras of the ragas

Raga	Vadi	Samvadi
Bahar	ma	sa
Bhairav	dha	re
Bhairabi	ma	sa
Bageshri	ma	sa
Bibhas	dha re	pa sa
Bihag	ga	ni
Desh	re	pa
Durgaa	dha	re
Hamer	dha	ga
Juanpuri	sa	pa
Jog	ma	sa
Kafi	pa	sa
Kalyani (Yaman)	ga	ni
Purvi	ga	ni
Sarang	re	pa
Todi	dha	ga

Vadi and Samvadi Swara: Every raga has two important kinds of swaras, the *Vadi* (most significant) and the *Samvadi* (next in significance). These are important in the construction of the Raga. *Vadi* is usually the swara which is most frequent, and often it is the swara on which the singer can pause for a significant time or stressing it. The note that is prohibited from being used in a raga is called *Vivadi*. The rest are referred to as *Anuvadi* (residual) notes. The concept of *Vadi* and *Samvadi* swara is an important characteristics of raga. Table 2 provides a list of *Vadi* and *Samvadi* swara for different ragas.

Arohi and Avrohi: The selection of swaras of a raga has unique ascending and descending progression. In an octave, the specific ascending and descending order in which swaras within a Raga are played is called the *Arohi* and *Avarohi*. Note sequence in the two progressions plays crucial role in raga composition. *Arohi* and *Avrohi* swara sequence for different ragas are shown in Table 3.

Combination of notes can be played to make the raga wonderful and sweet to listen. Such tunes match different sentiments and mood evoked in human mind and heart. Sentiments of ragas may match with the different feelings of

Table 3 Arohi, Avrohi sequence of the ragas

Raga	Arohi	Avrohi
Bahar	sa ma, pa Ga ma, dha, ni sa	re ni sa dha Ni pa, ma pa Ga ma, re sa
Bhairav	sa Re ga, ma, pa Dha, ni sa	sa ni Dha, pa ma ga, Re, sa
Bhairabi	sa Re Ga ma pa Dha Ni sa'	sa' Ni Dha pa ma Ga Re sa
Bageshri	sa Ga ma dha Ni sa'	sa' Ni dha ma pa dha ma Ga re sa
Bibhas	sa Re ga pa Dha sa'	sa' Dha pa ga Re sa
Bihag	'ni sa ga ma pa ni sa'	sa' ni dha pa Ma ga ma ga re sa
Desh	sa re ma pa ni sa'	sa' Ni dha pa ma ga re sa re' Ni dha pa, dha ma ga re, ga 'ni sa
Durga	sa re ma pa dha sa'	sa' dha pa ma re sa
Hamer	sa re ga ma dha ni sa'	sa' ni dha pa ma pa ga ma re sa
Juanpuri	sa re ma pa dha ma pa dha ni sa'	sa' ni dha pa dha ma pa ga re sa
Jog	sa ga ma pa ni sa'	sa' ni pa ma ga ma Ga sa
Kafi	sa re Ga ma pa dha Ni sa'	sa' Ni dha pa dha ma Ga re sa
Kalyani	'ni re ga Ma dha ni sa	sa' ni dha pa Ma ga re sa
Purvi	'ni Re ga Ma Dha ni sa'	sa' ni Dha pa Ma ga ma Re ga, Ma ga Re sa
Sarang	sa re ma pa ni sa'	sa' Ni pa ma re sa
Todi	sa re ga ma pa dha ni sa'	sa' ni dha pa ma ga re sa

natural phenomena like rain, storm, thunder, murmuring of river, undulation of waves, sense of infinity of universe and many others. Other feelings like loneliness, love, joy, detachment, rage, and sorrow may be realized if the ragas are played masterly. Role of the swaras and their combinations give rise to the flavors of raga.

3 Past work

Approaches for developing the automated systems to identify the raga in Indian classical music have strong resemblance with the raga recognition approaches of a human being. Human being follows either an intuitive or an analytic approach [8]. In intuitive approach, a listener relies on his vast experience of classical music. For a new piece of music, he identifies the raga by matching it with the already known tunes. This approach is focused on learn by example. On the other hand, the analytic approach is more knowledge oriented. A listener is well equipped with the knowledge of structure and grammar of classical music and recognizes a raga in a systematic manner by identifying the swaras (notes) and analyzing the note sequence. The properties like *arohi, avrohi, vadi, samvadi* etc. (discussed in Sect. 2) play important role in identifying a raga. In intuitive approach, a listener characteristics a raga by its overall acoustic impact which is the combined outcome of different properties of a raga. Thus, the properties are implicitly utilized in recognition. On the other hand, in analytic approach a listener consciously tries to identify the properties and thereafter utilizes the same to recognize a raga.

In an automated system it is important to design the descriptors from a music signal. In this process very often the knowledge of classical music properties are taken into consideration to ensure that in a way the descriptors can reflect the properties. Subsequently the descriptors are used to recognize the ragas based on machine learning techniques or by matching analytically. Thus, the automated raga identification systems are mostly hybrid in nature as it combines both the intuitive and analytic approach.

Considerable amount of past efforts focused on transcript oriented descriptors. Several researchers [4,15,24] have worked with hidden Markov model (HMM). Sequence of notes contributes in estimating the underlying structure of raga and it is exploited in modeling the ragas. It may be noted that such schemes require the segmentation of notes from a music signal and subsequent identification. This can be thought of as generation of transcript. Pandey et al. [15] automated the process of extracting the notes. But, it is heuristic and moreover applied on solo vocal. However, the performance of HMM based raga identification for Hindustani music [4] is not good enough. Sridhar and Geetha [25] also followed a similar methodology based on note sequence. Measure was taken to identify the fundamental frequency of the singer and used it to minimize the effect of instruments. In these transcript based schemes, corresponding to each raga template of note sequence is stored. For an unknown music clip, extracted note sequence is compared with those stored templates using string matching techniques or using classifiers like K-NN, SVM. Dutta et al. [6] have presented a raga verification technique using longest common segment set. The success of all these schemes depends heavily on the per-

formance of note extraction and identification (i.e. transcript generation).

Apart from transcript like descriptors, researchers also tried to develop features that can directly capture certain property of raga. An initial work was proposed by Chakravorty et al. [1]. In this work, instead of the music signal notation is used as input. It follows an analytic approach. At the first stage forbidden notes are identified from the notation and based on that a set of possible ragas are taken as the candidates for match. At the next level, *arohi-avrohi* section is extracted from the notation and lexically matched with those of the candidate ragas for final identification. Shetty and Achary [22] worked with *vadi* notes along with *arohi-avrohi* properties. In Indian classical music, melody is an important aspect. A raga can have number of frequently occurring melodic phrases where a phrase stands for a sequence of notes. Normally, these phrases are sung at the initial stage and acts as a clue to the listener. Several works [6, 18,19] are reported on melodic phrase based raga identification. But, proper extraction of melodic phrase is a big challenge.

A wide variety of low level features computed from the signal are also used as descriptors. Dighe et al. [5] relied on Mel-frequency cepstral co-efficient (MFCC), chroma and timbre features. Swara (Note) histogram based structural analysis [17] is utilized in raga identification. Pitch based features [2,8,9,11,12] are in wide use. Based on pitch histogram variants of pitch-class profile are presented in [2,9]. Pitch-profiles are obtained by weighting the quantized bins of the pitch histogram. Quantization is done in a manner that bins correspond to the notes. Bins are weighted once with the number of occurrences of corresponding note and once with the total duration of all the instance of that note to obtain the pitch profile. In [8], steady regions in the pitch contour are used as a pre-processing step to minimize the effect of instruments. In another variant [8], resolution is varied in forming the quantized histogram. Such pitch-profile lacks temporal information. Hence, along with pitch-profile, n-gram histograms are also considered in the work of Kumar et al. [12].

From the survey it is observed that most of the works are guided by the properties of ragas and accordingly focuses on the methodology. Transcript based approaches are more analytic but faces the challenge of automatic generation of transcript. For others, the descriptors reflect the characteristics of raga and at the next stage they rely on classifiers for the recognition. A non-linear support vector machine (SVM) that combines two kernels is used in [12] whereas K-NN classifier with KL divergence is tried in [8]. Clustering techniques are also used in recognizing the ragas [11]. Although different approaches and methodologies have been tried, it is still an active area of research.

4 Proposed methodology

Proposed methodology can be thought of as the combination of analytic and intuitive approach. It consists of two major modules namely *feature extraction* and *raga identification*. Feature extraction module follows an analytic approach as it tries to capture the compositional properties of a raga. On the other hand, raga identification module is intuitive. It relies on a supervised classifier. Like an experienced listener, it also gathers experience during the learning phase and utilizes the same in identifying a raga during test phase.

4.1 Feature extraction

Properties of a raga are described in Sect. 2. Proposed methodology focuses on the properties like *Vadi* and *Samvadi* swara, sequence of swara. Features are extracted from the music clip to capture these characteristics. In Indian classical music swara means a note in the octave. There are seven basic swaras of the scale are Sa, Re, Ga, Ma, Pa, Dha, and Ni. Among them, Re, Ga, Dha and Ni can be either *Suddha* (pure) or *Komal* (soft). Ma can also be either Suddha or Tivra (sharp). Thus, like Western Chromatic scale, Indian Classical music also has twelve swara. Features are computed based on the distribution of the strength (energy) of these swaras. We refer to this as swara profile and it serves as the foundation for designing the descriptors.

Swara profile: There are twelve swaras or notes in Western music and Indian classical music. Depending on the scale their frequency may vary. A swara/note in successive scales are one actave apart. Taking scales/octaves into consideration, there are 88 music notes (A0–C8) spreading over different scales or octaves. The swara profile stands for the signal energy distribution across a predefined set of tweleve swaras independent of their scale. It is also known as chroma scale profile. It is computed as follows [13,14].

- The audio signal is decomposed into 88 frequency bands using filter banks. The sub-bands correspond to MIDI pitches or the piano notes (A0–C8).
- Each sub-band signal is divided into number of frames with half overlapping. The hamming window function is applied on the frames and windowed output is used in subsequent step.
- For each sub-band, the magnitude spectrogram is obtained by applying Fast Fourier Transform (FFT) on windowed frames. The horizontal axis of the spectrogram corresponds to the time (or frame) and vertical axis corresponds to the energy.
- Magnitude spectrogram for each swara/note is obtained by adding up the spectrograms of the corresponding sub-bands. Twelve spectrograms thus obtained are nor-

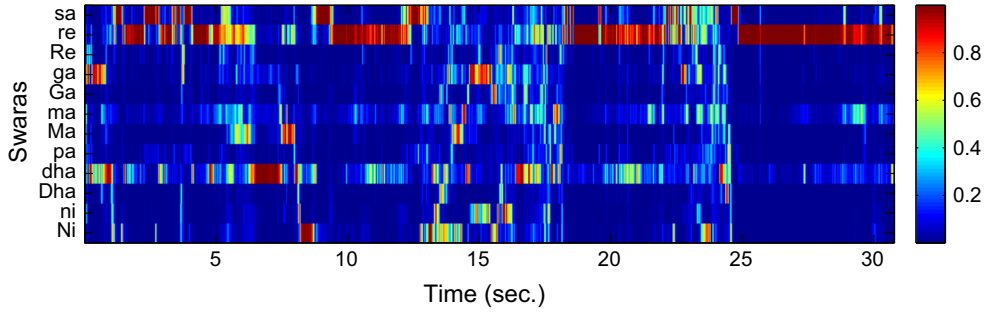


Fig. 1 Swara profile (chromatic scale profile) for Raga Malhar (vocal)

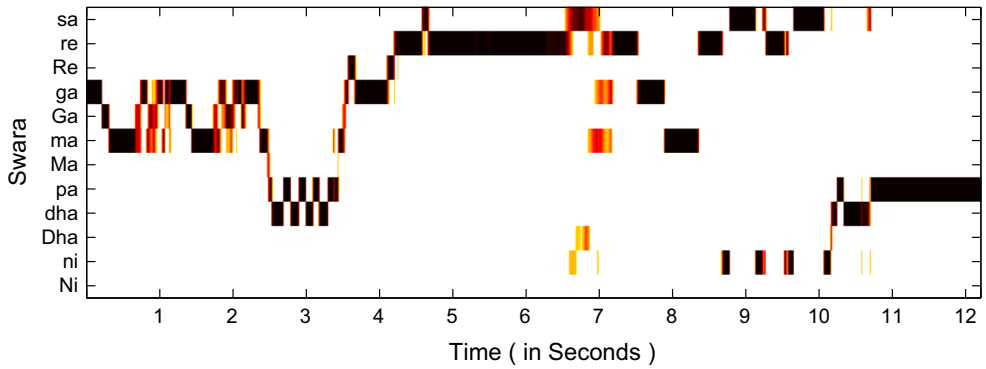


Fig. 2 Dominant Swara for Raga Kaunshi Kanada (flute)

malized to form the swara profile or chroma scale profile. The profile is finally captured into a chromagram where the vertical axis represents twelve swaras and time scale (frame number) is along the horizontal axis. A particular element in the chromagram stands for energy of the note or swara at a particular instance.

Figure 1 shows a sample swara profile of a vocal clip of Raga *Malhar* performed by Pandit Jasraj. The red color represents strong (high energy) swaras whereas blue color represents weak ones. The swara profile or chromagram of a raga reflects the strength of each swara in the raga on a temporal scale. A swara with very low strength (intensity value) indicates its absence in the profile. As a global visualization of the contribution of each swara in the raga is captured in the chromagram, it helps us in deriving the features to characterize a raga.

The different swaras in a raga have different levels of significance and that too varies with time. All these contribute to important property in characterizing a raga. As discussed in Sect. 2 *vadi* is the most frequent swara. The *samvadi* is the second-most prominent swara after *vadi*. A swara which is neither emphasized nor de-emphasized is called anuvadi. Swaras which are de-emphasized are referred to as being durbal or weak, while swaras which are excluded are called *vivadi*. *Vadi* swara, along with the *Samvadi* swara of a raga, usually brings out the uniqueness of the raga. In general, the dominance based categorization of the swaras in a raga is an important property. In order to represent this aspect we formulate the descriptors from the swara profile of a raga.

Occurrence histogram of dominant swaras: It is 12-dimensional descriptor that shows the frequency for each swara appearing as the most dominating one. It is computed as follows.

- Corresponding to each frame in the chromagram (swara profile), the swara with maximum energy is determined. It is the most dominating swara of the frame.
- A 12-dimensional occurrence-histogram showing the count for each swara is obtained. Count denotes number of times the particular swara is most dominating.
- Finally, the occurrence-histogram is normalized to form the features based on dominating swara.

Ideally, top two peaks in the occurrence-histogram correspond to *vadi* and *samvadi* note. The minima correspond to *vivadi* swaras. The histogram thus summarizes the relative significance of the twelve swaras. It may be noted that, in our work, individual swaras are not extracted from the signal. A frame may contain multiple swaras. As a result, the swara profile presents the relative energy of all the swaras in a frame. Hence only the most dominating swara from each

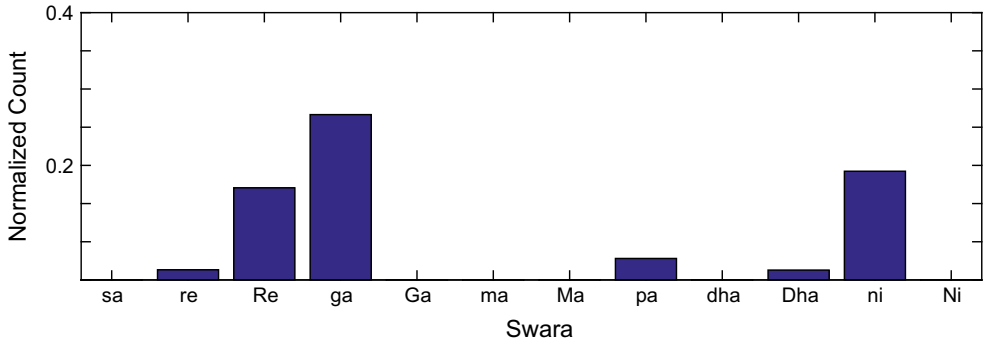


Fig. 3 Occurrence-histogram for Raga Maru Bihag (vocal)

frame is considered to form the occurrence-histogram. The time complexity for computing the feature from the chromagram is of $O(n)$ where n is the number of frames.

Figure 2 shows the dominant swaras at different time instances (frames) obtained by processing the audio signal of raga *Kaunshi Kanada*, a flute based performance of Pandit Hari Prasad Chaurasia. The darker color in that figure indicates strongly emphasized swaras. Figure 3 shows the occurrence-histogram for vocal clip on *Maru Bihag* raga. It also shows that *Gandhara* (ga) swara as *vadi* and *Nishada* (ni) as *Samvadi* and they correspond to top two peaks in the occurrence-histogram.

Strength distribution of swaras: It is also a 12-dimensional feature vector that captures the average strength of the swaras in the music signal. In each frame of the chromagram (swara profile), only the swaras with normalized energy higher than a threshold are considered. These are taken as the significant swaras in the frames. In our experiment a very low threshold is chosen. For the insignificant swaras in the frames, strength is taken as zero. Average strength of each swara is computed by considering all the frames. 12-dimensional feature vector thus obtained is normalized and taken as the strength distribution vectors. It incurs a time complexity of $O(n)$, n being the number of frames in the chromagram.

Figure 4 shows the strength distribution for different ragas. It is worth noting that ragas are also categorized into *jaati* based on the number of significant swaras present. *Audhav, Sadhav* and *Sampurna jaati* are examples of *jaati* with five, six and seven significant swaras respectively. It is observed in Fig. 4 that proposed strength distribution is able to detect the number of significant swaras present and thus it is atleast capable of classifying the ragas based on *jaati*. Two or more ragas may belong to same *jaati*. The relative strength distribution can be utilized in discriminating the ragas belonging to same *jaati*.

Features based on sequence of swaras: The main essence of a raga depends on the sequence of notes/swaras in the composition. Ragas are also correlated with emotion. The sequence of swaras results in to the arousal of different emotions like joy, sorrow, excitement. It has motivated us to capture the note sequence in the form of a co-occurrence matrix. It is a 12×12 matrix where (i, j)-th element denotes count of the co-occurrence of notes i and j. The steps for computing the matrix is as follows.

- Initialize each element of the matrix, $mat[i][j]$ with zero.
- At each time instance (frame) of the chromagram (swara profile), consider the swara with maximum energy as the dominating one.
- For each pair of consecutive time instance t and $t + 1$
  - Let d_1 and d_2 are the dominating swara at time instance t and $t + 1$.
  - $mat[d_1][d_2] = mat[d_1][d_2] + 1$.
- Normalize the matrix by dividing each element by the sum of all elements in the matrix.

It may be noted that only dominant swara of the frame is considered. As the matrix is prepared for the clip consisting of number of frames it can capture the distribution of the note sequence of a raga. The normalized matrix provides a sort of probability distribution of the occurrence of note pairs. The time complexity for computing the co-occurrence matrix is of $O(n)$ where n is the number of frames in the chromagram.

Figure 5 shows a sample co-occurrence matrices for raga *Marwa* and raga *Bageshree*. Darker a matrix element, more frequent is the corresponding note pair. Moreover, the effect of *Arohi* and *Avrohi* will also be embedded in the upper and lower triangle of the matrix. Thus, the matrix summarizes the note sequence of the raga along with its ascent and descent pattern.

4.2 Classification

In this work, we have used multi class Support Vector Machine (SVM) as classifier [3]. Essentially it is binary classifier that handles multi classes through a series of one versus

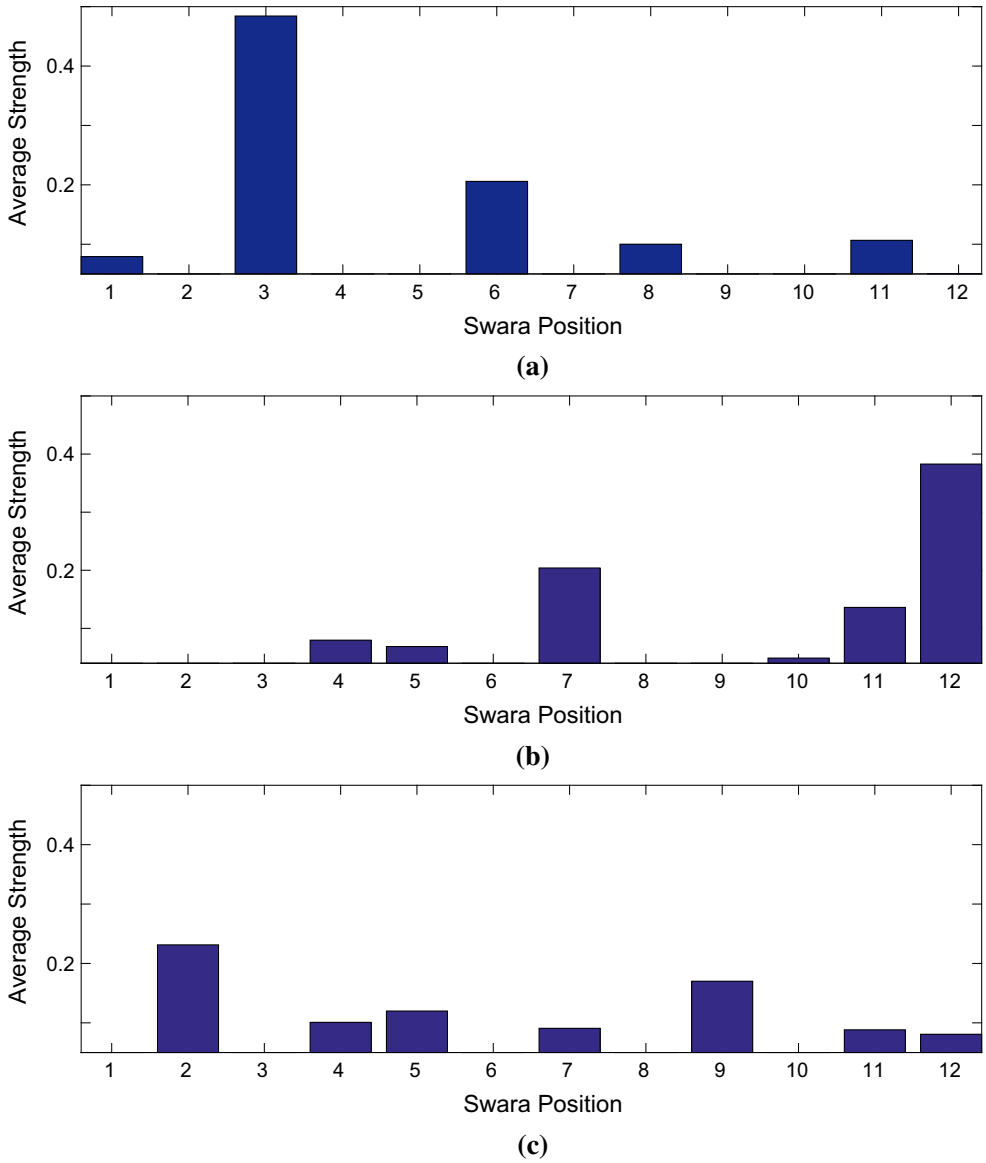


Fig. 4 Strength Distribution of Swaras for different ragas: **a** Raga Malkauns Jor (sitar)—Audhav jaati (five swaras), **b** Raga Bahar—Shadav jaati (six swaras) and **c** Raga Yaman—Sampurna jaati (seven swaras)

rest classification. In SVM, the goal is to learn the mapping: $X \rightarrow Y$ where $x \in X$ is descriptor or feature vector and $y \in Y$ is a class label (i.e. raga in our case). It tries to find the maximum-margin hyperplane separating two different class y_i and y_j. It is a complex optimization problem. Sequential Minimal Optimization (SMO) [16] is an iterative algorithm that solves such optimization problem. We have used SMO for training the SVM and it makes the process faster.

5 Experimental results

In order to carry out the experiment, we have prepared a dataset that reflects wide variety. Both, the vocal and instrumental based raga performances are included in the collection. The audio clips correspond to twenty four different ragas of Hindusthani classical music covering the performance of more than fifty renowned instrumental artists and more than twenty five great vocal artists. Table 4 provides a brief description of the dataset. All the recordings are sampled at 22,050 Hz and mono channeled.

Instead of working with the complete recording, we have tried to identify the raga based on a small part of it. We have considered clips of 45 seconds duration. As discussed in Sect. 4, each clip is represented by a 168-dimensional feature vector. It includes occurrence histogram of dominant swaras, strength distribution of swaras and co-occurrence of note pairs. First two categories of the features are global and they

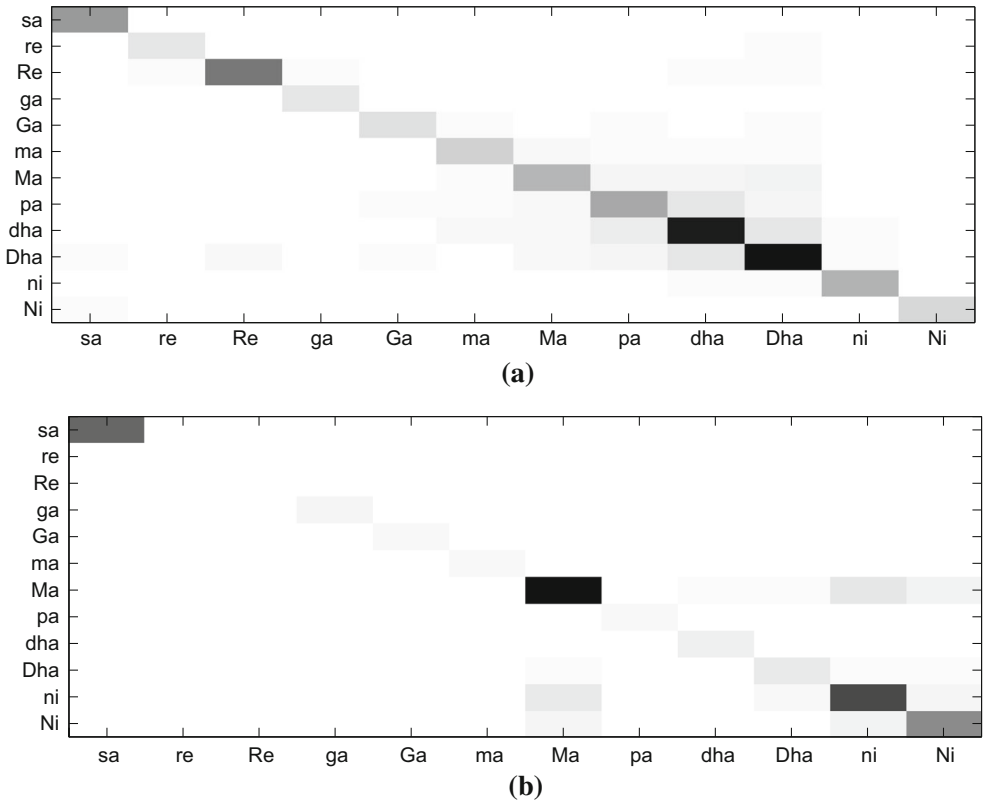


Fig. 5 Co-occurrence matrix of Swaras for different ragas: **a** Raga Marwa and **b** Raga Bageshree

Table 4 Description of the dataset

Type	The dataset contains both vocal and instrumental based ragas
Collection	The dataset contains 1648 raga clips. The duration of each clip is 45 s. Among them, 1190 raga clips from instrumental and 458 raga clips from vocal performances
Ragas	Bageshri, Bahar, Bhairabi, Bhairav, Bibhas, Bihag, Desh, Durga, Hamer, Jaunpuri, Jog, Kafi, Kalyani (yaman), Kanada, Kedar, Khamaj, Kirwani, Lalit, Malhar, Malkauns, Marwa, Purvi (Purvagauda), Sarang, Todi

do not capture temporal information. Co-occurrence of note pairs incorporates temporal aspects to some extent. Finally, SVM classifier is used for classification. The proposed methodology is tested separately on vocal and instrumental dataset. Confusion matrices for the vocal and instrumental collections are shown in Tables 5 and 6 respectively. Classification accuracy of the proposed system based on ten fold cross validation is summarized in Table 7.

We have compared the performance of our system with two other systems [8,12] on the same dataset. Koduri et al. [8] relied on pitch based features. First of all pitch contour is determined following the methodology proposed by Salamon et al. [20]. Stable regions in the pitch contours are then extracted and based on the same twelve dimensional (each corresponds to a swara) octave folded pitch histogram is formed. The $p_{instance}$ and $p_{duration}$ are each 12-dimensional feature vectors. These are weighted pitch histogram. Each bin in the histogram is weighted by the number of instances of the swara in the clip to obtain $p_{instance}$. For $p_{duration}$ weight is the duration of corresponding swara in the clip. For recognition, modified KL divergence is used as the distance measure and k-NN is used as classifier. Vijay et al. [12] in their work have considered pitch based $p_{instance}$ and $p_{duration}$ as features and KL-divergence as corresponding distance measure. Additionally n-gram distribution of notes is considered. To compare such distributions radial basis function kernel is used. Finally, SVM framework that combines kernel over two sets of features and corresponding distance measures has been used for raga identification. We have implemented both the systems and carried out the experiments on our vocal and instrumental dataset. The performance of all the systems is presented in Table 8. It is observed that performance of the proposed methodology is superior.

Table 5 Confusion matrix for vocal collection

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Raga
1	75.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	3.7	0	0	0	5.6	1.8	Kanada
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Bibhas
3	0	0	78.1	0	0	0	0	0	0	0	0	0	0	0	0	12.5	0	0	0	0	0	9.4	0	Bihag
4	11.1	0	5.6	72.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11.1	0	Sarang
5	0	0	0	0	83.3	0	0	0	0	0	0	0	0	0	0	16.7	0	0	0	0	0	0	0	Desh
6	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	Durga
7	0	0	20	0	0	0	60	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	Hamer
8	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	Kafi
9	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaunpuri
10	25	0	0	0	0	0	0	0	0	50	0	0	0	0	0	12.5	0	0	0	0	0	12.5	0	Jog
11	2.9	0	0	0	0	0	0	0	0	0	82.4	0	0	0	0	14.7	0	0	0	0	0	0	0	Bageshri
12	27.7	0	5.6	0	0	0	0	0	0	0	0	61.1	0	0	0	5.6	0	0	0	0	0	0	0	Kalyani
13	10	0	0	15	0	0	0	0	0	0	0	0	35	0	0	30	0	0	0	0	0	10	0	Kedar
14	12.5	0	0	0	0	0	0	0	0	0	0	0	0	50	0	25	0	0	0	0	0	0	12.5	Khamaj
15	14.3	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	14.3	0	21.4	0	Lalit
16	22	0	0	0	0	0	0	0	0	0	0	0	2	0	0	66	0	0	0	0	0	10	0	Malhar
17	10	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0	0	10	0	Malkauns
18	9.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	81.8	0	0	0	9.1	0	Marwa
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62.5	37.5	0	0	0	Purvi
20	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	5	0	0	0	70	0	15	0	Todi
21	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	10	50	0	0	Bahar
22	6	0	3	0	0	0	0	0	0	0	0	0	0	0	0	6	6	0	0	0	0	77.3	1.6	Bhairav
23	9.1	0	0	9.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.1	72.7	Bhairabi

Table 6 Confusion matrix for instrumental collection

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Raga
1	94.7	0	0	0	0	0	0	0	0	0	0	0	0	2.1	0	0	1.1	0	0	2.1	0	0	0	Kanada
2	0	89.6	0	0	2.6	2.6	0	0	0	0	2.6	0	0	0	0	2.6	0	0	0	0	0	0	0	Bihag
3	0	0	90.4	0	0	0	0	0	0	0	3.9	0	3.9	0	0	1.8	0	0	0	0	0	0	0	Sarang
4	0	0	0	79.2	0	0	0	0	0	0	0	0	16.6	0	0	4.2	0	0	0	0	0	0	0	Desh
5	0	0	0	0	93.8	0	0	0	0	0	3.1	0	0	0	0	0	0	0	0	3.1	0	0	0	Durga
6	0	0	0	0	0	71.9	0	0	0	0	0	0	6.3	0	0	0	0	0	0	0	18.8	3	0	Hamer
7	0	0	0	0	0	0	84.9	0	0	4.3	0	0	4.3	0	0	0	0	0	0	2.2	0	0	4.3	Kafi
8	0	0	0	0	75	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaunpu.
9	0	0	1.9	0	0	0	1.9	0	86.5	0	0	0	3.9	0	0	3.9	0	0	0	1.9	0	0	0	Jog
10	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	4.8	0	0	0	1.2	0	0	0	Bageshr
11	2.9	0	0	0	0	0	1.5	0	0	2.9	75	0	1.5	0	0	2.9	0	0	0	1.5	0	0	11.8	Kalyani
12	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	7	0	0	0	0	0	7	0	Kedar
13	5	0	1.3	0	0	1.3	0	0	0	0	2.5	0	82.5	0	0	3.7	0	0	0	3.7	0	0	0	Khamaj
14	0	0	0	0	0	0	4.3	0	0	0	2.2	0	0	91	0	2.2	0	0	0	0	0	0	0	Kirwani
15	6.3	0	0	0	0	0	0	0	0	6.3	0	0	0	0	75	0	0	9.3	0	3.1	0	0	0	Lalit
16	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	91	0	1	0	0	1	3	0	Malhar
17	0	0	0	0	0	0	0	0	2.2	0	0	0	2.2	0	0	0	91	0	0	4.3	0	0	0	Malka.
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	Marwa
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	67	0	0	0	Purvi
20	5.8	0	0	0	0	0	0	0	0	0	0	0	1.2	0	0	1.2	0	0	0	89.5	0	0	2.3	Todi
21	0	0	0	0	0	18.8	0	0	0	0	9.4	0	0	0	0	6.2	0	3.1	0	0	62.5	0	0	Bahar
22	6.8	0	2.7	0	0	0	0	0	2.7	0	0	0	4.1	0	0	2.7	0	3	0	0	0	77	0	Bhairav
23	2.6	0	0	0	0	0	0	3.8	0	0	20.5	0	6.4	0	0	10.3	0	0	0	0	0	0	56.4	Bhaira.

Table 7 Classification accuracy (in %) of proposed system

Raga	Instrumental dataset	Vocal dataset
Bahar	62.50	50.00
Bhairav	77.03	77.27
Bhairabi	56.41	72.73
Bageshri	94.05	82.35
Bibhas	–	100.00
Bihag	89.47	78.13
Desh	79.17	83.33
Durgaa	93.75	75.00
Hamer	71.86	60.00
Juanpuri	25.00	100.00
Jog	86.54	50.00
Kafi	84.78	50.00
Kanada	94.68	75.92
Kedar	85.71	35.00
Khamaj	82.50	50.00
Kirwani	91.30	–
Kalyani (Yaman)	75.00	61.11
Lalit	75.00	50.00
Malhar	91.03	66.00
Malkaunsh	91.31	75.00
Marwa	100.00	81.82
Purvi	33.33	62.5
Sarang	90.38	72.22
Todi	89.54	70.00
Overall	84.29	70.52

Table 8 Performance comparison: classification accuracy (in %) of different systems

System	Instrumental dataset	Vocal dataset
Koduri et al. [8] using $p_{instance}$	54.96	49.78
Koduri et al. [8] using $p_{duration}$	58.91	50.22
Vijay et al. [12]	62.99	55.20
Proposed system	84.29	70.52

6 Conclusion

We have presented a simple but novel scheme to identify the ragas in Hindusthani classical music. Features are designed to reflect the compositional properties of a raga. First of all pitch based swara (note) profile is formed. It is utilized to generate occurrence histogram of dominant swaras and strength distribution of the swaras. Such descriptors have strong correlation with the properties that categorizes the role of swaras in the ragas. Note sequence is an important property and note co-occurrence matrix captures this temporal aspect to an extent. Thus, features are devised following an analytic approach adopted in manual recognition. Training of the SVM classifier emulates the intuitive approach of a human being where identification relies on prior knowledge gathered by listening the music. Proposed methodology thus combines analytic and intuitive approach. Experiment has been carried out with a diversified dataset and compared the performance with other systems. It is observed that proposed system works better. In future, domain knowledge may be utilized to analyze the classification errors and to take measures for further improvement.

Acknowledgements Rajib Sarkar thanks University Grants Commission (UGC), India for his fellowship.

References

1. Chakravorty, J., Mukherjee, B., Datta, A.: Some studies on machine recognition of ragas in indian classical music. J. Acoust. Soc. India **XVII**(3 & 4), 1–4 (1989)
2. Chordia, P., Rae, A.: Raag recognition using pitch- class and pitch-class dyad distributions. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 431–436 (2007)
3. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
4. Das, D., Choudhury, M.: Finite state models for generation of Hindustani classical music. In: Proceedings of International Symposium on Frontiers of Research in Speech and Music, pp. 59–64 (2005)
5. Dighe, P., Agrawal, P., Karnick, H., Thota, S., Raj, B.: Scale independent raga identification using chromagram patterns and swara based features. In: Proceedings of International Conference on Multimedia and Expo Workshops, pp. 1–4 (2013)
6. Dutta, S., Murthy, H.A.: Raga verification in carnatic music using longest common segment set. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 605–611 (2015)
7. Katz, J.: Dhrupad: tradition and performance in Indian music. Ethnomusicol. Forum **14**(1), 113–115 (2005)
8. Koduri, G.K., Gulati, S., Rao, P., Serra, X.: Rga recognition based on pitch distribution methods. J. New Music Res. **41**(4), 337–350 (2012)
9. Koduri Gopala-Krishna, S.G., Rao, P.: A survey of raaga recognition techniques and improvements to the state-of-the-art. In: Sound and Music Computing, pp. 33–40 (2011)
10. Krishnaswamy, A.: Melodic atoms for transcribing carnatic music. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 345–348 (2004)
11. Kumar, K.P., Rao, M.S.: Raaga identification using clustering algorithm. In: International Conference on Electrical, Electronics, and Optimization Techniques, pp. 2077–2081 (2016)
12. Kumar, V., Pandya, H., Jawahar, C.V.: Identifying ragas in Indian music. In: International Conference on Pattern Recognition, pp. 767–772 (2014)
13. Müller, M., Ewert, S.: Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In: Proceedings of International Conference on Music Information Retrieval, pp. 215–220 (2011)
14. Müller, M., Kurth, F., Clausen, M.: Audio matching via chroma-based statistical features. In: International Conference on Music Information Retrieval, pp. 288–295 (2005)

15. Pandey, G., Mishra, C., Ipe, P.: Tansen: a system for automatic raga identification. In: Proceedings of International Conference on Artificial Intelligence, pp. 1350–1363 (2003)
16. Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MST-TR-98-14. Microsoft Research (1998)
17. Pranay, D., Harish, K., Bhiksha, R.: Raga verification in carnatic music using longest common segment set. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 605–611 (2013)
18. Rao, P., Ross, J.C., Ganguli, K.K., Pandit, V., Ishwar, V., Bellur, A., Murthy, H.: Classification of melodic motifs in raga music with time-series matching. J. New Music Res. **43**(1), 115–131 (2014)
19. Ross, J.C., Rao, P.: Detection of raga-characteristic phrases from Hindustani classical music audio. In: Proceedings of CompMusic Workshop, pp. 133–138 (2012)
20. Salamon, J., Gomez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans. Audio Speech Lang. Process. **20**(6), 1759–1770 (2012)
21. Sharma, M.: Tradition of Hindustani Music. A.P.H. Publishing Corporation, New Delhi (2006)
22. Shetty, S., Achary, K.K.: Raga mining of Indian music by extracting arohana-avarohana pattern. Int. J. Recent Trends Eng. **1**(1), 362–366 (2009)
23. Singh, J.: Interview with Pandit Jasraj. https://www.youtube.com/watch?v=VPshheRL69M (2016). (You Tube; Accessed 10 Sept 2016)
24. Sinith, M., Rajeev, K.: Hidden Markov model based recognition of musical pattern in south Indian classical music. In: Proceedings of International Conference on Signal and Image Processing, Hubli, India (2006)
25. Sridhar, R., Geetha, T.V.: Raga identification of carnatic music for music information retrieval. Int. J. Recent Trends Eng. **1**(1), 571–574 (2009)
26. Viswanathan, T., Allen, M.: Music in South India: The Karāak Concert Tradition and Beyond: Experiencing Music, Expressing Culture. Global Music Series. Oxford University Press, Oxford (2004)
27. Wade, B.: Chz in khyl: the traditional composition in the improvised performance. Ethnomusicology **17**(3), 443–459 (1973)