Multi-Aspect Music Classification using Low Level Features

Bikram Rana
Registration No. - 133676 of 2015-2016

Examination Roll No. - MCA186014

Supervisor: **Prof. Sanjoy Kumar Saha**

Department of Computer Science & Engineering
Jadavpur University
Kolkata - 700032

This project report is submitted for the partial fulfillment of the degree of
Master of Computer Application

May 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Bikram Rana
Registration No. - 133676 of 2015-2016
Examination Roll No. - MCA186014

Jadavpur University

Department of Computer Science

Jadavpur, Kolkata – 700032

CERTIFICATE

This is certify that the project entitled **"Multi-aspect Music Classification using Low Level Features",** submitted by **Bikram Rana** is a record of bona-fide work carried out by him, in the partial fulfillment of the requirement for the award of Degree of **Master of Computer Application** of the Department of Computer Science and Engineering, Jadavpur University.This work is done during the academic year 2017-2018, under my guidance.

(**Prof. Sanjoy Kumar Saha**)

Project Supervisor
Computer Science and Engineering
Jadavpur University, Kolkata - 700032

Countersigned:

(Dr. Ujjwal Maulik)

Head of Department
Computer Science and Engineering
Jadavpur University, Kolkata - 700032

(Prof. Chirnajib Bhattacharjee)

Dean
Faculty Of Engineering and Technlogy
Jadavpur University, Kolkata - 700032

Jadavpur University

Department of Computer Science

Jadavpur, Kolkata – 700032

CERTIFICATE OF APPROVAL

The foregoing project report entitled "**Multi-aspect Music Classification using Low Level Features**" is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood by this approval the undersigned do not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn therein but approve the report only for the purpose for which it has been submitted.

Internal Examiner

External Examiner

Acknowledgements

I would first like to thank my guide, Prof. Sanjoy Kumar Saha of Computer Science and Engineering department,Jadavpur University for consistently helping me through my entire project work and guiding me every time I needed. I would also like to thank Rajib Sarkar, without whose participation and input I could not have successfully completed the work. Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Automated classification of music signal is an active area of research. Currently classification is performed manually. Automatic classification can assist or replace the human user in this process and would be a valuable addition to music information retrieval systems. In addition, it provides a framework for developing and evaluating features for other type of content-based analysis of musical signals. It can act as the fundamental step for various applications like archival, indexing and retrieval of music data. In this work music is classified based on fundamental aspects of music like singer, emotion and genre. Low level features are extracted as descriptors. SVM with SMO is used for classification purpose. The classification result for individual as well as the combination of all three aspects are satisfactory.

Table of contents

1 Introduction 1

2 Past work 4

3 Methodology 8

3.1 Feature Extraction : Emotion 8

3.1.1 MFCC based feature 8

3.1.2 Spectral Flux (SF): 9

3.1.3 Spectral Rolloff (SR): 10

3.1.4 Spectral Centroid (SC): 10

3.1.5 Spectral Spread (SSP): 10

3.1.6 Spectral Slope (SSL): 10

3.1.7 Spectral Flatness Measure (SFM): 11

3.1.8 Spectral Crest Factor (SCF): 11

3.1.9 Spectral Kurtosis (SK): 11

3.1.10 Spectral Skewness: 12

3.2 Feature Extraction : Singer 12

3.2.1 Extraction of Vocal Component 12

3.2.2 Spectrogram based vocal-print extraction 13

3.2.3 MFCC based feature extraction 13

3.3 Feature Extraction : Genre 14

3.3.1 Pitch based feature 14

3.4 Classifier 15

3.4.1 Support Vector Machines (SVM): 15

4 Results 16

4.1 Dataset 16

4.2 Result and Discussion 16

4.2.1 Genre Classification 17

4.2.2 Singer Classification 17

4.2.3 Emotion classification 18

4.2.4 Classification of Genre, Singer and Emotion together 18

5 Conclusion 20

References 21

Chapter 1

Introduction

Music is created/performed with a vast range of instruments and vocal techniques ranging from singing to rapping; there are solely instrumental pieces, solely vocal pieces (such as songs without instrumental accompaniment) and pieces that combine singing and instruments. In many cultures, music is an important part of people's way of life, as it plays a key role in religious rituals, ceremonies, social activities and cultural activities ranging from amateur karaoke singing to playing in an amateur funk band or singing in a community choir. People may make music as a hobby, or work as a professional musician or singer. Music is broadly defined, but for the sake of our thesis we define music as an art-form that combines both vocal and instrumental sound to produce something that evokes emotion. When we talk about music, we are mostly interested in three things; The artist/s, the genre and the emotion/mood associated with the music.

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications. Those involved in MIR may have a background in musicology, psychoacoustics, psychology, academic music study, signal processing, informatics, machine learning, computational intelligence or some combination of these. Analysis can often require some summarizing, this is achieved by feature extraction, especially when the audio content itself is analyzed and machine learning is to be applied. The purpose is to reduce the sheer quantity of data down to a manageable set of values so that learning can be performed within a reasonable time-frame. One common feature extracted is the Mel-Frequency Cepstral Coefficient (MFCC) which is a measure of the timbre of a piece of music. Other features may be employed to represent the key, chords, harmonies, melody, main pitch, beats per minute or rhythm in the piece. There are a number of available audio feature extraction tools[

Musical **Genre** does not have a strict definitions or boundaries and can be highly subjective. Simply put genre is a label created by humans to categorize vast collections of music.A music genre or sub-genre may also be defined by the musical techniques, the style, the cultural context, and the content and spirit of the themes. Geographical origin is sometimes used to identify a music genre, though a single geographical category will often include a wide variety of sub-genres. With increasing numbers of music files on the web, automatic music information retrieval has gained more importance as a way to structure and organize them.

One another characteristics that usually pops up when taking about music is the **emotion/mood** that is evoked when we hear a musical piece. like with the genre, **mood/emotion** is highly subjective and maybe even more so. The ability to perceive emotion in music is said to develop early in childhood, and improve significantly throughout development. The capacity to perceive emotion in music is also subject to cultural influences, and both similarities and differences in emotion perception have been observed in cross-cultural studies A music that once gave us a happy feeling may seem depressing depending upon the situation we're in. Generally more than one emotion is associated with a piece of music. Musical structure of a music plays a very important role in how we perceive song. For example an uplifting/happy music may differ in musical structure from one that is depressing/sad. We can use this knowledge to classify music based on emotion/mood.

Manual classification by music experts is one of the way of organizing music and is by far also the more reliable. Automatic classification can potentially automate this process and more importantly provide a basis of developing new features for musical information retrieval.

Chapter 2

Past work

The work of Tzanetakis et al. [1] is considered as the pioneering work on automatic music genre classification. State of the art of music genre classification [2] shows that automatic music genre classification demands more attraction and still is an active research area. Panagakis et al. [3] proposed feature based music genre classification method. They have extracted MFCCs, chromagram and wavelet transform based summarization of spectrogram (Auditory Cortical Representations) features. They proposed a joint sparse low-rank classification (JSLRR) based classifier. The JSLRR based approach is an alternative of sparse and low-rank representation to reduce noise and specify the subspace where data defiled by outliers. Huang et al. [4] proposed a self-adaptive harmony search (SAHS) method to choose best feature descriptors. They have extracted intensity, pitch, timbre, tonality, and rhythm based features from the music excerpts. The SVM classifier used for classification purpose. Markov and Matsui [5] proposed Gaussian Processes (GP), a Bayesian nonparametric model for music genre classification. They have extracted MFCCs, line spectral pairs (LSP),timbre, spectral crest factor (SCF), spectral flatness measure (SFM) and chromagram features. Finally they have shown that the performance of the Gaussian Processes classification model is better compared to SVM. Schindler and Rauber [6] proposed an audio and visual based approach to classify music genre. They extracted music related features such as Statistical Spectrum Descriptors (SSD), Rhythm Patterns, Rhythm Histograms, MFCCs and Chroma. Color statistics and emotion related features taken from music videos as visual information. In their work, Nanni et al. [7] extracted a set of texture descriptors from spectrogram in order to music genre classification. To derive the features, three different representations (mel-scale divided, linear divided and whole) of the spectrogram is used. Later, in [8] they have combined visual features with the textures to improve the performance. They have used SVM for classification.

It is not always effective to to label a music excerpt by exactly one emotional tag and it may have multiple emotional aspects. One way of representing the emotional aspect of music clips is to place them in a two dimensional plane, as proposed by Thayer [9] and Russell [10]. The two dimensions in the plot are pleasure/valence and arousal. Russell's work is considered as the pioneer in Music Emotion Recognition field to describe human emotions. Thayer adopted Russell's emotion model. This two-dimensional emotion plane can be divided into quadrants or octant denoting four or eight emotional state. According to Thayer's emotion plane, emotion of a music can be depicted by a two dimensional model where the X and Y axes are valence and arousal respectively. So, regression model for both valence and arousal can be trained to find the position of the song in the plane which gives us an impression about the emotional aspect of the music. Gomez et al. [11] proposed feature based multi-label music emotion detection system. A set of features are used as descriptor. The descriptor includes mean and standard deviation of spectral centroid, spectral roll-off and mel-frequency cepstral coefficients (MFCCs). The classification algorithm k-Nearest Neighbors (kNN) is used to classify song clips according to their emotional labels. Chen et al. [12] proposed a regression based method. They have extracted MFCC coefficients, tonal intensity value, linear predictor coefficients of the spectral envelope, spectral flux and spectral shape descriptors like spectral centroid, spectral spread, spectral skewness, and spectral kurtosis. The valence and arousal values of each music clip is regressed with ground truth using Gaussian mixture model (GMM). Koch et al. [13] proposed a method for recommendation of on-line videos. The low level audio features are extracted to determine the emotion. MIRtoolbox [14] is used to extract the low level audio features and for classification support vector machines (SVM) is used. Zhang et al. [15] proposed a feature based emotion detection approach where features like root mean square (RMS) energy, MFCC's, zero crossing rate (ZCR), fundamental frequency (f_0) are considered along-with voicing probability and statistical features. They have used Random Forest for classification. Media-Eval competition [16] considers emotion in music as a task. Participants deal with the prediction of static and dynamic emotion [17].

Various properties of music signal and low level features are discussed in [18] which can be utilized in singer identification. One early attempt [19] borrowed the idea of *eigenface* used in face recognition and proposed the idea of *eigenvoice* in the context of speaker recognition. But singing voice and speech voice differ significantly [20]. Hence speaker characterization technique can not be applied directly in characterizing singing voice. Tsai and Lee [20] proposed two alternative solutions. In one approach speech model undergoes an adaptation process to deal with singing voice. The second approach presented a transformation to be applied on singing voice and converted into speech voice. Finally, speech driven model is used. Gaussian mixture model (GMM) was used for classification. Motivated by the work of Cano et al. [21], a compact representation of music track signature is proposed in [22]. Normalized Mel frequency cepstral co-efficients (MFCC) are computed the frames in a track. Such co-efficients vectors are grouped following equal sized clustering. Cluster centres are concatenated to form the track signature. Based on the signatures of number of tracks the singer characteristics are designed.

In recent years number of works [23–26] have put emphasis on the extraction of singing voice segment from the song. This is in order to reduce the impact of non-vocal component in the process. Sofianos et al. [23] used Azimuth Discrimination and Resynthesis (ADRess) and Independent Component Analysis (ICA) to develop a hybrid method for singing voice separation.

Tsai and Lin [24] studied the spectrogram of solo singing voice and the same for solo singing voice mixed with accompaniment. By observing the spectrograms, a relationship between the solo singing voice and combined signal was established. The knowledge gathered is utilized in removing the effect of non-voice component. But, it suffers from the fact that obtaining the solo singing voice signals for number of singers is difficult. Huang et al. [25] relied on the observation that non-vocal musical segment is more repetitive than the vocal part. Finally, a methodology based on robust principal component analysis is presented for separating the singing voice segments. Yang [26] proposed an algorithm called multiple low-rank representation (MLRR) to separate the singing voice from background music. It exploits the fact that in a magnitude spectrum of a song signal the non-vocal part corresponds to low-rank component and the singing voice is represented by the sparse component. As the vocal part often is mapped to the low-rank component the separation becomes difficult. MLRR tries to address this problem.

Su and Yang [27] proposed a system based on voice timbre. As preprocessing, singing voice part is extracted and then instead of spectral magnitude, phase information is used for modeling the voice timbre. To be specific, sparse feature from the negative derivative of phase (with respect to frequency) is used. SVM is used for classification. Hu and Liu [28] proposed a system where vocal segment is extracted based on computational auditory scene analysis (CASA). Finally, gammatone frequency cepstral co-efficient (GFCC) are computed by analyzing the pitch content. Tsai et al. [29] worked in compressed domain and proposed a methodology for computing MFCC directly from MP3 file. GMM is used for classification.

The survey indicates that although variety of approaches have been tried by the researchers, still the problem is open. Also, researchers proposed different features and methods for the classification of single music aspects.

Chapter 3

Methodology

The proposed methodology focused on multi-aspect classification of music signal. We have considered three different music aspects - *Singer, genre* and *Emotion*. First of all, music is classified using the three aspects independently. Then, all the three aspects are combined together to see the performance of the proposed system. Suitable features has been extracted from the music excerpts based different aspects. These are detailed in the following sections.

3.1 Feature Extraction : Emotion

The following MFCC and spectral features are extracted and used for emotion classification.

3.1.1 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. The steps for computing MFCC are elaborated in [30]. 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

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

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 co-efficients. In our work, frame size is taken as 256. First thirteen co-efficients of all the frames are considered. The frame level co-efficients may be concatenated to represent the signal characteristics in detail. But the dimension becomes prohibitive. we have considered mean and standard deviation of the 13 co-efficients over the frames.

3.1.2 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:

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

It captures changes in the power of spectral components over the successive frames.

3.1.3 Spectral Rolloff (SR):

It is defined as the q^{th} percentile of the power spectral distribution [31]. 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.

3.1.4 Spectral Centroid (SC):

The Spectral Centroid of an audio signal represents the center of gravity of the spectral power. SC is a 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.

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

3.1.5 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

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

3.1.6 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

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

where, $\mu_{\mathscr{S}}$ 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.

3.1.7 Spectral Flatness Measure (SFM):

It is the proportion of geometric mean and arithmetic mean of a magnitude spectrum [32, 18], as shown below,

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

For uniform (flat) distribution of power spectral component it provides higher value.

3.1.8 Spectral Crest Factor (SCF):

It is the measurement of the quality of a acoustic signal [18]. It is computed as the proportion of highest of the power spectrum with total power spectrum.

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

3.1.9 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. SK indicates how Gaussian the magnitude spectrum distribution looks like. It is calculated as

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

where, $\mu_{\mathscr{S}}$ is the mean of spectral magnitude and $\sigma_{\mathscr{S}}$ is the standard deviation of the spectrogram.

3.1.10 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

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

Here also, mean and standard deviation of individual frame level features are considered at the clip level.

3.2 Feature Extraction : Singer

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.

3.2.1 Extraction of Vocal Component

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.

It is observed that a vocal segment has more energy compared to a non-vocal segment. The non-vocal segments can be removed based on the energy distribution. 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.

3.2.2 Spectrogram based vocal-print extraction

From the spectrogram we have extracted the vocal-print of a singer. To construct the spectrogram, the 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.

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 (1 Hz to 1500 Hz)is divided into number of bands of width 100Hz. Average spectral power in the bands are concatenated to form the 15-dimensional feature vector representing the vocal-print.

3.2.3 MFCC based feature extraction

The MFCC features are extracted from the vocal segments only. The procedure for extracting MFCC features already discussed in section 3.1.1.

3.3 Feature Extraction : Genre

For genre classification, we have considered pitch based features along with the feature set used for emotion classification (discussed in section 3.1).

3.3.1 Pitch based feature

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 [33], 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.

The steps for deriving the features are as follows.

- For each MIDI band, quantize the STMSP values as follows.
  - Calculate average (μ) and standard deviation (σ) of STMSP values in the band.
  - Quantization levels are taken as $\mu + k\sigma$ where k varies from -2 to $+2$ with step size 0.5.
- For each band prepare the normalized histogram of quantized STMSP values.
- Concatenate the histograms of all the bands.

For each pitch-band, STMSP values are quantized into ten bins. It has been observed that the contribution of the first three bins are marginal. Hence those are ignored and finally, 616 dimensional histogram is obtained. From this, 15 dimensional feature is selected applying Principal Component Analysis (PCA) and used as the descriptor.

3.4 Classifier

3.4.1 Support Vector Machines (SVM):

We have used Support Vector Machines for classification. SVMs [34] are widely used in classification, regression or novelty detection problem. SVMs are large-margin classifiers. Given training data containing f features with only two possible output labels, SVM finds a hyperplane in f-dimensional hyperplane that maximizes margin between two classes. This margin is calculated as the difference between the decision border line and the nearest input vectors. This binary classification strategy can be extended to solve multi-class classification problem with one against one approach. Sequential Minimal Optimization (SMO) [35] is used to train the SVM. The SVM training algorithm requires to solve a very large quadratic programming (QP) optimization problem. SMO divides this QP problem into a series of small sub-problems, which are then solved analytically. In this way SMO avoids the expensive QP optimization during the training of SVM.

We have used the implementation of the classifiers from the WEKA [36] framework.

Chapter 4

Results

4.1 Dataset

In order to carry out the experiment, we have prepared a dataset that reflects variety. The dataset consists of 202 music clips. The dataset broadly annotated into three different categories - *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* 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.

4.2 Result and Discussion

All the experiments are carried out using SVM classifier with SMO. 50% data is used for training and remaining 50% data used for testing. The results are represented below using a confusion matrix, which gives a visualization of the performance of our proposed system. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class.

At first we have tested the performance of the proposed system for the three basic categories - Singer, Emotion and Genre classification independently. Then, we have tested the performance of the proposed system for all combination of the three basic categories. The findings are reported in details in the following sub sections.

4.2.1 Genre Classification

For genre classification, out of 202 excerpts, 186 excerpts were correctly identified. The classification accuracy is 92.08%. The confusion matrix for Genre based classification is shown in Table 4.1.

Table 4.1 Confusion matrix for Genre based classification

	Folk	Tagore	Devotional	Classical
Folk	56	1	1	3
Tagore	6	30	0	0
Devotional	2	2	39	0
Classical	0	0	1	61

4.2.2 Singer Classification

For Singer classification, out of 202 excerpts, 198 excerpts were correctly identified. The classification accuracy is 98.02%. The confusion matrix for Genre based classification is shown in Table 4.2.

Table 4.2 Confusion matrix for Singer based classification

	Abbasuddin	Asha	Anita	Jasraj	Kishore	Anup
Abbasuddin	45	0	0	0	1	1
Asha	0	12	0	0	0	0
Anita	1	0	13	0	0	0
Jasraj	0	0	0	62	0	0
Kishore	0	0	0	0	24	0
Anup	0	0	0	0	1	42

4.2.3 Emotion classification

For Emotion classification, out of 202 excerpts, 150 excerpts were correctly identified. The classification accuracy is 74.26%. The confusion matrix for Genre based classification is shown in Table 4.3.

Table 4.3 Confusion matrix for Emotion based classification

	Joy	Peaceful	Romantic	Sadness
Joy	7	0	4	2
Peaceful	5	54	1	12
Romantic	0	4	10	5
Sadness	0	12	7	79

4.2.4 Classification of Genre, Singer and Emotion together

For all three aspects (genre,singer and emotion), 144 excerpts out of 202 were correctly identified, which results an accuracy of 71.2%. The result is shown in Table 4.4.

In the following S1-S6 represent six singers, E1-E4 represents four singers and G1-G6 represent four genres.

Table 4.4 Confusion matrix for classification of Genre, Singer and Emotion combined together

	G1S1E1	*G1S1E2*	*G1S1E3*	*G1S1E4*	*G2S2E1*	*G2S2E2*	*G2S2E3*	*G2S2E4*	*G1S3E4*	*G4S4E1*	*G4S4E2*	*G4S4E3*	*G4S4E4*	*G2S5E1*	*G2S5E2*	*G2S5E3*	*G2S5E4*	*G3S5E2*
G1S1E1	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1S1E2	0	2	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
G1S1E3	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G1S1E4	0	0	3	27	3	0	0	0	0	0	0	0	0	0	0	0	0	0
G2S2E1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
G2S2E2	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
G2S2E3	0	0	0	0	1	0	2	1	0	0	0	0	0	0	0	0	0	0
G2S2E4	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0
G1S3E4	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0
G4S4E1	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0
G4S4E2	0	0	0	0	0	0	0	0	0	0	12	0	7	0	0	0	0	0
G4S4E3	0	0	0	0	0	0	0	0	0	0	2	3	2	0	0	0	0	0
G4S4E4	0	0	0	0	0	0	0	0	0	1	6	6	19	0	0	0	0	0
G2S5E1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
G2S5E2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0
G2S5E3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	0
G2S5E4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	13	0
G3S6E2	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	37

Chapter 5

Conclusion

Despite the fuzzy nature of genre or emotion boundaries, classification can be performed automatically with results significantly better than chance, and performance comparable to human classification. Using the proposed features set classification of 92.08% based on genres, 98.02% based on singers and 74.02% based on emotions, has been achieved in a dataset consisting of four musical genres, six singers and four emotions respectively.The success of the proposed features for classification testifies to their potential as the basis for other types of automatic techniques for music signals such as similarity retrieval, segmentation and audio thumb-nailing which are based on extracting features to describe musical content.

Applications such as Audio fingerprinting, content-based querying and retrieval, music recommendation, song/artist popularity estimation, are already very popular field of research.

References

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

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

[3] 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.

[4] 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.

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

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

[7] Loris Nanni, Yandre Costa, and Sheryl Brahnam. Set of texture descriptors for music genre classification. In *22th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2014.

[8] 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.

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

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

[11] Lucía Martín Gómez and María Navarro Cáceres. Applying Data Mining for Sentiment Analysis in Music. In *Proc. PAAMS*, pages 198–205, 2017.

[12] 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.

[13] Christian Koch, Ganna Krupii, and David Hausheer. Proactive Caching of Music Videos Based on Audio Features, Mood, and Genre. In *Proc. MMSys*, pages 100–111, 2017.

[14] Olivier Lartillot and Petri Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *Proc. DAFx*, pages 237–244, 2007.

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

[16] Mohammad Soleymani, Anna Aljanaki, Yi-Hsuan Yang, Michael N Caro, Florian Eyben, Konstantin Markov, Björn W Schuller, Remco Veltkamp, Felix Weninger, and Frans Wiering. Emotional analysis of music: A comparison of methods. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1161–1164. ACM, 2014.

[17] Felix Weninger, Florian Eyben, and Björn W Schuller. The TUM Approach to the MediaEval Music Emotion Task Using Generic Affective Audio Features. In *MediaEval*, 2013.

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

[19] R Kuhn, J Junqua, P Nguyen, and N Niedzielski. Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Trans. on Speech and Audio Processing*, 8(6):695–707, nov 2000.

[20] Wei-Ho Tsai and Hsin-Chieh Lee. Singer Identification Based on Spoken Data in Voice Characterization. *{IEEE} Trans. on Audio, Speech & Language Processing*, 20(8):2291–2300, 2012.

[21] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A Review of Audio Fingerprinting. *VLSI Signal Process. Syst.*, 41(3):271–284, nov 2005.

[22] 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.

[23] Stratis Sofianos, Aladdin M Ariyaeeinia, and Richard Polfreman. Towards effective singing voice extraction from stereophonic recordings. In *ICASSP*, pages 233–236. IEEE, 2010.

[24] Wei-Ho Tsai and Hao-Ping Lin. Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification. *IEEE Trans. on Audio, Speech & Language Processing*, 19(5):1196–1205, 2011.

[25] P.-S. Huang, S D Chen, P Smaragdis, and M Hasegawa-Johnson. Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, 2012.

[26] Yi-Hsuan Yang. Low-rank representation of both singing voice and music accompaniment via learned dictionaries. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, nov 2013.

[27] Li Su and Yi-Hsuan Yang. Sparse modeling for artist identification: exploiting phase information and vocal separation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, nov 2013.

[28] 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.

[29] 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.

[30] Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR*, 2000.

[31] 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.

[32] 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.

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

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

[35] John Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical report, apr 1998.

[36] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, nov 2009.