CONTENT BASED MUSIC INFORMATION RETRIEVAL (CB-MIR) AND ITS APPLICATIONS TOWARDS MUSIC RECOMMENDER SYSTEM

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

Y V SRINIVASA MURTHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA (NITK),

SURATHKAL, MANGALORE,

KARNATAKA - 575 025, INDIA.

DECEMBER 2018

Dedicated to my
Grand Father (Late Sri Jaya Simhudu), Grand Mother, Parents, Teachers, Research Supervisor, In-Laws, Mavayya (Uncle), Wife, Daughter, Brother, My Friend (late) P. Srinivasu and All my dear friends.

DECLARATION

by the Ph.D. Research Scholar

I hereby **declare** that the Research Thesis entitled **Content-based Music Information Retrieval (CB-MIR) and its Applications Towards Music Recommender System** which is being submitted to the **National Institute of Technology Karnataka (NITK), Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Computer Science and Engineering** is a **bonafide report of the research work carried out by me**. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

(**135067 CS13F05, Y V Srinivasa Murthy**)
(Register Number, Name & Signature of Research Scholar)
Department of Computer Science and Engineering

Place: NITK, Surathkal.
Date: December 12, 2018

CERTIFICATE

This is to *certify* that the Research Thesis entitled **Content-based Music Information Retrieval (CB-MIR) and its Applications Towards Music Recommender System** submitted by **Y V Srinivasa Murthy**, (Register Number: **135067 CS13F05**) as the record of the research work carried out by him, is *accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

Dr. Shashidhar G. Koolagudi
Research Supervisor
(Name and Signature with Date and Seal)

Chairman - DRPC
(Name and Signature with Date and Seal)

Acknowledgements

Firstly, I would like to express my sincere gratitude to my guide, **Dr. Shashidhar G. Koolagudi (Associate Professor, Department of CSE)**, for his continuous guidance and motivation, because of which I have been able to solve any issues that I face. It is due to his influence that I have gained patience and an insight on simple living.

Besides my guide, I would like to thank **Prof. U. Sripathi Acharya (Professor, ECE (RPAC Member))** and **Dr. Mohit P. Tahiliani (Assistant Professor, CSE (RPAC Member))** who have been a continuous source of encouragement to me and have been always ready to assist me with constructive suggestions.

My sincere thanks also goes to **Prof. Santhi Thilagam (HOD, CSE.)** and **Dr. Alwyn Roshan Pais (Chairman, DRPC)** who encouraged me all the time with positive statements.

I can't miss mentioning the faculty of CSE Department, **Dr. K.C. Chandrasekaran**, **Dr. B. Annappa**, **Dr. Manu Basavaraj**, **Dr. Jeny Rajan**, **Dr. Basavaraj Talawar** and **Dr. B.R. Chandavarkar** who have always been ready to support me when I needed it.

I can't forget the one who welcomed me to NITK with a smile, **Late. Prof. K. C. Shet**.

A special thanks to **Dr. Sujatha D. Achar, MACS Department**, who blessed me all the time with positive wishes.

This would be incomplete without mentioning my sister, **Smt. Saumya Hegde (Assistant Professor, CSE Department)**, who supported me in all the aspects of my Ph.D. life.

Heartfelt thanks **Mrs. Jayashree Koolagudi** for enquiring my thesis status all the time and caring all the time.

I would also like to thank the technical and supporting team **Smt. Yashwanthi, Smt. Seema Shivaram, Vairavanathan, Kamath, Mohini, Vikranth, Yash-**

wanth, Ravi and **Arun** for providing all the facilities to do my research in a smooth way. Without such high-end infrastructure and continuous power supply, it would have not been possible for me to finish my research.

Heartfelt thanks to the speech group of NITK, **SIMPLE (Speech, Image, and Music Processing Learning Environment)**, the members including **Pravin B. Ramteke**, who always helped in all the aspects, **Manjunath Mulimani**, **Nagaratna B. Chittaragi** and **Fathima Afroz** for their help in explaining unknown concepts, coding, adjusting my duties during presentations, and a lot more assistance they have rendered in many personal matters too.

Thanks to the faculty and research scholars of CSE, IT & ECE departments who made me happy all the time, especially **Dr. Srihari**, **Dr. Ram Mohan Reddy**, **Raghavendra**, **Dr. Likewin Thomas**, **Dr. Manoj Kumar**, **Sachin D. Patil**, **Nikhil**, **Pramod Yelmewad**, **Bheemappa**, **Vishal**, **Rashmi**, **Bane Raman Raghunath**, **Dr. Sumith**, **Dr. Ganesh Reddy**, **Shiva** and **Ambikesh**.

My sincere thanks to the organizers of **Workshop on Speech and Image Processing (WISP)** and **Winter School on Speech and Audio Processing (WiSSAP)** for giving valuable suggestions on the research problem that I had chosen. A special thanks to **Prof. Sreenivasa Rao Krothapalli, Professor, IIT KGP**, **Prof. S.R.M. Prasanna**, and **Prof. K.S. R. Murthy** for their valuable comments on the work during my research.

A special thanks to the members of my **special group (Friends forever), Charitha, Manjunath** and **Susanna** for keeping me happy all the time and tracking my thesis status every day.

I would like to express my gratitude to all the professors who served as **Dean Academics (Prof. Sumam David, Prof. Katta Venkata Ramana and Prof. Sai Dutta)** for their support in all the academic regulations.

My sincere thanks to the **Directors of NITK (Prof. Swapan Bhattacharya** and **Prof. Uma Maheswara Rao)** for their support, without which it would not have been possible to conduct this research.

My sincere thanks to the members of **Academic section team**, especially **Dayanand** and **Prathibha** for their continuous support in processing applications.

I sincerely thank **Mrs. Nisha Shetty** for language corrections.

Special thanks to the **Security Guards of NITK** for their daily wishes with a smile.

Their smiles have helped me to forget the pressures of research.

Prior to NITK, many friends encouraged me to pursue my research in NITK. One special person among them is **P. Srinivasu**. This never ending list includes **Aditya, Naidu, Dr. Suresh, Suresh Chandra Satapathy, Saranya, Dr. K. Tammi Reddy, Prof. M. Kamaraju, Prof. G.V.S.N.R.V. Prasad, Prof. Y Srinivasu, Dr. T Srinivasa Rao, S.V.G. Reddy, Dr. Santhi Chilukuri and Madhusudhan**. I am eternally grateful to all of them for their support.

Thanks to my B.Tech, Diploma and SSC friends **Ramjee, Uma, Rama Rao, Ramakrishna, Raghuveer Arja,** and **Siva** for supporting me mentally and financially.

A family is the most special asset in anyone's life. Thanks to my **Maternal Grand Parents (Sri. Veeranki Jayasimhudu and Dhanalaxmi)**, **Grand Parents (Yarlagadda Butchayya and Ushagani)**, **Parents (Yarlagadda Srinivasa Rao and Vijaya Bharathi)**, **In-laws (Chebolu Srinivasa Rao and Baby Kanaka Durga)**, **Uncle(Veeranki Satyanarayana)**, who is my first teacher, **Aunt (Smt. Satyavathi)**, **Brother (Yarlagadda Chanikya)**, **Brother-in-laws (Pothuraju, Kondala Rayudu and Gowtham)** for their efforts in making my life special and worthwhile.

Last, but not the least, profound gratitude to my wife, **Yarlagadda Sugandha**, who bore all my pressures and shared my happy and sad moments. Thank you so much. A special thanks to my lovely daughter, **Yarlagadda Hamsini**, for making all my days happy. Even in the toughest situations, her smile means a lot to me and gives me strength.

Finally, Thanks to all the **GOD**s for keeping me healthy during my research career.

Place: Surathkal **Y. V. Srinivasa Murthy**

Date: December 12, 2018

Abstract

Music is a pervasive element of human's day-to-day activities. Most of the people love to listen to music all the time for handling their stress and tensions. Some are capable of creating the music. The importance of music for human beings has exploited the advancements in technology resulting in an enormous number of digital tracks. However, a majority of tracks are available with an inadequate meta-information. The meta-information is limited to the song title, album name, singer name and composer. Now, the question is how to organize them effectively in order to retrieve the relevant clips quickly, without proper meta-information like genre, lyrics, raga, mood, instrument names, etc. The process of labelling the meta-information manually for millions of tracks of the digital cloud is practically not possible. Hence, an area of research known as music information retrieval (MIR) has been introduced in the early years of 21^{st} century. However, it acquired much attention of researchers since 2005 with the support of Music Information Retrieval Evaluation eXchange (MIREX)[1] competition. There are several works that have been proposed for various tasks of MIR such as singing voice detection, singer identification, genre classification, instrument identification, music mood estimation, lyrics generation, music annotation and so on. However, the main focus is on Western music, and only a few works are reported on Indian songs in the literature.

Since Indian popular songs are contributing to a major portion of the global digital cloud, in this thesis, an attempt has been made to develop a few useful MIR tasks such as vocal and non-vocal segmentation, singer identification, music mood estimation and development of music recommender system in Indian scenario. Efforts have been put to construct relevant databases with a possible coherence for all the tasks mentioned above. Results include comparative analysis with standard datasets such as MIR-1K and *artist20* are given. For each of the four tasks, some novel approach has been presented in this thesis.

First, the task of vocal and non-vocal segmentation has been choosen to locate the

[1] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

onset and offset points of singing voice regions. A set of novel features such as formant attack slope (FAS), formant heights from base-to-peak (FH1), formant angle values at peak (FA1),formant angle values of valley (FA2), and singer formant (F5) have been computed and used for discriminating vocal and non-vocal segments. Also, an attempt has been made to develop a feature selection algorithm based on the concepts of genetics, known as genetic algorithm based feature selection (GAFS). The list of observations made out of this experimentation using selected features on the Indian and Western databases has been reported. Second, the task of singer identification (SID) has been considered. A database with the songs of 10 male and 10 female singers has been constructed. The songs are taken from two popular cine industries of Indian subcontinent named *Tollywood (Telugu)* and *Bollywood (Hindi)*. Various timbral and temporal features have been computed to analyze their effect on singer identification with different classifiers. However, the feature based systems are found to be less effective, and hence the trending convolutional neural networks (CNNs) have been used with spectrograms of song clips as inputs.

Identifying mood of the song has been considered as a third objective for this thesis. Six different moods are identified based on the analysis done on the combination of *Russell's* and *Thayer's* models (Saari and Eerola, 2014). We have developed, a two-level classification model for music mood detection. In the first stage, songs have been categorized into energetic or non-energetic songs. The actual class label has been predicted in the second stage. The performance of the system is found to be better in this case compared to development of single phase classification recommender system has been taken up using the labels like the title of a track, singer name(s), mood of a song, and duration. The graph structure based recommendation system has been proposed in this work to estimate the similarity in the listening patterns of same listeners. A graph has been constructed for every user by considering songs as nodes. Further, the similarities are estimated using the adjacency matrices obtained on listening patterns. This approach could be more appropriate for improving the performance of song recommender systems.

Keywords: Convolutional Neural Networks, Genetic algorithm based feature selection (GAFS), Graph based collaborative filtering, Formant Analysis, Music information retrieval, Music mood estimation, Music recommender system, Singer identification, Singing voice detection, and Vocal & non-vocal segmentation.

Contents

List of Figures

List of Tables

Abbreviations and Nomenclature

Abbreviations

AB.Tree	AdaBoost.Tree
AFTE	Auditory Filter-bank Temporal Envelops
AGG-DTM	Aggregating Song Level DTMs
ANN	Artificial Neural Network
AR	Amplitude Regression
AR	Auto Regression
ASD	Attack-Sustain-Decay
ASE	Amplitude Spectrum Envelope
AUC	Area Under ROC
BALS	Boundary Alignment Linear Scaling
BDS	Boosted Description Stumps
BPM	Beats Per Minute
CB-IR	Content-based Information Retrieval
CB-MIR	Content-based Music Information Retrieval
CBIR	Content based Image Retrieval
CDTW	Continuous Dynamic Time Warping

CF	Crest Factor
CMS-MFCCs	Cepstral Mean Subtraction based MFCCs
COMUS	Context-based Music Recommendation
CS	Chord Sequence
DAR	Diagonal Auto Regression
DCT	Discrete Cosine Transformation
DFT	Discrete Fourier Tranformation
DMM	Dirichlet Mixture Model
DNN	Deep Neural Networks
DTM	Dynamic Texture Mixture
DTW	Dynamic Time Warping
DWCH	Daubechies Wavelet Coefficient Histograms
DWPT	Discrete Wavelet Packet Transform
DWT	Discrete Wavelet Tranformation
EDR	Edit Distance on Real Sequence
EM	Expectation Maximization
EMD	Earth Mover's Distance
ENS	Echo Nest Song
ENT	Echo Nest Timbre
F0	Fundamental frequency
FFT	Fast Fourier Transform
FMCV	Frame MDCT Coefficient Vector

FP	Fluctuation Patterns
FT	Fourier Transformation
GLM	Generalized Linear Model
GMM	Gaussian Mixture Models
GTCC	Gamma Tone Cepstral Coefficient
HC	Harmonic Coefficient
HEM-DTM	Expected Maximization for DTM
HMM	Hidden Markov Model
HPCP	Harmonic Pitch Class Profile
ICA	Independent Component Analysis
ICM	Indian Classical Music
IGM	Inter Genre Similiarity
IIGM	Iterative approach of Inter Genre Similarity
IIR	Infinite Impulse Response
IMIRSEL	International Music Information Retrieval System Evaluation Laboratory
KNN	K-Nearest Neighbour
KTA	Key Transposition Algorithm
KTRA	Key Transposition Recursive Alignment
LCSS	Longest Common Sub-sequence
LDA	Linear Discriminative Analysis
LDTW	Local Dynamic Time Warping
LPC	Linear Predictive Coding

LPCC	Linear Predictive Cepstral Coefficients
LS	Linear Scaling
MAR	Multi-variate Auto Regression
MDCT	Modified Discrete Cosine Transformatino
MET	Mask Estimation Technique
MFCC-EMD	MFCC-Empirical Model Decomposition
MFCCs	Mel Frequency Cepstral Coefficients
MGR	Multi-variate Gaussian Regression
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MLL	Multi-label Learning
MP3	MPEG Layer-3
MRS	Music Recommender System
MSC	Modulation Spectral Contrast
MSD	Million Song Dataset
MSV	Modulation Spectral Valley
MUMS	McGill University Master Samples
NLP	Natural Language Processing
NLS	Note-based Linear Scaling
NLSH	Note-based Locality Sensitive Hashing
NRA	Note-based Recursive Align
OSC	Octave based Spectral Contrast

OSCC	Octave Scale Cepstral Coefficients
PCD	Pitch Class Distribution
PCDD	Pitch Class Dyad Distribution
PCP	Pitch Class Profile
PDF	Probability Density Function
PLSH	Pitch-based Locality Sensitive Hashing
PMCV	Phoneme MDCT Coefficient Vector
PTD	Propotional Transportation Distance
QBE	Query-by-Example
QBH	Query-by-Humming
QBIC	Query-by-Image Content
QBT	Query-by-Text
RC	Rhythmic Coefficients
RCEPS	Real Cepstral Coefficients
RH	Rhythmic Histogram
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
RP	Rhythmic Patterns
RWC-MDB	Real World Computing -Music Database
SC	Spectral Contrast
SCF	Spectral Crest Factor
SF	Spectral Flux

SFM	Spectral Flatness Measure
SH	Spectrum Histogram
SHD	Singing or Humming Discrimination
SMBGT	Subsequence Matching with Boundary Gaps and Tolerances
SMN	Semantic Multi-nominal
SPSF	Stereo Panning Spectral Features
SR	Spectral Roll-off
SRBM	Sparse Restricted Boltzman Machine
SSD	Statistical Spectrum Descriptor
STFT	Short-Time Fourier Transformation
SVM	Support Vector Machines
WLP	Warped Linear Prediction
ZCR	Zero Crossing Rate
ZT	Z-Transformation

Chapter 1

Introduction

> *Music gives a soul to the universe, wings to the mind, flight to the imagination and life to everything.*
>
> — Plato

1.1 Background

Music is a paramount element of a human's day-to-day activities. Everyone shows their love towards listening to music, except a negligible few while some are even capable of creating it. The importance of music in human life has taken advantage of technology advancements, resulting in an enormous number of music tracks. As a consequence, the complexity of searching a relevant track has been increased phenomenally. The present search engines are capable of providing information based on the input keywords provided. However, the difficulty in forming a text query for extracting information from multimedia data such as image, audio, and video has created an opportunity to design a system called content-based information retrieval. In particular, the process of extracting information based on passing an image as a query instead of text is called content-based image retrieval (CBIR). Several sophisticated tools for CBIR have already been designed and created in the past few decades, to extract information from images. Query-by-Image content (QBIC) and Image compass are the notable ones (Weihs et al., 2007). However, similar content extraction exercises have not been primarily reported in the literature on music information retrieval (MIR). The complex structure of a music signal[1] when compared to an image is the reason for not developing a sophisticated tool for music information

[1]The terms *'audio signal'* and *'music signal'* are interchangeably used in the entire thesis.

retrieval (MIR). Since the tool is expected to extract the information from an audio signal, it is popularly called as *content-based music information retrieval* (CB-MIR).

Though the number of digital tracks are extensively available in online and offline music stores, the information provided to develop the system on CB-MIR is inadequate. Most of the available labelled information is limited to the title of track, album name, singer, composer details, etc. only a few of them are labelled with the genre information. In a majority of the cases, genre labelling is done based on experts opinion. Moreover, many times the same song[2] is found to be mapped onto more than one genre. The answer to this question of what kind of information can be extracted, from music has remained an unanswered one for several years, from the time of the expansion of the digital cloud. Furthermore, it is practically impossible to manually label the meta-information for every track of digital cloud and may take several human years to do the same while new tracks are continuously being added to the existing repository. Hence, an idea of automatic music information retrieval system has been proposed by keeping in mind its importance for future users (Kassler, 1966).

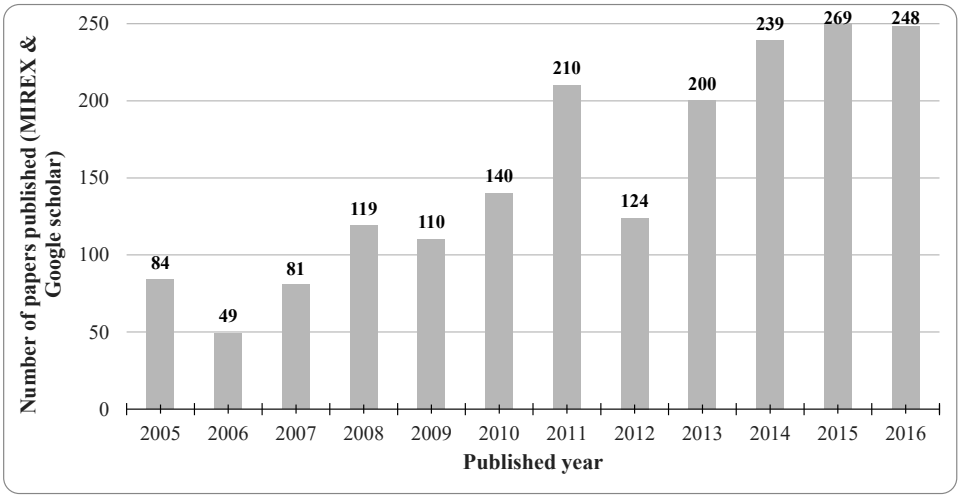

Figure 1.1: Increasing trend of research activities in CB-MIR in the last 10 years.
The above data is collected from `http://www.music-ir.org/mirex/wiki/MIREX_HOME` and `http://scholar.google.com` using the keyword Music Information Retrieval that contains all articles published in both conferences and journals

The clearly intended research towards music information retrieval has been mainly observed since the beginning of the 21^{st} century, because of the involvement of Music Information Retrieval Evaluation eXchange (MIREX)[3]. MIREX is an annual contest where researchers propose techniques for MIR and these are evaluated with the coordination

[2]The words *song* and *track* are interchangeably used in the complete thesis.
[3]`http://www.music-ir.org/mirex/wiki/MIREX$_$HOME`

of International Music Information Retrieval Evaluation Laboratory (IMIRSEL). Some of the prominent tasks that are received by MIREX are generally published in a prestigious musical conference called International Symposium on Music Information Retrieval (ISMIR). It is not surprising that the number of articles published on MIR has reached 1700 till 2016. An increase in the number of papers received every year indicates the importance of sophisticated MIR system for future generations. The same information has been depicted in the Figure 1.1 which shows the year wise growth in research in the area of developing MIR systems.

In this work, a few important broad categories have been addressed from the list of tasks identified by ISMIR, which are shown in Figure 1.2.

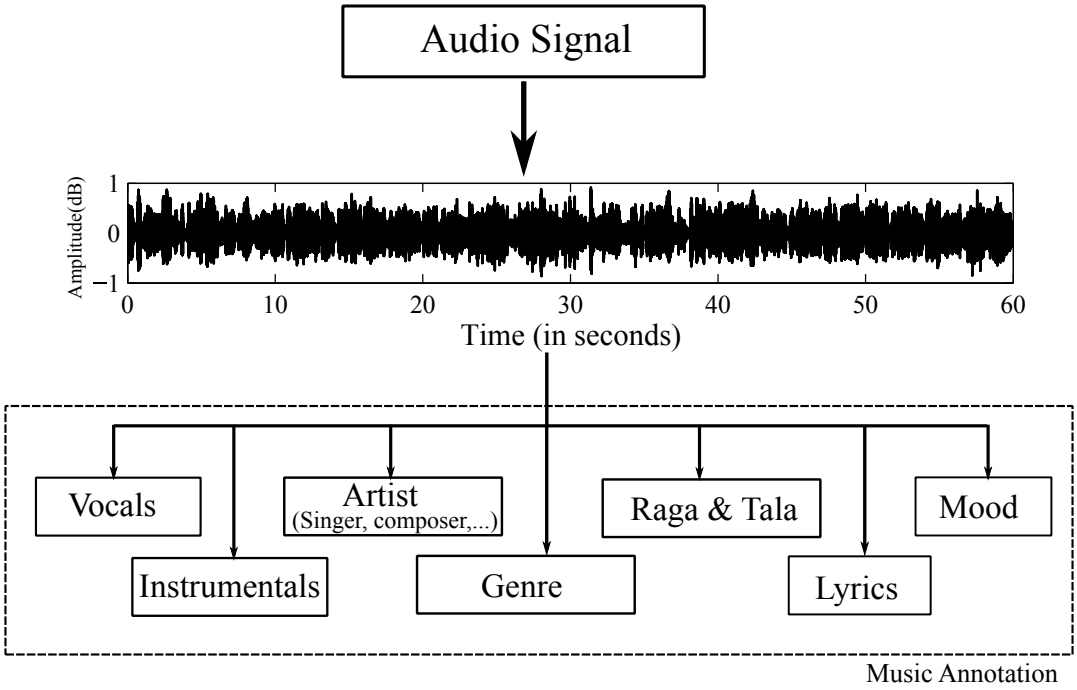

Figure 1.2: The possible information that can be extracted from an audio signal by developing an MIR system.

- *Vocal and Non-vocal segmentation:* A music signal contains several useful information. Of these, the most important portions are vocal, where the voice presents with background music, and non-vocals where it contains instrumental information. The very preliminary need of an MIR system gives the onset and offset locations of either vocal or non-vocal portions. The meta-information related to MIR can be found in vocal portions, non-vocal portions, or in both the regions. If vocal region is enough to extract meta-information, then there is no meaning in processing the whole signal. For instance, a small vocal portion is enough to recognize a singer

information instead of processing vocal and non-vocal regions.

- *Artist Identification:* It is a process of identifying either the singer, composer or instrumentalist information from the song. This is further extended to identify gender of a singer, song category (solo/ duet/ trio/ chorus), tracking a singer throughout the song and so on.

- *Genre Recognition:* This is the process of estimating the style or genre of a song clip. Many classes are identified to create the taxonomy for genre (Li and Ogihara, 2005). It is hard to accept that the research towards categorizing genre classes is still under progress.

- *Raga Identification:* Raga is a melodic framework of an audio clip. One can identify a raga based on the *tonic* frequency or *shadja*. Since there are 72 parent ragas and many child ragas in ICM, the process of automatic raga identification helps in several applications like online tutor systems, categorizing concerts, etc.

- *Music Mood Estimation:* Mood of the song is an important information that decides the present state of a person. Since it is very difficult to label all the moods provided in *Thayer's* and *Russel's* models, a majority of the works have concentrated on only valence and arousal moods. The process of labelling relevant moods for the songs have several applications in both, music categorization and pathological applications.

- *Instrument Identification:* Instruments are important components of a music clip that decides genre, mood and composer. They also help in annotating the music clip. The number of instruments considered to compose an audio clip can be estimated through the approach called, independent component analysis.

- *Lyrics Transcription:* Lyrics are another important portion of an audio clip. Provision of lyrics information facilitates the query-by-text application even for music. If a user remembers a portion of the song, then the same can be given as a query to extract the needed information.

- *Query-by-humming/singing (QBSH):* This approach is essential in implementing an application called, query-by-content. In this case if a user remembers a portion of the song, in this case, the same can be hummed to extract the relevant audio.

- *Music Annotation:* This is an ultimate solution for MIR which converts an entire audio clip in the form of text, by labelling each segment with relevant information like instrument name, singer name, composer name, genre, mood, raga, lyrics, etc.

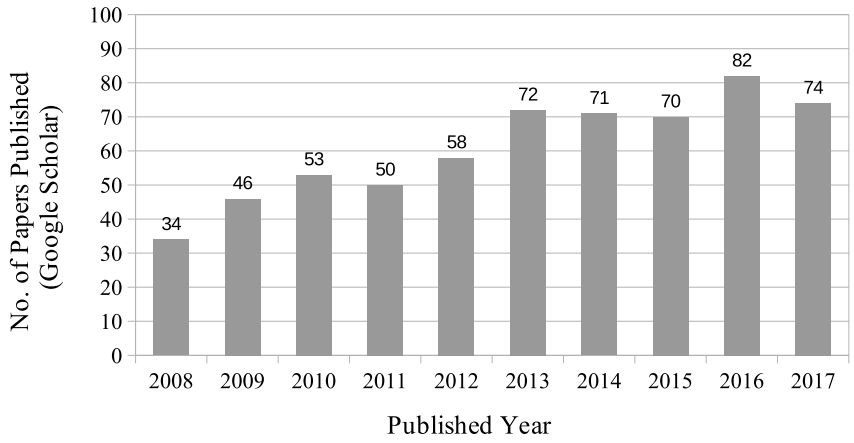

Figure 1.3: Increasing in the publications received for developing a Music Recommender System. The information has been collected from the the *Google Scholar* using the commands *"allintitle: music recommendation"* and *"music recommender"*.

It is quite easy to create personalized music collections if proper meta-information has been provided with the MIR system. However, a massive growth in the number of tracks in personalized music collections has deteriorated their value. It means that the tracks with personalized interest are listened to only once, or not at all, due to huge increase in the number of tracks. Due to this, many of the impressive songs are losing priority and their identity may be lost soon. Hence, there is a need to focus on improvisation in generating personalized music collections. An obvious solution is a recommender system, which was implemented by (Shardanand, 1994) for the first time in 1994. However, there has been a growth in the development of music recommender systems only in the later years of automatic MIR implementation. Minimal growth has been found till 2007, with the number of articles published on various aspects of MRS being 72, which is very less (Herrada, 2009). However, a consistent growth has been found later, with a variety of techniques. The same information has been depicted in Figure 1.3, which shows the year-wise increase in the publications of MRS. The information provided in the figure has been taken from *Google Scholar* website.

1.2 Motivation

The increasing number of tracks in the digital cloud needs high capacity storage devices. Fortunately, growth of hardware technologies is efficiently taking care of this rise in data quantity. However, the thought of saving every song in a personal storage is an impractical solution, as everyone is not interested in all the songs available. A better approach can be the process of identifying songs based on their contents rather than blindly relying upon the limited meta-data which is available at present with the tracks. The issue with many online music stores such as; *Gaana*, *iTunes*, *last.fm*, *Universal Music*, *Wynk* and so on is to categorize the song repository based on the music contents such as; lyrics, composer, mood, singer, language, gender, raga, etc., and provide them to the listeners based on their listening behaviour. In this scenario, the significance of music information retrieval (MIR) has gained much attention of the researchers where a variety of useful information can be obtained for every track of digital cloud (Jensen, 2010).

Basically, the tasks of MIR have been broadly categorized based on the aspects of objective and cultural information. The subcategories based on objective and cultural aspects are listed in Table 1.1. Of which, the list of objective aspects is primarily associated with the intrinsic properties of an audio clip such as melody, singer information, gender of a singer, rhythm, harmony, information about instruments, etc. There could be a straight away approach for developing an algorithm in the case of objective information. In contrast, categorizing the audio tracks based on cultural information is an ambiguous task. It is quite difficult to decide whether the information mapped to audio clip is appropriate or not. For instance, the process of labelling mood of a song clip is highly subjective. One can argue with a different mood name for same song (Pachet and Cazaly, 2000). Similarly, genre of a music clip is highly dependent on the geographical area, music culture, and many other parameters (Scaringella et al., 2006).

In the past two decades, several research works have been received on MIR. However, the focus has been mainly on Western music. Considering that Eastern music (*especially* Indian music) is also highly contributing to the digital cloud, there is a need to develop the suitable MIR systems for Eastern/Indian tracks as well. Moreover, appropriate MIR approaches are yet to be developed for almost all objective and cultural aspects.

The motivation for developing an MIR system is to maintain a proper catalogue for all the songs and to recommend songs based on user's interest. The popular Indian music websites such as *Gaana*, *iTunes*, *Wynk*, and *Youtube.com* are either recommending songs

Table 1.1: The division of MIR tasks based on the objective and cultural aspects.

Objective aspects	**Cultural aspects**
Singer recognition	Mood estimation
Composer identification	Genre classification
Gender recognition	Song similarity
Instrument identification	Rhythm similarity
Estimating beats per minute (BPM)	Singer performance assessment
Chord recognition	Music recommendation
Lyrics transcription	

based on text found in the title or merely by repeating the playlist. A majority of the listeners merely use the *'shuffle'* option which selects the songs randomly for playing from the playlist. New playlist can never be automatically generated based on listeners' needs in this case. Hence, we are motivated to develop an MIR system for Indian popular songs, which extracts essential information from a given audio clip and uses that for song recommendation. Since the singer and mood are two primary attributes that play a significant role in recommender systems, they are considered as primary objectives in this thesis. As the vocal and non-vocal segmentation is essential for locating singing voice segments, the task of vocal and non-vocal segmentation has been considered as an initial activity. Further, recommender systems based on graph structure analysis have been considered to recommend songs with specific analysis on their listening behaviour.

1.3 Challenges in MIR and MRS

Key challenges that generally need attention while developing MIR systems are addressed in this section. The first critical issue is the need for databases with original music tracks. A majority of the systems developed till now have used the artificial databases (Lesaffre, 2006). For instance, singer voices have been recorded externally instead of getting them from audio clips for singer identification (Bartsch and Wakefield, 2004). The performance that is obtained with such systems cannot be considered for real-time applications. The reason may be essential twined information like instruments, background support, chorus, etc, along with vocals within the audio clip is ignored in artificial databases.

Moreover, it is strenuous to separate the singing voice alone from background complex instrumentals (Lehner and Widmer, 2015; Vincent, 2006). Ideal MIR systems are

expected to segment an each source of musical sound. However, it is unlikely due to their frequency overlapping between many instrumental sounds. This overlapping nature is a basic challenge for not developing an ideal system to label each instrument information in a given polyphonic audio clip. However, repetition in a pattern is the intrinsic characteristic of music that helps to remove the background music to some extent (Rafii and Pardo, 2013; Thomas et al., 2016). Unlike speech processing, in which a majority of the portion is voice alone, an audio clip is a mixture of several components like intro, verse, chorus, bridge, and outro. Primarily, music clips may be categorized into vocal and non-vocal portions. The vocal regions contain either singing voice or chorus, possibly along with background accompaniment. It is very rare to find pure vocals in a given audio clip. Hence, the systems developed using studio recorded artificial databases, miserably fail in the case of real-world scenarios. In addition, labelling the subcategories of cultural information such as genre, mood, rhythm, etc, is also a challenging task. An MIR system with a few essential components such as vocal and non-vocal segmentation, singer identification, mood estimation, genre classification is sufficient for many applications including music recommender system.

Many issues are to be considered while developing an ideal MIR system. Music listening behaviours of an individual change frequently and abruptly. Sometimes, the listener in repeatedly listening to the same song, several times. Later, this interest may fade. These are typical unpredictable problems that are to be faced while developing a recommender system. Moreover, recommending a music track is not as easy as, recommending the books or any other products (Herrada, 2009). The concept of collaborative filtering technique, where it is assumed that individuals with similar listening behaviour listen to the same songs, may not be suitable in many cases. Hence, a different approach that adapts individual user's listening behaviour is needed. The future of recommender systems is highly dependent on perceived accuracy, than the predicted one. This is also an important aspect to be monitored while developing a system for music recommendation.

1.4 Applications of MIR and MRS

Music information retrieval (MIR) is a trending research field getting motivation from its diversified and commercial applications. Some important applications of both MIR and MRS are presented in this section.

1.4.1 Music Information Retrieval (MIR)

The basic purpose of designing a MIR system is to support users in selecting the required audio based on some machine driven measures. There are two such measures reported in the literature based on the specificity (Casey et al., 2008). The high specificity system is expected to retrieve exact matching information in an audio clip. If specificity is low, then retrieval of the clip with statistically similar properties is enough (Grosche et al., 2012). However, it should be possible to retrieve audio clips based on album, singer, genre information and other relevant information in any case. The signal level similarity helps to develop low specificity system. One significant application of low-specific MIR is assessing the candidates' performance objectively during live singing performances (Biswas et al., 2018), where judges can assess the performance subjectively. It is also possible to assess plagiarism content in an audio clip by comparing a song clip with the available original audio clip. Similarity check can further help in providing copyright information to an audio clip. The well-known applications designed for the task of similarity measurement are Shazam[4], MusicID[5], and Vericast[6] (Wang et al., 2003a).

Name of the music clip is considered as an important piece of information, while categorizing audio clips in the case of traditional classification systems. The process of cataloguing the music clips based on their objective similarity using dynamic time warping (DTW) sequence algorithms, is always a suitable and robust approach instead of their track or album names (Müller et al., 2006). Similarly, collecting the songs of a particular singer and keeping them as one cluster is also a kind of cataloguing (Cunningham et al., 2012). Many times, we observe that the same song is available in the repository with variations in instrumentation, harmony, key, rhythm or structure. The process of identifying such similar songs is useful in grouping the similar tracks together; generally called as cover song detection (Serra et al., 2010). Identifying cover song is another important application of MIR system.

An approach of similarity measurement leads to retrieval of a music clip based on the input query in the form of music itself instead of just keywords. Many a times, it is practically not possible to provide keywords for each portion of an audio clip, as the listeners may not be able to remember and reproduce the keywords; rather they hum or sing a portion of the song. Query-by-humming/singing (QBSH) is another important

[4]http://www.shazam.com
[5]http://www.gracenote.com/music/recognition/
[6]http://www.bmat.com/products/vericast/

application of MIR system where the user can conveniently frame a query for searching the needed song. Many commercial systems are already supporting QBSH on limited data. MUSART testbed (Dannenberg et al., 2007) and SoundHound[7] (Salamon et al., 2013) are some notable ones. The impact of MIR is to extract musical clips based on semantic queries framed by listeners. For instance, a user can frame a query like "List all the clips that have a tempo of 110 bpm at C major scale" (Isaacson, 2002). It is possible to develop a system that can handle such type of queries using semantic information (Knees et al., 2007; Turnbull et al., 2009).

In contrast to the above mentioned low specificity information, systems with some high specificity information such as singer, mood, genre, instrument(s), raga, and lyrics, is also essential and helps in cataloguing and indexing music clips. Cover song albums also can be generated by collecting clips of similar high-specificity information. For instance, if anyone is interested in listening to all songs of *S.P. Balasubrahmanyam* (SPB, a famous Indian singer), then all SPB songs can be grouped by providing a suitable cover label. Estimating prevailing mood of a song further helps to judge the mental state of a listener at that moment. All high-specificity information along with relevant low-specificity information greatly help in developing a sophisticated music recommender system.

1.4.2 Music Recommender System (MRS)

The primary goal of music recommender system (MRS) is to propose a set of music tracks based on the listening behaviour of a user. A recommender system is capable of meeting four specific requirements called accuracy, diversity, transparency, and serendipity (áOscar Celma, 2010; Ricci et al., 2011). They have been categorized based on the similarity among the recommended playlist and user's preferences; user's satisfaction with comments, trust in the process of recommendation, and feeling towards surprising recommendations. Recommendation based on listener's behaviour further helps to get the knowledge of user's present mental state. For instance, if the user is continuously listening to the sad songs, then it is psychologically an indication of the depressed, hurt or inferior state. Such analysis assists pathologists and psychiatrists to take necessary steps for the treatment of such listeners.

Two primary objectives of MRS are playlist generation and recommending music for visual presentations. The MRS system should keep up the above mentioned four re-

[7] `http://www.soundhound.com`

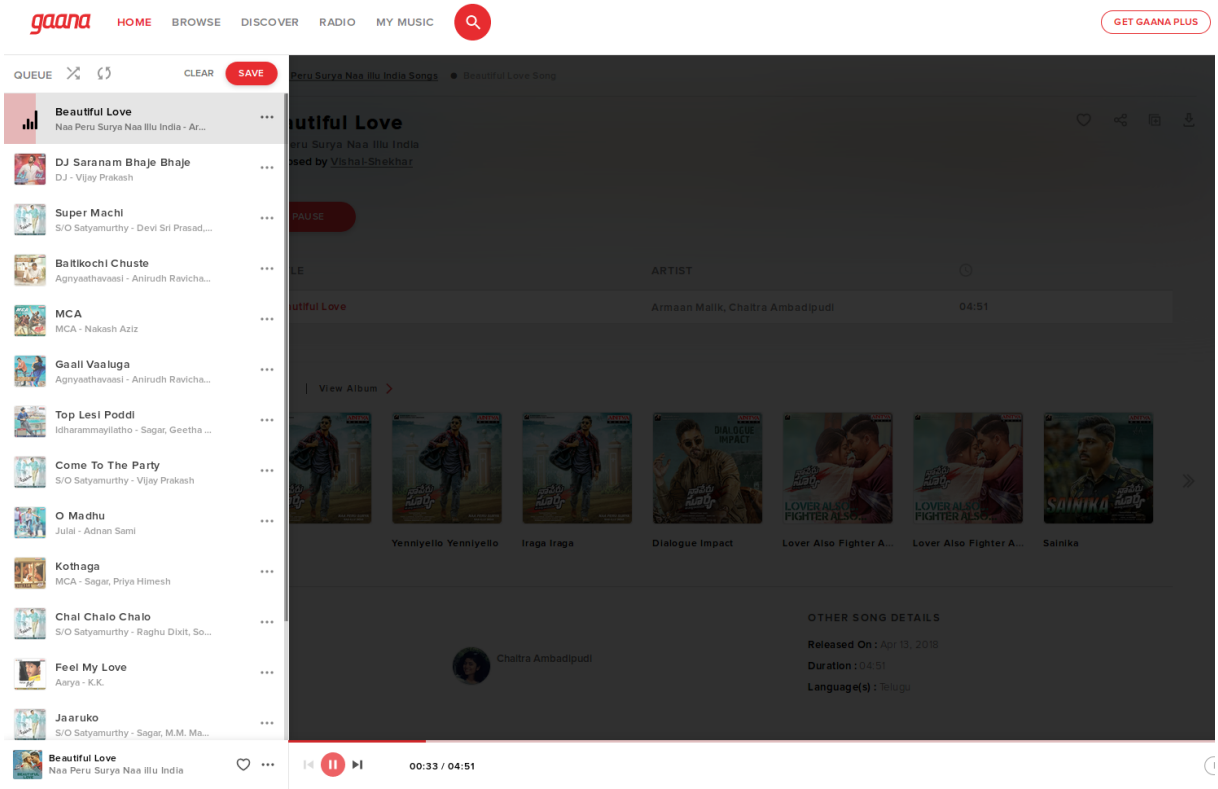

Figure 1.4: An instance of *Gaana* website that shows the nature of present recommender system in filling the playlist based on manual linking.

quirements of the user in mind while filling the playlist. Commercially available popular websites such as *Last.fm*[8], *Pandora*[9], *Gaana*[10], *iTunes*[11], and *Wynk*[12] presently recommend songs to the users based on manual similarity metrics that are available as an information in the digital cloud. However, this information is available for a limited number of songs. In this case, when the user accesses a song, all other linked songs appear as the list of recommended playlist. The same list repeatedly appears whenever the user listens to a particular song. These kinds of recommendation systems may not sustain for a long time due to overlooking the properties of diversity and serendipity. For instance, *Gaana* website exactly provides the exact play list for the specific song played twice at two different instances in the gap of three months. There is no difference found in the recommended play lists that is obtained with the song played once in the mid of April 2018 and in the second week of June of the same year. The list has been displayed in Figure 1.4. The selected song is "*Beautiful Love*" from the album name "*Naa Peru Surya Naa Illu India*" of Indian *Tollywood*.

[8] `http://www.lastfm.com`
[9] `http://www.pandora.com`
[10] `http://www.gaana.com`
[11] `https://www.apple.com/in/itunes/music/`
[12] `https://www.wynk.in/music`

The recommender systems may also be helpful in recommending songs to a car driver to keep him alert while driving. Listening to lullabies while driving may lead consequences that can be fatal (Baltrunas et al., 2011). In such cases, recommender systems would be helpful in filtering out unwanted lullabies and also recommend energetic songs to keep driver awake and alert. Presenting visual perception for a pattern of background music thrills the listener. Future MRSs must address this issue as well.

1.5 Scope of the Thesis

This work focuses on the importance of content-based music information retrieval (CB-MIR) mainly for music recommendation. Comparatively, MIR research that has been reported in the context of Indian songs is too less. However, contribution of Indian songs to the growth of digital cloud is considerably high. Hence, the songs of two popular Indian cine industries namely, *Tollywood* and *Bollywood*, are considered for the work presented in this thesis.

This work has focused on four individual components of music information retrieval, of which two are chosen under the category of objective information and the other two are from cultural information. Since the aim of this work is to propose a new approach for music recommender system (MRS), the meta-information which is highly useful for MRS has been obtained through automated approaches. Four different databases with the tracks of *Tollywood* and *Bollywood* have been constructed with sufficient clips and care has been taken to provide possible coherence among them. The performance obtained is based on the experimentation done on closed set database. A little deviation in the performance has been observed with the open testbed. The scope of each research objective is given below:

- The task of vocal and non-vocal segmentation has been addressed by considering a variety of music clips that include different male and female singers, instruments, repeated patterns, pure vocals, vocals with background support, different moods, etc. The accuracy which has been obtained is based on frame-level features. The formant features are newly proposed. The genetic algorithm based feature selection approach is found to be better when compared to other methods explored. It may take more number of iterations to provide an optimal feature vector if the database size is increased.

- The data of twenty singers has been taken for the task of singer identification. Size of the database considered for singer identification is larger when compared to that for vocal and non-vocal segmentation. The performance reported in this work is strictly limited to the list of singers considered.

- Six basic moods are considered in this work. The moods are identified based on *Thayer's* and *Russell's* model. The database used in this work is an expanded version of the one used for singer identification and vocal & non-vocal segmentation.

- In this thesis, a dataset with 1000 songs; listened by 500 listeners has been collected to develop a music recommender system.

1.6 Thesis Outline

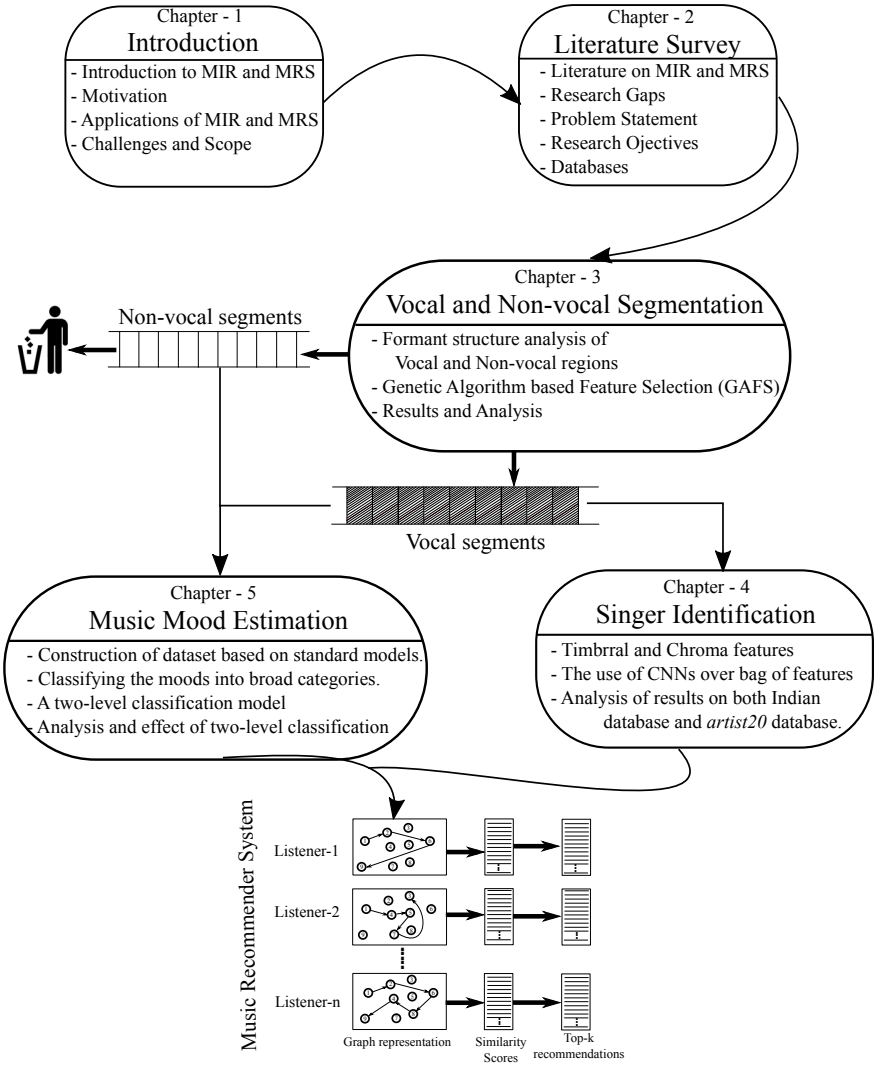

Figure 1.5: A brief thesis outline which describes about the organization of remaining chapters.

The remaining chapters of the thesis are organized as follows.

The detailed literature survey on the various tasks of MIR has been given in **Chapter 2**. It covers the various approaches proposed for each task of MIR, limitations, and future scope of some important research works for developing a MIR system. A few important tasks such as vocal and non-vocal segmentation, singer identification, music mood estimation, and recommender system are considered for detailed survey. The remaining tasks are condensed for this thesis as there is less focus on them for this thesis. The research gaps that are identified from the selected MIR tasks are listed that motivated us to frame a problem statement with relevant objectives.

In **Chapter 3**, the importance of vocal and non-vocal segmentation for the rest of MIR tasks has been detailed. Then, the process of extracting suitable features for segmenting vocal and non-vocal regions, based on the formant structure, is discussed. Further, the explanation on the approach which is proposed for selecting the relevant features using genetic algorithms is given. The detailed analysis of the results obtained for vocal and non-vocal segmentation has been given, followed by summarizing the chapter.

In **Chapter 4**, the portions of vocal regions have been taken for the task of singer identification. The experimental analysis has been given for 20 Indian singers, using both traditional bags of features approach and trending convolutional neural networks. Further, the performance comparison between Indian singers and the singers of artist20 database has been given. Further, the chapter summarizes the work done for singer identification.

In **Chapter 5**, the importance of mood estimation, to estimate the listener's behaviour, has been detailed. The effort behind the selection of moods with some opinion scores has been given. The explanation on the proposed two-level classification and its efficiency over direct classification has been given. Further, the chapter summarizes the work done on music mood estimation.

In **Chapter 6**, the role of graph structures in recommending music to listener's based on the collaborative filtering has been detailed. Different similarity measures that are considered for estimating the affinity among the listeners have been explained. Further, the chapter has been summarized.

In **Chapter 7**, the conclusion of the works done on four tasks of MIR have been detailed. Further, some important future directions have been given that may help the future researchers to select the suitable problem and to know the status of a MIR system.

Finally, the reference research works that have been used as guides us in implementing

the proposed system are given followed by the list of papers presented based on this work. A brief outline of the thesis has been depicted in Figure 1.5 which gives clear information about the organization of the remaining chapters.

Chapter 2

Literature Survey

> *Music can change the world because it can change people.*

— Bono

This chapter covers in detail the existing literature on the selected subcomponents of an MIR. It also touches upon the motivation for considering them in this thesis. Singing voice detection, artist identification, genre identification, raga identification, instrument identification, mood estimation, and music annotation are the major components of MIR. Of these, singer and mood are two important attributes that affect the recommendation performance in MIR. Hence, the literature with respect to singer identification and mood estimation has been detailed in this chapter. Prior to that, the need of singing voice detection and the works done are also detailed. In addition, works done on music recommender system are also explained here. This chapter also includes sections on datasets used in MIR works, feature extraction process and classification models.

2.1 Datasets used in Various MIR tasks

The difficulty, of arranging the tracks into different categories, is increasing, with an increase in the number of digital tracks. The task-relevant tracks are highly essential for building a sophisticated MIR system (Casey et al., 2008). For instance, different instrument clips are preferred for developing the task of instrument identification instead of clips with audio and polyphonic sounds. In this regard, it is useful if the benchmark datasets are identified and listed for use in future research. Identifying task-specific datasets helps the researchers in comparing their works with the state-of-art systems.

In many cases, the copyright issue of commercial audio clips is of paramount cause which leads to the use of existing datasets for research. Several datasets have been found for the tasks of MIR based on few notable sources such as wiki[1], ISMIR[2], Colinraffel.com[3], and audio content analysis websites[4]. Most of the datasets mentioned in these sources do not contain precise information. Many of them are not available publicly; some datasets have very few clips to use for experiments. However, from the exhaustive list, we have identified some prominent datasets that are publicly available and highly used in the past two decades. Table 2.1 lists these datasets with some necessary information.

Table 2.1: Highly used and publicly available datasets with some useful information.

Sl. No.	Datasets	Ref.	#Clips	Purpose‡	Sl.No.	Datasets	Ref.	#Clips	Purpose‡
1.	RWC	Goto et al. (2003)	465	IR	12.	1517-Artists	Seyerlehner et al. (2010)	3,180	AI
2.	GTZAN	Sturm (2013)	1,000	GC	13.	MIR-1K	Hsu and Jang (2010)	1,000	VOD, AI
3.	USPoP	Lidy and Rauber (2005)	8,752	QBH	14.	OMRAS2	Fazekas et al. (2010)	1,52,410	MA
4.	BallRoom	Tsunoo et al. (2009)	698	GC	15.	TagATune	Hamel et al. (2012)	25,863	MA
5.	ISMIR2004	Cano et al. (2006)	1,458	GC	16.	CAL10K	Tingle et al. (2010)	10,271	MA
6.	103-Artists	Schedl et al. (2005)	2,445	AI	17.	UNIQUE	Seyerlehner et al. (2010)	3,115	QBH
7.	Homburg	Homburg et al. (2005)	1,886	GC	18.	MSD	Bertin-Mahieux et al. (2011)	10,00,000	GC
8.	Codaich	McKay et al. (2006)	26,420	GC	19.	MusiClef	Orio et al. (2011)	1,355	AI, MA
9.	LMD	Silla Jr et al. (2008)	3,227	GC	20.	Ext.BallRoom	Marchand and Peeters (2016)	4,180	GC
10.	Artist20	Ellis (2007)	1,000	AI	21.	AudioSet	Gemmeke et al. (2017)	20,84,320	AEI
11.	CAL500	Turnbull et al. (2007)	500	MA	22.	FMA	Benzi et al. (2016)	1,06,574	GC

‡AEI - Audio Event Identification, AI - Artist Identification, GC - Genre Classification, IR - Instrument Recognition, MA - Music Annotation, QBH - Query-by-Humming, and VOD - Vocal onset Detection.

An effective MIR system can be built if the task-specific benchmark datasets are available. It is observed from the literature that the datasets available are less complex, and are recorded with limited scope. The datasets with incomplete and monotonic information are not suitable for many real-time applications. Considering the literature, the genres of eastern countries are less focused, especially Indian categories. As they contribute to a major portion of the digital music world, it is essential to develop a sophisticated MIR systems on Indian categories.

2.2 Features and Classification Models

Audio songs are mainly available in the form of high-quality audio CDs', recorded with a sampling frequency of around 44.1 KHz., in offline and online stores. Direct processing of these high-quality audio songs for information retrieval consumes large memory and processing time. Generally, numeric features that resemble the signal characteristics and

[1]https://en.wikipedia.org/wiki/Wiki
[2]http://www.music-ir.org/mirex/wiki/MIREX_HOME
[3]http://colinraffel.com
[4]https://www.audiocontentanalysis.org/data-sets/

compactly represent the original audio songs are extracted. There are enormous number of features that have been introduced with the support of various signal processing techniques and statistical methods. A majority of them are used to characterize the music. Some additional features are also introduced to model the music signal in a better way. Based on this, a hierarchy has been provided by (Scaringella et al., 2006) for audio features. In this article, features have been categorized into three classes, namely timbre, pitch, and rhythm. Later, the taxonomy has been revised by (Weihs et al., 2007) who categorized them into short-term, long-term, semantic and compositional feature sets. Although the features mentioned in the article are mainly based on few concepts of music research such as music similarity, transcription, and cognitive psychology, they cannot be generalized and used for all MIR tasks. Hence, the two taxonomies have been combined and enhanced in (Fu et al., 2011) to present a generalized hierarchy for audio features. The features discussed in this chapter are mainly classified into (i) Low-level, (ii) Mid-level and (iii) High-level features, as shown in Fig. 2.1(a).

In general, low-level features are extracted from the smaller audio segments, of length $10 \sim 100$ milliseconds, known as frames. The mid-level features are extracted from a note level (Kitahara, 2010). Low-level features carry abstract characteristics of a frame. They cannot represent the characteristics of an entire signal. Mid-level features can mimic the characteristics of an entire signal or set of segments. They can be computed on longer segments or by applying statistical operations on low-level features (Dittmar et al., 2007; Peeters, 2004; Wolter et al., 2008). High-level features provide semantic information such as annotations, which are useful for labeling the clip and help for easy retrieval. The combination of low and mid-level features is used to decide the high-level information such as genre, mood, instrument, artist and so on. The following subsections describe various low and mid-level features.

2.2.1 Low-level features

These are common block-based features used for various tasks related to music. They are further classified into timbre and temporal features.

The number of vibrations caused, to produce sound waves in a second is known as *pitch* (also called *fundamental frequency (F0)*) of a note. Strength of the signal can be measured by computing the sum of squares of samples called *energy* of the signal. Timbre is the quality of a musical tone which helps to differentiate the voice or instruments even

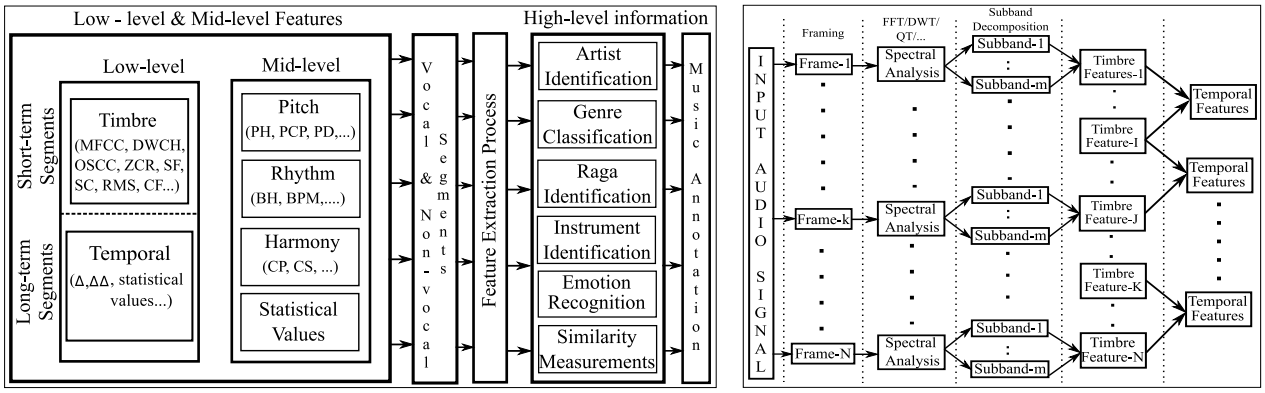

Figure 2.1: (a) Audio feature classification as Low-level, Mid-level and High-level information. (b) Process of extracting low-level features.

when their pitch and energy are the same. For instance, if a guitar and a piano are playing the same note on same scales then timbre of those instruments helps to classify them. In psycho-acoustics, timbre is defined as the voice quality of a musical note, sound, tone color, or tone quality that distinguishes various kinds of sound sources (Erickson, 1975; Popper and Fay, 2014). Timbre features are generally computed from the frames of length 10-100 ms. The main advantage of this approach is the technical simplicity and availability of well established methods, to process stationary signals in terms of effectiveness and complexity. The general process for low-level feature extraction is shown in Fig. 2.1(b). Initially, the input signal is divided into chunks of frames and are transformed into frequency domain using various transformations such as Fourier transform, constant Q-transform, wavelet transform and so on. A sub-band decomposition technique is applied to the frequency domain signal. Each sub-band is analyzed to extract the timbre features of a frame. A combination of timbre features is used to extract temporal features. The low-level features can be extracted from both, time and frequency domains. The important features found in the literature that are extracted from time-domain information are root mean square (RMS) energy, zero crossing rate (ZCR) (Tzanetakis and Cook, 2002; Li et al., 2003; Bergstra et al., 2006; Mörchen et al., 2006), and crest factor (CF)[5] (Helen and Virtanen, 2005). To analyze the signal in the frequency domain, various transformations such as discrete Fourier transformation (DFT) (Ahmed et al., 1974), discrete wavelet transformation (DWT) (Bruce et al., 2002), and constant Q-transformations (Schörkhuber and Klapuri, 2010) are applied. From the spectrum obtained through transformation, it is possible to extract features like spectral roll-off (SR), spectral centroid (SC), spectral flux (SF), and bandwidth using statistics. Instead of applying Fourier transformation (FT) on complete

[5]Crest factor is a ratio of amplitude peak and RMS value and is obtained as $CF = \frac{peak(|signal|)}{rms(signal)}$.

signal, the short-time Fourier transformation (STFT) can be used to extract potential features such as mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), Daubechies wavelet coefficient histograms (DWCHs), spectral flatness measure (SFM) (Allamanche et al., 2001; Benetos et al., 2006), spectral crest factor (SCF) (Cheng et al., 2008), and amplitude spectrum envelope (ASE) (Kim and Sikora, 2004; Lee et al., 2009). To obtain MFCCs, the sub-bands during spectrum computation are linearly spaced up to 1000 Hz and are logarithmically spaced at higher frequencies. MFCCs can model music patterns better than other spectral features (Mörchen et al., 2006). Moreover, segments of variable length are used to extract cepstral features based on inter-beat segments which are more relevant than the traditional equal sized block processing approach. This has led to the invention of new kinds of features, i.e. OSCCs (Maddage et al., 2004). This approach is extended to extract relevant features directly from MP3 files with slight modifications in discrete cosine transformation (DCT); this process is called modified discrete cosine transformation (MDCT). An attempt has also been made to extract the features from both recording channels (left and right) since the vocals of both channels are common and the non-vocals vary for most of the times. To identify the spectral distribution in both channels, the stereo panning spectral features (SPSF) have been introduced (Tzanetakis et al., 2007, 2010). However, many of the timbre features mentioned above are adopted from the works of speech processing. As there are several distinctions between speech and music, the features used in speech processing may not be suitable for extracting the distinct timbre effect for efficient music analysis.

The temporal variations in the signal help in several music classification tasks. Temporal features are a kind of low-level features extracted on top of timbre features and useful to observe the temporal feature transformation of the given signal. Generally, to compute the temporal features, statistical parameters such as mean, variance, co-variance, and kurtosis are applied on large number of local windows (Tzanetakis and Cook, 2002). The means and variances have been computed from a timbral texture to form a feature vector, called *MuVar* (Tzanetakis and Cook, 2002). The means of covariance values are computed from a covariance matrix to form a feature vector called *MuCov* (Mandel and Ellis, 2005). *MuVar* and *MuCov* are also explored in the literature to observe the temporal variation in the signal (Li et al., 2003). The same operations are performed on the frames of larger lengths and named $MuVar^2$ and $MuCov^2$. Normally, the block based process is used to extract temporal features considerably increasing the computational complexity. Hence,

feature integration is done using the other techniques such as amplitude regression (AR), multi-variate auto regression (MAR), multi-variate Gaussian regression (MGR), and diagonal auto regression (DAR) to reduce the complexity issues (Mandel and Ellis, 2005). Along with the other available techniques, probabilistic models are also used to extract the temporal features. One such model is hidden Markov models (HMM) (Weihs et al., 2007; Reed and Lee, 2009) which models the time series data using hidden states. In HMM, each frame is treated as a state that helps to provide the feature set for the current frame based on the output probabilities of the previous frames.

2.2.2 Mid-level features

Human ears can perceive the intrinsic properties of any music with the help of integrated biological mechanism. Generally, low-level features that have been discussed in previous subsection have failed to capture much of the required information from a given music clip. Thus, mid-level features have been introduced and they are mainly used for the tasks such as QBE, Query-by-Singing/Humming (QBSH), cover song detection, raga identification, and so on. There are three broad categories of mid-level features namely: (i) pitch− the fundamental frequency, (ii) rhythm− the recurring pattern of tension, and (iii) harmony− a mixture of notes that are played simultaneously and successively to produce chords and chord progressions (Zentner, 2003). In music processing, pitch plays an important role for different applications such as QBH and raga identification. Other factors such as context, loudness and timbre also influence the pitch. In the musical context, the pitch is not a single fundamental frequency (F0) since every instrument has its own harmonic frequency series. Multi-pitch estimation is necessary for such cases. Few algorithms (Tolonen and Karjalainen, 2000; Klapuri, 2003) have been designed especially to estimate the multiple pitch values. These algorithms are helpful in extracting the pitch values at frame level and song level using pitch histograms (PHs). The PHs are used to recognize the genre and mood of a song with the additional support of MFCCs and other perceptual features. Along with the PHs, other different features such as pitch class profile (PCP) can also be used for music processing. The first note of the C major scale is the note C. If it is pitched around 261.63 Hz, then the low-C and high-C is around 65.40 Hz. and 1046.50 Hz. respectively (Casey and Slaney, 2007). Though there are several variations in pitch frequency, all the variations are considered in the same pitch class (Krumhansl, 2001). PCPs and harmonic pitch class profiles (HPCPs) are helpful in extracting the chroma

(pitch class) features. Chroma features are helpful in analyzing the melody of a song including *gamakas*[6] (Sarala and Murthy, 2013).

The occurrence and recurrence of patterns can be discriminated using rhythmic features. These features are mostly helpful in recognizing the repeated pattern in a song clip. The most repeated pattern in any song is known as a beat (Geist et al., 2012). The features such as beats per minute (BPM) and tempo are useful to estimate the beat locations. Another way of computing beat features is by taking the envelope of an auto-correlation of a given input signal. The regularity in peaks of the auto-correlation signal helps to compute beat histograms (Tzanetakis and Cook, 2002). In the literature, rhythmic features are also used for mood estimation tasks (Feng et al., 2003a; Macy, 2001). The results indicate that the mood of a song is highly correlated to the rhythm. It is normally observed in music clip that each mood is roughly associated with some value of a scale (Yang et al., 2010, 2008).

The third important mid-level feature is harmony which can be recognized through several factors. Of these, one is chord sequence (CS). Harmony is quite different from melody since melody obtains the horizontal information and harmony obtains the vertical information of a song (Kuusi, 2009). Melody is the linear succession of musical notes and is a combination of rhythm and pitch. Harmony is the combination of simultaneous notes or *chords*. The CS can be extracted by some chord detection algorithms found in the literature (Turnbull et al., 2007; Gómez and Herrera, 2004; Jensen et al., 2009). These sequences are also helpful in detecting the multiple fundamental frequency values present in the chord since a chord is the combination of more than one note played together. The harmony features are used in the literature for the cover detection of a song (Bello, 2007) and song similarity (Ellis and Poliner, 2007). Although mid-level features can capture the intrinsic properties of a music clip such as pitch, BPM, melody, harmony, rhythm, etc., they alone are sometimes not sufficient enough to achieve good results. The combination of both low-level and mid-level features can give better results (Kitahara, 2010). In pattern recognition applications, it has been difficult to establish a strong correlation between specific tasks and features. In such cases, a set of features is used initially, and later, feature selection techniques such as elimination, correlation, and so on, are applied to reach the optimum feature set (Li and Ogihara, 2006; Shen et al., 2006, 2009; Fu et al., 2010; Ness et al., 2009; Barrington et al., 2008).

[6] A '*gamaka*' is an ornament which gives soothing effect for the *raga* of ICM.

2.2.3 High-level features

Sometimes, it is also possible to extract the features directly from the complete signal. They can be artist name, album name, genre, raga of a song, mood of a song, singer name, instrument names, song title, etc. However, it is not possible to characterize the signal at once due to its stochastic nature. Hence, low and mid-level features are together useful to identify the high-level information which is mentioned above.

The features specified in low-level and mid-level categories are task specific. However, the task of vocal and non-vocal segmentation is the prior task for any music information retrieval, to avoid irrelevant portions for a particular task. Hence, some suitable feature vectors can be considered to accurately segment vocal and non-vocal segments. Further, task related features may be considered to perform extract meta-information shown in fourth block of 2.1 (a). All of the extracted high-level information are further considered to annotate each portion of music clip.

The process of selecting a suitable classification model is the next important step while developing MIR system. There are three categories of classification models namely (i) unsupervised, (ii) semi-supervised, and (iii) supervised. Since the audio data is highly non-linear and a majority of them are classification problems, unsupervised classification models may not handle them effectively. Several supervised classification models such as artificial neural networks (ANNs), Gaussian mixture models (GMMs), support vector machines (SVM), AdaBoost (AB), generalized linear models (GLMs), k-nearest neighbor (KNN), sparse restricted Boltzmann machine (SRBM), and so on, have been considered for a variety of MIR tasks. Since the classifier selection is completely dependent on the constructed feature vector (Murthy and Koolagudi, 2018b), it is highly difficult to suggest a single classifier for the specific task. The performance of the system with different classification models is given in respective sections of this thesis. For instance, if the data falls under normal distribution, then GMM is the better classifier.

The following sections detail the related works taking place in the selected subtasks of MIR for this thesis.

2.3 Vocal and Non-vocal Segmentation

An audio signal is a combination of pure vocals, instrumental region, silent regions (SIL), and vocals with background instruments. Since a majority of the users are interested

in listening to popular songs, identifying the popular song structure is an interesting research assignment. In this section, two important issues are observed while processing audio clips, and are discussed along with their possible solutions. Identification of SILs is the foremost pre-processing step in any speech and audio-processing task. Generally, the length of the silence portion in a song is negligible, because more than 99% of the audio songs are occupied by either a singing voice or an instrumental sound. The second issue is segmenting the vocal and non-vocal regions as the music signal, and is the complex cohesion of these two components.

As vocals are usually accompanied with background music, segmentation becomes a challenging task. Segmentation is a prerequisite for singer identification, emotion recognition, instrument classification, lyrics transcription, and so on. One of the interesting commercial applications of vocal and non-vocal segmentation is the *karaoke* system. *Karaoke* is a Japanese word, which means only music track without vocals. This is helpful for music enthusiasts to learn singing for many existing compositions or to use the tracks in concerts for simulating reality or ro sing with the existing instrumental composition of a particular song. Presently, the extraction of karaoke tracks is being done manually during recording, which requires a lot of manual effort and time. Segmentation of vocal and non-vocal regions is an essential step in designing an automated *karaoke* system.

For automating segmentation, several approaches have been reported in the literature. Initial attempts have been made, to analyze the signal in time domain, by using simple features such as energy, ZCR, and so on; these values get a sudden jump when vocal region appears (Zhang, 2003). However, when vocals are accompanied by background music. it does not always hold true since the drum sound comprises high energy components when compared to vocals In addition, it is understood that analysis of a music signal in its time domain is not sufficient for accurate segmentation. Spectral analysis is also essential and is employed.

Different kinds of transformations, including Fourier transform (FT), are available to represent time domain signal in frequency domain. Note that the most of the energy of vocals formant falls in the frequency range of 200 Hz. to 2000 Hz. Therefore, suppressing other frequency values helps to locate the singing voice segments. This can be done by using any of the available infinite impulse response (IIR) filters such as Butterworth, Chebyshev and so on (Kim and Whitman, 2002). This approach can be used to separate the background accompaniment, that helps in locating the vocal segments easily. However,

Table 2.2: Summary of works on vocal and non-vocal segmentation. *(Note: Only some relevant and widely cited articles are listed).*

Sl. No.	Title of the article	Composition of Database	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
1	Artist detection in music with Minnowmatch Whitman et al. (2001).	82 clips (male and female)	FFT values and MFCCs	85.10	Two classifiers namely SVM and ANN, are used for segmenting singing voice segments.	As FFT gives good discrimination for vocal and non-vocal portions, statistical operations on FFT values may improve the accuracy.	Accuracy of the system comes down with the increase of database.
2	Singer identification in popular music recordings using voice coding features Kim and Whitman (2002).	20 full-length songs	Chebyshev-IIR and Harmonicity	55.40	Chebyshev-IIR filter is applied to enhance vocal regions and attenuate other frequency regions. Later, harmonicity is applied to detect singing voice segments.	Frequency analysis of singer and non-vocal regions may improve the accuracy of detecting singing voice locations.	It is assumed that formant energy always falls below 4 KHz. Due to the advancements in technology and music rendering, distinguishable/useful formants may be extracted up to 12 KHz.
3	Automatic singer identification Zhang (2003).	English and Chinese clips	Energy, ZCR and SF	70.00	A sudden increase in the value can be observed for specified features when singing voice starts.	Identifying similar kind of time-domain features may reduce the complexity issues	The sudden change in the specified values can be found in case of pure vocals. As the background accompanies vocals in a majority of vocals, the approach could not be practical.
4	Singer identification based on vocal and instrumental models Maddage et al. (2004).	110 tracks (English and Chinese)	OSCCs	83.58	Inter-beat frames are considered instead of fixed size frames to compute cepstral coefficients and named them as OSCCs. Better performance is observed with OSCCs when compared to traditional MFCCs.	There is a need to develop a system that can divide the signal into variable length frames instead of shorter and fixed length frames. It may be helpful in reducing complexity issues.	The proposed system is not suitable to identify all the vocal and non-vocal regions. The OSCCs may confuse to segment using inter-beat segmentation due to vocals involvement.
5	Singing voice separation from monaural recordings Li and Wang (2006).	Popular English songs	Intonation and Viterbi algorithm	89.44	Inverse comb filtering is applied to reduce the background accompaniment and later, vocal frames are identified by observing high energy levels when vocal region starts.	A thorough analysis of filtering techniques may help in reducing the background accompaniment which further helps in properly detecting the vocal onset detection.	The dataset contains very few songs and may not be sufficient to generalize the results.
6	Automatic singer identification based on auditory features Cai et al. (2011).	140 clips (English)	MFCCs	92.10	At first step, low-pass filter is applied to suppress background accompaniment. Sparse representation classifier (SRC) is used to locate the vocal segments. MFCCs are used as features	Reduction of background score and enhancement of the singing voice may increase the performance.	The detailed explanation is not found on using the SRC classifier.
7	Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations Murthy and Koolagudi (2015).	300 clips (small and longer) clips	MFCCs, stat{pitch} and vibrato	87.05	Baseline MFCCs, statistical values of pitch and vibrato features were used to observe the variations in vocal non-vocal regions.	Signal level analysis on popular songs may give some repeated patterns that may be helpful in locating singing segments.	It is observed that the computational complexity increases if the clip length is longer.
8	A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks Lehner et al. (2015).	149 clips	Fluctrogram Analysis and spectral features	89.06	Fluctrogram is introduced to compute the pitch, and available spectral features such as MFCCs, Spectral contraction and flatness are added to improve the performance.	Proper analysis on fluctogram may give suitable temporal features that help in improving the accuracy of vocal onset detection.	For experimentation, only a single genre is considered, which is not sufficient to rely on the approach.

it is very difficult to observe the frequency range of vocal and music in a mixed clip (Vembu and Baumann, 2005). The change in the shape of a spectrum of vocal and non-vocal regions has created much research interest in recent times. By analyzing the spectrum, formants and MFCCs are computed and used to understand the characteristics of vocal and non-vocal segments (Regnier and Peeters, 2009).

In some works, it is found that *fluctogram* gives more information when compared to a spectrogram. The sub-semitone and pitch-continuous fluctuations can be viewed using a simple cross correlation followed by shifting operation. The resultant of this operation gives a new visual representation named *fluctogram*. Features based on the *fluctogram* have been extracted; however, very little effort has been made in this direction (Dittmar et al., 2015; Lehner et al., 2015). Prominent human formant values are mainly observed in the range of 2 - 3 kHz. Basic cepstral features such as MFCCs also carry important music/vocal information (Cai et al., 2011; Li and Wang, 2006; Mesaros et al., 2007). While computing MFCCs, the length of the frames is always fixed. Frames of variable lengths are introduced (Maddage et al., 2004) based on inter-beat-times to improve the performance of a system, known as OSCCs. The results convince that the OSCCs are more suitable for music modeling than the MFCCs. Frequency analysis along with temporal behavior is considered for vocal characterization by using Δ (velocity) and $\Delta\Delta$ (acceleration) features of MFCCs. Similarly, there are other features found in literature such as Δlog energy, modulation energy, harmonic coefficients (HC), and Δ MFCCs for locating the singing voice (Chou and Gu, 2001). Vibrato in the singing voice is also useful in locating the vocal segments efficiently (Murthy and Koolagudi, 2015; Mauch et al., 2011). The trending deep neural networks (DNNs) are also used in some works to separate the source information (Simpson et al., 2015). Table 2.2[7] summarizes the research contributions to locate the singing voice segments along with their limitations and scope of improvement.

In majority of the cases, the task of locating singing voice segments has been considered as a sub-task for singer identification. It is true that the small portion of the singing voice is enough for such tasks. A few works have concentrated on segmenting the complete music that may be helpful for the applications like *Karaoke* (Shenoy et al., 2005; Murthy and Koolagudi, 2015). Since the high dimensional feature vector takes more computational time for segmentation, feature dimensionality reduction is also a necessary task. Some works have concentrated on optimizing the features using feature selection algorithms

[7]Expansions for the acronyms are given in Appendix.

that help in selecting the suitable features for locating vocal segments (Ramona et al., 2008). Nevertheless, there is a good scope for an accurate system that segments the vocal and non-vocal regions in a given song clip of any kind.

2.4 Artist Identification

Artist information is one important attribute available within a music clip. Singer identification, recognition of composer, and artist identification of a concert are the variations in artist identification. Majority of the times, it is possible to observe the unique styles (singing/ performing or writing/ composing) of the artist while they are performing. Through the implicit learning capability of humans, the differences can be discerned by listening to a sample audio clip (Fu et al., 2011). If a person is familiar with a specific singer's tone, it is possible to recognize the singer by a small piece of the audio clip. At present, music stores are utilizing the efforts and expertise of music professionals to label the singer information for the unknown songs of their music databases. However, it is practically difficult to manually label millions of tracks available in the digital market manually, and sometimes, it becomes unreliable. The complex audio signals do not give any singer specific information by simply looking at them (Kim and Whitman, 2002). The applications of automation of singer identification task include music recommendation, cataloging and indexing. It can also be used in issuing copyrights for tracks to avoid music plagiarism.

Singer identification is a $one - in - n$ class classification problem as it deals with identifying a singer among n possible singers. The difficulty is to handle large music database. In this scenario, "singer similarity" based approaches are more useful and suitable. In the mutual phase, similar singers may be grouped together by using clustering algorithms. One important constraint is that the singers maintain similar voice patterns and common characteristics while rendering songs although the occasion is different. The scope of this thesis is limited to the literature on singer identification.

Traditional speech processing techniques for speaker identification (Rabiner and Juang, 1993; Becchetti and Ricotti, 2008) may not be suitable for the task of singer identification (Tsai and Lee, 2011). In spontaneous speech, the pitch of a speaker involuntarily changes with factors such as emotion, loudness, and so on. Whereas in singing, controlled pitch modulation is necessary for melody. Singers are trained to vary pitch while rendering music and have control on vocal parameters including respiratory system, laryngeal muscle

activity, articulation, and so on. In simple terms, singers are trained to vary the vocal parameters systematically; this gives evidence to recognize the singers through the analysis of voice parameters (Björkner, 2006; Shen et al., 2009).

In the literature, several techniques have been proposed on singer identification (Kim and Whitman, 2002; Shen et al., 2009; Zhang, 2003; Patil et al., 2012; Mellody and Wakefield, 2000; Tsai et al., 2008; Zhang and Packard, 2003; Maddage et al., 2004; Liu and Huang, 2002; Tsai and Wang, 2006) and artist identification (Berenzweig et al., 2002; Whitman et al., 2001; Kim et al., 2006). Some of the important approaches for singer and artist identification given in the literature are presented below:

In many works, MFCCs are used as base-line features for singer modelling as they are already well-established features for speaker identification (Reynolds, 1994; Logan et al., 2000). Compared to speech, music contains more high-frequency components (many instrumentals) in the frequency range of 200 to 15000 Hz. To have the expected soothing effect, the music signal is maintained at a very high sampling frequency (above 40 KHz). A slight modification in MFCC extraction process produces tweaked MFCCs, which are used for singer identification by using complete frequency bandwidth (up to 22000 Hz.) (Whitman et al., 2001). Cepstral Mean Subtracted MFCCs (CMSMFCCs) have been proposed to improve the classification accuracy as they can capture the variations among singers (Patil et al., 2012). These features are computed by subtracting the cepstral mean from each vector of MFCCs. Moreover, the temporal behaviour of MFCCs is considered to study the singing pattern variations among singers through Δ and $\Delta\Delta$ MFCCs (Berenzweig et al., 2002). OSCCs have also been proposed for singer identification, where the cepstral features are computed on frames of variable lengths (Kim and Whitman, 2002; Zhang, 2003), which helped to characterize the harmonic structure of a singer. To compute OSCCs, framing is done based on inter-beat duration rather than traditional fixed length frames.

In general, specific vibrato and pitch profiles are followed by the singer while performing (Sundberg and Rossing, 1990). Therefore, features that resemble human perception have a high role in many music processing applications. One such approach is warped linear prediction (WLP), where all coefficients are extracted at warped scale (Strube, 1980; Harma and Laine, 2001). A warped scale is closely related to the logarithmic one and highly resembles the functioning of a human ear. Hence, warped linear prediction coefficients (WLPCs) are used (Kim and Whitman, 2002) to recognize the singer successfully.

The results of the above works convey that the WLPCs exhibit better singer characterization as compared to the conventional LPCs. In general, the same kind of instruments are used to provide the background to the singer while he/she is performing in concerts. Hence, the performance of the singer identification may be improved with the combination of non-vocals instead of vocals alone. In some works, the LPCs are utilized to dense the cepstral coefficients for the task of singer identification (Zhang and Packard, 2003). From the literature, it may be observed that warped LPC based cepstral coefficients can be explored further for singer identification.

Primarily, the following points are to be considered while developing an application for singer recognition. Commercially available audio files are always accessible in compressed formants (*e.g.* MP3), whereas a majority of the works in the literature are experimented on raw files (*e.g. wav*). MPEG Audio Layer-3 (MP3) is one of the techniques used to compress the audio files. Identifying and extracting the features from MP3 clips help in designing a real-time system for music processing. A few works are only reported in which features are directly extracted from MP3 clips (Liu and Huang, 2002). New features and approaches are essentially required to extract the singer relevant information from MP3 clips. Another important issue in singer identification is locating multiple singers and identifying the overlapped regions. Existing systems are helpful in characterizing and recognizing a single singer. Many a times, the length of duets and trios is much more than that of the solo regions in the song. Thus, there is a need for the approaches to recognize multiple singers, track the location of singers, etc. This approach is helpful to those who are learning to sing songs on empty (vocals absent) tracks (Kim and Whitman, 2002; Tsai and Wang, 2004; Fujihara et al., 2005; Tsai et al., 2008). A summary of the literature with their limitations and scope in artist identification is depicted in Table 2.3.

Singing voice mostly occupies a place between the dominant musical instrument and speech (Mesaros and Astola, 2005b,a). The spectrogram of a singing voice reflects vowels with a harmonic structure. Hence, the harmonicity helps in recognizing the singer from a given clip. At the same time, the features based on articulatory techniques are also helpful in determining the singer as they outperform in speaker identification tasks (Loui, 2015). The above statements hint at combining music and speech related features to improve the singer recognition accuracy. Considering the fewer efforts, singer identification has to be explored with wider dimensions at least in the context of Indian music. Singing quality of an artist has a direct correlation with one's timbre. Hence, estimating the timbre will

Table 2.3: Excerpts of the articles published on the issue of Artist Identification. *(Note: Only some relevant and widely cited articles are listed).*

Sl. No.	Title of the article	Composition of Database	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
1	Artist detection in music with Minnowmatch. Whitman et al. (2001)	82 clips (male and female)	FFT values and MFCC	85.10	Artist classification is done with two classifiers. SVM gives good performance when compared with NN for more artists.	Artist's timbre can be detected using the statistical operations on FFT.	Database with fewer artists gives good accuracy.
2	Singer identification in popular music recordings using voice coding features. Kim and Whitman (2002)	NECI Minnowmatch testbed	LPC and WLPCs	45.30	Warped scale is introduced and combined with linear scale to extract LPCs.	Features that are extracted using variable-length frames and perceptual scales may be helpful in developing real-time systems.	A little bit of improvement is found when compared to traditional LPCs. However, the mentioned performance may not be sufficient to standardize the system.
3	A singer identification technique for content-based classification of MP3 music objects. Liu and Huang (2002)	200 clips (male and female)	PMCV and FMCV	66.00	DCT is applied on frames of MP3 clips and named as MDCT. Phone and frame level features are extracted for experimentation.	Feature extraction on MP3 files (compressed) may be useful for the tasks of MIR which is to be thoroughly explored.	The database with few clips has been considered for experimentation and less accuracy is observed.
4	Automatic singer identification. Zhang (2003)	45 (English and Chinese) clips	LPC and MFCCs	80.00	Singing voice locations are identified automatically. Further, GMM classifier is used to classify the singers.	Increase in database size and understanding the voice qualities of singers may be helpful for singer identification.	Database size is very small and it may difficult to model all modulations of singers' using it.
5	Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. Tsai and Wang (2006)	260 (solo and duet tracks)	MFCCs	82.80 (solo)	Solo and duet clips are considered to extract multiple singers' information.	The system may be extended to locate the singer information and track the singer.	Performance of locating target singer and tracking target singer is not as per expectation.
6	Automatic singer identification based on auditory features. Cai et al. (2011)	140 clips	MFCCs, LPCCs, and GTCCs	90.00	Combination of three cepstral features are used to improve the performance.	As cepstral features are highly correlated to human perception, they can be used to characterize the singer.	It is observed that the performance gets degraded with the increase in the number of singers.
7	Combining evidences from mel-cepstral features and cepstral mean subtracted features for singer identification. Patil et al. (2012)	500 (14 M and 6F)	MFCCs CMS-MFCCs	84.50	Cepstral mean is subtracted to observe the temporal variation of singers' information.	Temporal fluctuation estimation may be helpful to identify singer	Vocal locations are manually marked, which may not meet the real time applications.

benefit the task of singer identification. Moreover, the vocal tract, and excitation level features along with rhythmic features of a performer may further be useful in detecting the singer more accurately.

2.5 Music Mood Estimation

Music mood[8] estimation based on music patterns is another important aspect of CB-MIR which helps to recommend or fill the playlist based on users' emotional needs. The aim is to categorize the songs based on emotional patterns such as happy, angry, sad, and so on. Emotions are difficult to process because of inherent complications. It is impractical to compare the performance of the systems due to lack of benchmark dataset. Recently, MIREX has created a standard dataset that is used to check the reliability and effectiveness of the works received for their competition. However, the dataset is not generalized to cover all important categories. The effort to create a benchmark dataset in the context of Indian popular music is almost nil. Hence, there is a need to create a standard dataset and develop an approach, which can classify music based on emotions. The task of mood estimation is highly ambiguous due to many psychological aspects related to the emotions of a song. In the literature, some approaches such as *Thayer's* model (Thayer, 1989), *Hevner's* model (Hevner, 1936), and TWC model (Tellegen et al., 1999) have tried to address the issues in emotion processing. All these models are designed by expert psychologists and used by various MIR scientists; however, these models lack the support of listeners. In the recent works, listeners' opinion is collected through majority voting which makes an open ballot available to collect the options from a variety of users, and then a majority can be used to label the emotion of the song (Hu and Downie, 2007; Downie et al., 2008).

The features used in literature for the task of mood estimation are almost similar to those used for genre classification. From the analysis of the literature, it is clear that the low-level spectral features such as MFCCs, LPCCs, Δ features, etc., are helpful in estimating the mood of an audio clip (Li and Ogihara, 2006, 2003; Korhonen et al., 2005; Lu et al., 2006). Some experiments have also been conducted using rhythmic features for categorizing the emotions in music (Feng et al., 2003a). These features are combined with other low-level features to improve the performance (Chua, 2008). The perceptual observation that the smoothness is in placid emotions such as happy or sad is high compared

[8]The words '*mood*' and '*emotion*' are interchangeably used in this thesis.

Table 2.4: Excerpts of the articles published on the issue of Emotion Recognition from Music. *(Note: Only some relevant and widely cited articles are considered).*

Sl. No.	Title of the article	Composition of Database	Emotional Classes	Feature(s)	Accuracy %	Approach	Future Scope	Limitations
1	The 2007 MIREX audio mood classification task: Lessons learned Downie et al. (2008).	600 Clips	Five emotions	Temporal, Tonal and Loudness	52.65	Human-based classification is done and later compared with the system performance.	Analysis of the clips for all categories of emotions with mean opinion score may give better reliability.	The performance of system is not up to the mark with specified features.
2	Music mood and theme classification - a hybrid approach Bischoff et al. (2009).	*Allmusic.com* and *Last.fm*	Four moods and Four themes	Audio features	62.50	The hybrid approach is proposed to group the songs based on the emotions in them.	The theme and mood hierarchy is not standardized. Generalized hierarchy may be helpful to classify songs.	The process of feature extraction and feature selection is not explained properly.
3	SMERS: Music emotion recognition using support vector regression Han et al. (2009).	165 Clips	Eleven emotions	Scale, Energy, Rhythm and Harmonics	94.55	Eleven emotions have been classified with the support of support vector regression (SVR).	Proper selection of perceptual features may be useful to detect the emotions in a better way.	SVR is showing better performance compared to other non-linear classifiers. However, task related features are to be selected.
4	Lyric-based song emotion detection with affective lexicon and fuzzy clustering method Hu et al. (2009).	981 Chinese Clips	Valence and Arousal	NLP and Fuzzy clustering	60.38	NLP is applied to recognize the words and distribution among valence and arousal is done using fuzzy clustering.	Extracting lyrics may be helpful along with the support of signal processing approaches for mood estimation.	Fuzzy clustering alone is considered, which may be the reason for less accuracy.
5	Music emotion classification and context-based music recommendation Han et al. (2010).	120 Clips	Eleven emotions	Low-level features and COMUS Ontology	61.80	COMUS is used an ontology to estimate the users' present emotional state based on past behavior and low-level features that are applied for song mood estimation.	Multi-mood estimation may be possible with low-level features as a song contains more than one emotion.	The system gives less accuracy due to improper estimation of users' mood. Moreover, low-level features alone may not be sufficient.
6	An approach of genetic programming for music emotion classification Bang et al. (2013).	488 western clips	Five Emotions	Timbre, Tonality and Chord	74.4	Two-level classification is applied to identify class of emotion and actual emotion.	A light is to be thrown on evolutionary approaches to reduce the complexity issues and increase the performance.	The performance of system gets degrading when the number of classes is increasing.
7	Audio songs classification based on music patterns Sharma et al. (2016).	300 clips	Seven Emotions	MFCC, stat{Pitch} and Vibrato	82.00	Modulated features are used to detect emotions and mean opinion score is collected.	Consideration of vocal and non-vocal regions may improve system accuracy.	Increase in database may reduce the accuracy because the features may not be sufficient to discriminate emotions.

to strong emotions like anger (Mion and Poli, 2008). To estimate the smoothness among the changing multiple sounds, articulation-based features, are more useful (Chua, 2008). Some experiments are conducted on tempo-based features, and the tempo of angry clips is faster than that of placid emotions. This analysis supports the necessity of the use of articulation and rhythmic features for mood estimation.

The process of low-level feature extraction demands more time as these features are extracted at every 20∼40 milliseconds. To resolve this problem octave-based spectral contrast (OSC) is explored (Jiang et al., 2002). These features are extracted at every spectral sub-band instead of fixed small-length segments known as frames (Jiang et al., 2002; Lu et al., 2006; Yang et al., 2008). The results of this approach convince that the OSCs are better than traditional MFCCs for music mood estimation. Moreover, the emotion classification in music is multi-label learning (MLL) problem because, a song may contain more than one emotions in it (Li et al., 2003). To address this issue, sophisticated algorithms are introduced with the support of kNN classifier (Trohidis et al., 2008). The literature also reports the efforts to identify the mood of a song based on the instrumental region (Agarwal et al., 2018). In fact, the emotion of a song clip can be recognized by focusing on vocal as well as non-vocal regions. Table 2.4 gives some overview of the existing literature with possible future directions. Based on the literature, the development of a reliable system for emotion recognition from songs based on the analysis of both vocal and non-vocal regions may give better performance.

2.6 Music Recommender System (MRS)

The process of generating a playlist with relevant audio clips for the listener is called music recommendation. Based on the literature, MRSs are categorized into three classes on the basis of collaborative filtering, content-based recommendation, and hybrid recommendation.

Collaborative filtering does not depend on the audio content of the song. They monitor the listener's previous profile in order to extract the information, such as ratings of audio clips, listening statistics, sequence of songs listening and other similar behaviour (Celma and Serra, 2008; Jawaheer et al., 2010; Levy and Bosteels, 2010). Further, the semantic tags such as artist information, song title, genre, etc of audio clips are considered for music recommendation. The information about the tags has been taken from social tagging websites (Celma and Serra, 2008; Herrada, 2009; Schedl et al., 2011). In

the case of popular songs The approaches used in collaborative filtering are successful in predicting the relevant tracks for listeners. However, their performance is found to be miserable in the long tail for unpopular songs. The reason could be the "cold-start" problem due to the unavailability of user ratings and meta-information in the beginning (Celma and Serra, 2008). Content-based music recommendation can overcome this problem (Hyung et al., 2014; Bogdanov et al., 2013). In general, the which are content-based recommendation approaches construct a feature vector using a timbral, tonal, temporal, and/or high-level information from the audio clips. Further, some similarity measures are considered to find the distance among the audio clips that give a report about the list with similar nature (Bogdanov et al., 2010; Pohle et al., 2009). This approach is similar to query-by-singing/humming (QBSH). The success rate achieved by MIR community using content-based recommendation is high. However, there is a concern about computational complexity as each clip has to be compared with all others to obtain the similarity measures (Downie et al., 2010). Further, the research has been focused on hybrid recommendation which is a combination of both collaborative and content-based recommendations (Su et al., 2010).

The task of music recommendation with content-based approaches has coaxed with the issues of computational complexity. Hence, an effective collaborative recommendation system with semantic information that can overcome the issue of "cold-start" problem is essential.

2.7 Research Gaps

- Sophisticated algorithm that separates vocal and non-vocal segments is not implemented yet despite being an important prerequisition for implementing many other issues of MIR.

- Lack of standard datasets for real-time application development is the other issue. Creating a generalized dataset for varied requirements of MIR is an impossible task. However, it is possible to collect an exhaustive dataset for a particular regional audio clips.

- Based on the literature, it is found that the work that has been done on Indian music is very meager. except some works on *raga* identification. Since contribution of Indian music to the digital cloud is high, there is an essence to focus on Indian

music in all aspects.

- The task of vocal and non-vocal segmentation is considered as a singing voice detection in most of the existing literature works. There is ample scope of research for developing an efficient system for vocal and non-vocal segmentation which should be independent of singer, music, genre, etc. Identification of task specific features for vocal and non-vocal segmentation is also an issue to be addressed in multiple dimensions with an acceptable depth.

- Source (instrumental sounds) separation is a challenging issue which is addressed in sleak. Though some efforts are made to separate the source from vocals, a sophisticated system is yet to be built. Separation of source gives clear vocal information which cut down the complexity of other MIR tasks.

- Majority of the experiments conducted on singer identification are mainly based on the studio recorded voice of a specific singer which may not suit in real-world environments. Along with singer identification, MIR systems have to address many other issues like processing background accompaniment, mood of a singer, age, gender, duets, chorus, and so on.

- There are many works reported on music mood estimation. However, the performance of automatic mood detection system is not commercially acceptable. The majority of the works have tried to categorize the songs into two basic moods namely valence and arousal. The human perception with respect to mood is quite different. There is a need of the system which is capable of characterizing and categorizing the music into 6-8 basic moods/emotions.

- The two existing approaches for music recommendation namely collaborative and content based filtering have their inherent drawbacks. The first one is suffering from recommendation accuracy & '*cold-start*' problem, and the other one is from computational complexity issues. Focusing on some other formal method based recommendation can resolve to some extent.

2.8 Problem Statement and Objectives

Based on the research gaps identified from the literature review, the research problem for this work has been defined as follows.

This work aims at designing a music information retrieval (MIR) system for Indian songs which extracts useful information such as locating singing voice, singer information, and mood of a music. Further, this work also proposes an approach for music recommender system (MRS) using graph structures.

I *To develop a module for the task of vocal and non-vocal segmentation:* Features based on formant structure have been proposed to discriminate vocals and non-vocals effectively. Genetic algorithm based feature selection has been proposed to extract the useful features and ignore the rest. $n-$point moving window has been proposed to accurately segment vocal and non-vocal regions.

II *To recognize a singer from a smaller snippet using feature-based and deep learning approaches:* An attempt has been made to classify and categorize 20 Indian singers. Two different approaches of which, one is based on features and the other is using trending convolutional neural networks (CNN) have been proposed.

III *To estimate the mood of a song clip using two-level classification approach:* Six different moods have been selected based on the analysis of standard models. A two-level classification model of which, one is categorizing the given audio clip into energetic and non-energetic and, the other classifies into actual mood using CNNs.

IV *To propose an approach that recommends songs using collaborative filtering by considering singer, emotion and user statistics as key parameters:* A database has been constructed based on the listening behaviour for around 100 users. A graph has been constructed to understand their listening patterns. Further, to recommend songs, graph similarity metrics have been considered.

2.9 Datasets Considered for this Thesis

Four different datasets have been constructed for different experimental setups with respect to each objective. The list of datasets and their details are listed in Table 2.5. It also contains the information about some standard datasets that are considered for comparison. Since the emotional and MRS database have been considered with respect to Indian music, It is difficult to find similar datasets for comparison. The clips considered for Indian database are taken from the two popular Indian cine industries named *Tollywood* and *Bollywood*. Care has been taken to include all varieties of information that

include various singers, genders, moods, instruments, and scenarios. The details of each dataset with possible information are given in the following points.

Table 2.5: The details of datasets considered in this thesis.

Sl.No.	Name of the Dataset	Details of the dataset	Standard dataset for comparison (if available)	Objective	Source of data
1	Tollywood and Bollywood Popular Songs (TBPS)	#Clips: 500 vocal & 500 non-vocal #albums: 100 #singers: 20	MIR-1K dataset	Objective - I	Audio CDs Music websites
2	Indian popular singers database (IPSD)	#clips: 100 clips for each singer #singers: 20 #Male: 10 & #Female: 10	artist20 dataset	Objective - II	Audio CDs Music websites
3	Moods of Indian songs (MIS)	#categories: two #moods: six hr #clips: 50 for each mood	–	Objective - III	Audio CDs Music websites
4	MRS database	Listener's behaviour Songs information Listener's information	–	Objective - IV	Constructed

1. *Tollywood and Bollywood Popular Songs (TBPS):* Two sets of audio clips have been considered for segmenting the vocal and non-vocal regions. One set contains trainings and the other set holds testing clips. The training set has been created with 500 vocal and 500 non-vocal clips of length 3∼5 seconds each. The test set has the clips of length 1∼5 minutes each. Combinations of {*vocal, non-vocal, vocal*}, {*non-vocal, vocal, non-vocal*}, {*vocal, non-vocal*}, and {*non-vocal, vocal*} have been considered while creating the test dataset. All the clips are sampled at 44,100 Hz and quantization is 16-bits.

2. *MIR-1K dataset:*Further, the comparison has been done with standard dataset called Multimedia Information Retrieval – 1K (MIR-1K)[9]. The standard MIR-1K dataset contains 1000 clips that incorporate both vocal and non-vocal regions. Majority of the works of well-known MIREX (Music Information Retrieval Evaluation eXchange) have utilized this dataset for experimentation. However, the songs of this dataset do not contain different variations. The concept of REpeating Pattern Extraction Technique (REPET) (Rafii and Pardo, 2013) has been applied to suppress the source information on MIR-1K clips, as they maintain notable background patterns.

[9]`https://sites.google.com/site/unvoicedsoundseparation/mir-1k`

Table 2.6: Details of the singers whose songs are collected and included in the proposed IPSD. *Note:* Gender: M → Male and F → Female. Language: T → Telugu and H → Hindi.

Sl. No.	Singer Name	Gender	Language	Sl. No.	Singer Name	Gender	Language
A	Arjith Singh	Male	Hindi	K	Mohammed Rafi	Male	Hindi
B	Asha Bhosle	Female	Hindi	L	Mukesh C Mathur	Male	Hindi
C	Geetha Madhuri	Female	Telugu	M	S. Janaki	Female	Telugu
D	K.J. Yesudas	Male	Telugu	N	S.P. Balasubramanyam	Male	Telugu
E	K.S. Chitra	Female	Telugu	O	S.P. Sailaja	Female	Telugu
F	Karthik	Male	Telugu	P	Shreya Ghoshal	Female	Hindi
G	Kousalya	Female	Telugu	Q	Sunidhi Chauhan	Female	Hindi
H	L.R. Eswari	Female	Telugu	R	Udit Narayan	Male	Hindi
I	Latha Mangeshkar	Female	Hindi	S	V. Ramakrishna	Male	Telugu
J	Mano	Male	Telugu	T	V.R. Ghantasala	Male	Telugu

3. *Indian Popular Singers Database (IPSD):* Twenty different singers from two cine industries have been selected. Around 100 clips are chosen for each singer from a variety of albums. The details of singers including their name, gender, and language are given in Table 2.6. Care has been taken to include different emotions as well. The length of each clip in training set is 3∼5 seconds. However, care has been taken while collecting test set which includes the clips of lengths 60s, 30s, 10s, and 5s. The reason for collecting different lengths is to estimate the effect of the performance of the proposed system on longer clips and shorter clips as well.

4. *artist20 dataset:* Efforts have been made to construct training and testing set for standard *artist20*[10] data set as well. In a majority of MIR tasks, *artist20* is highly used for the task of singer identification since 2005. Twenty different artists have been considered to build the *artist20* dataset with 1413 tracks in total. The songs have been taken from six different albums of each artist that include two genres (rock and pop) (Ellis, 2007).

5. *Music Mood Dataset:* It is very difficult to label the song clip with a particular mood due to ambiguities. In this work, certain analysis has been done on both *Russell's* (Russell, 1980) and *Thayer's* (Thayer, 1989) models and six different moods namely happy, anger, energetic, sad, devotional, and romantic are identified. They have been identified based on analysis in two categories called *Energetic* and *Non-energetic*. For each mood, around 50-100 song clips have been collected based on their availability. An opinion from music professionals and mean opinion scores (MOS) have been taken to label the audio clips. Care has been taken to include the different emotions of same singer in the database. As there is a lack of standard

[10] `https://labrosa.ee.columbia.edu/projects/artistid/`

datasets for the task of mood estimation, similar kind of Western datasets are not available for comparison.

6. *MRS dataset:* The set of 1000 song clips are chosen with complete details that include song name, album name, duration, singer name, and mood (*energetic* or *non-energetic*). The listening behaviour of 50 user's has been taken by collecting their login time, song name, duration, and frequency. For initial experimentation, these two datasets have been considered, as there is a lack of proper database from the popular Indian musical websites such as Gaana[11] and Raaga[12]. The details of the datasets for each objective are given in Table 2.5.

2.10 Summary

This chapter gives detailed literature on the works of MIR. Since the focus is on a few important aspects, such as vocal and non-vocal segmentation, singer identification, and music mood estimation, a complete literature on these aspects has been given. In addition to this, details on some standard datasets available for implementing various MIR tasks, and literature on music recommender system are also presented. Further, few research gaps that have led to the problem formulation are given along with specific objectives of the thesis. Moreover, the details of the datasets that have been considered for experimenting the objectives are also presented. Chapter 3 discusses on the approach proposed to solve the foremost objective, called vocal and non-vocal segmentation.

[11] `https://gaana.com/`
[12] `https://www.raaga.com/`

Chapter 3

Classification of Vocal and Non-vocal Segments

> *If I cannot fly, let me sing.*
>
> — Stephen Sondheim

3.1 Introduction

Though the number of digital tracks is huge, proper meta-information is not available to all songs/albums. It is hard to categorize/use the digital cloud adequately if proper meta-information is not available. A song is a combination of different portions that appear in a specific order namely intro, verse, bridge, chorus, and outro (Nwe and Li, 2007; Thomas et al., 2016). The song starts with an intro and ends with an outro. There is an involvement of several components that include vocals (singing voice with background accompaniment), non-vocals (pure instrumental sounds), and chorus (multiple singers voice) to make the song complete. The details of all these components have not been tagged to each of the existing tracks of digital cloud. It is important to annotate the necessary meta-information such as singer(s), instruments used, genre, language, emotion(s), composer, lyrics, and so on for a song clip. Hence, several efforts have been made by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) to build an efficient music information retrieval (MIR) mechanism through a centralized evaluation system named Music Information Retrieval Evaluation eXchange (MIREX)[1].

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

The ideal MIR system has to extract the meta-information automatically if relevant portion of a clip is given as an input. For instance, it is worthless to analyze the non-vocal portions of a clip for the task of singer identification. The process of labelling the vocal onset and offset points manually for a millions of tracks is impractical. Hence, there is a necessity to develop an automated system which can automatically segment the vocal and non-vocal regions.These segmented (vocal/non-vocal) tracks would serve as inputs to other MIR tasks.

3.1.1 Applications

The task of segmenting vocal and non-vocal regions in a music clip has many commercial applications. One such important application is *Karaoke*. The term *Karaoke* is termed from the combination of two Japanese words *Karappo* (means *empty*) and *Okesutura* (means *orchestra*). It means, *Karaoke* is an empty track without vocals. These tracks are highly useful for amateur singers to tune themselves to the track. *Karaoke* is also useful for the singers during their live performances (Berenzweig et al., 2002; Kim and Whitman, 2002; Zhang and Packard, 2003; Wang et al., 2003b, 2004). It is also possible to reduce the number of accompanying musicians while performing on stage by providing *Karaoke* tracks. In addition to that, the research on MIR will benefit greatly from the segmentation of vocal and non-vocal regions (McVicar et al., 2014).

3.1.2 Challenges

Since the nature of an audio signal is highly complex due to its stochastic behaviour and the involvement of multiple components, it is quite difficult to segment the vocal and non-vocal portions. It is easy to distinguish vocals from non-vocals if pure vocals are available but, it is rare to get pure vocals without background accompaniment. Hence, the complexity of the task increases making it difficult to determine the characteristics of vocals.

3.2 Research Gaps

- Time-domain features such as ZCR, SF, STE, etc., are used in literature for the task of singing voice detection. They may not be able to accurately detect the singing voice due to dominant background accompaniment.

- The majority of the features found in the literature are adapted directly from the influences of speech processing. For instance, features used for speaker identification are directly explored for singer identification. However, certain analysis has to be done to select suitable features for the task of music processing.s

- Lack of proper database for implementing a sophisticated system for vocal and non-vocal segmentation.

3.3 Proposed Methodology

The step by step process of the present work is depicted in Figure 3.1. This section details the process behind feature extraction, and classifiers configuration. The details of feature selection using GAFS is explained in subsequent section.

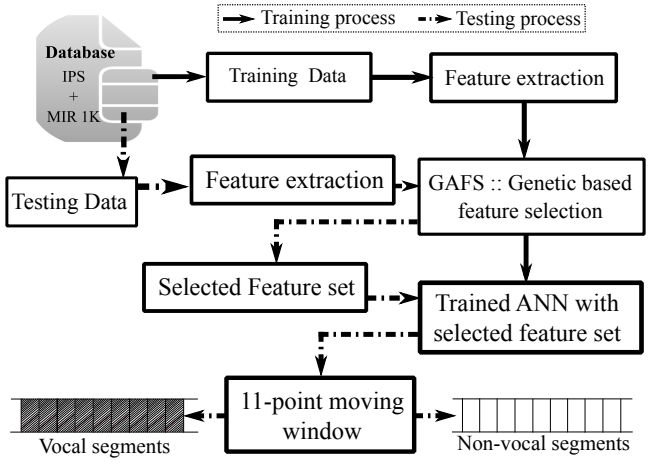

Figure 3.1: The proposed flow diagram for vocal and non-vocal segmentation.

3.3.1 Feature Extraction

Different features have been considered in this work. Important ones among them are MFCCs, LPCCs, FDLP, statistical values of pitch, jitter, shimmer, and formant based features. Following subsections discuss the process of extracting them one after the other.

A Mel-Frequency Cepstral Coefficients (MFCCs)

The popular and widely used features that played a major role in many speech and audio processing tasks are Mel frequency cepstral coefficients (MFCCs) (Logan et al., 2000; Ellis, 2007; Tzanetakis and Cook, 2002). The compact representation of spectral envelope provided by MFCCs is helpful in characterizing the speech and speaker information.

MFCCs also perform well in many music processing tasks (Eghbal-Zadeh et al., 2015). There are several variations for MFCCs such as velocity, acceleration, cepstral mean subtraction (CMS), and so on. However, standard MFCCs are considered for this work as better baseline performance has been obtained with them.

B Linear Prediction Cepstral Coefficients

The linear prediction coefficients (LPCs) are the coefficients of all-pole filter that can be used to model the vocal tract system. The process of computing LPC coefficients (LPCCs) is simple and can be done directly from the LPCs given below:

$$cep(m) = \begin{cases} 0 & ; m < 0 \\ \ln(G) & ; m = 0 \\ a_m + \sum_{q=1}^{m-1} \left(\frac{q}{m}\right) c(q) a_{m-q} & ; 0 < m \leq p \\ \sum_{q=m-p}^{m-1} \left(\frac{q}{m}\right) c(q) a_{m-q} & ; m > p \end{cases} \tag{3.1}$$

Where $cep(m)$ is cepstral coefficient, a_q represents linear prediction coefficients, and q is the order of prediction (Wong and Sridharan, 2001).

C Frequency Domain Linear Prediction based Features

Many transformation techniques such as Fourier transform, Wavelet transform, constant Q-transform, and so on are available to assist while converting the signal from time-domain to frequency-domain. The concept of linear prediction has been introduced because of its importance in many speech processing applications such as formant analysis, fundamental frequency estimation, spectrum and cepstrum analysis, source estimation and so on (Markel and Gray, 2013). The technique of linear prediction can capture temporal variations in the case of spectral domain. The work proposed by (Ganapathy, 2012), known as frequency domain linear prediction (FDLP), is designed to construct the short-time feature vector using DCT components that consume less time when compared to the other existing approaches. The FDLP coefficients perform well in many speech applications even the signal is supported with background random noise including reverberant speech. Since they can clearly capture the temporal variations, they are suitable for the tasks of MIR specially to discriminate vocal and non-vocal regions like other cepstral features. A 39-dimensional feature vector has been constructed for the present work.

D Statistical Values of Pitch, Jitter and Shimmer

The spectrograms and pitch contours of vocal and non-vocal portions are to be properly analyzed for the task of segmentation. It is also known that singers voluntarily change their pitch to provide a melody component while rendering music and the same may not be found with the non-vocals. Based on this observation, statistical variations in pitch such as minimum, maximum, standard deviation (σ) and mean (μ) have been estimated and analyzed with different correlations, histograms and GAFS. In addition, cycle-to-cycle variations of pitch and amplitude have been used, also known as Jitter and Shimmer respectively. The process of computing jitter and shimmer has been depicted in Figure 3.2.

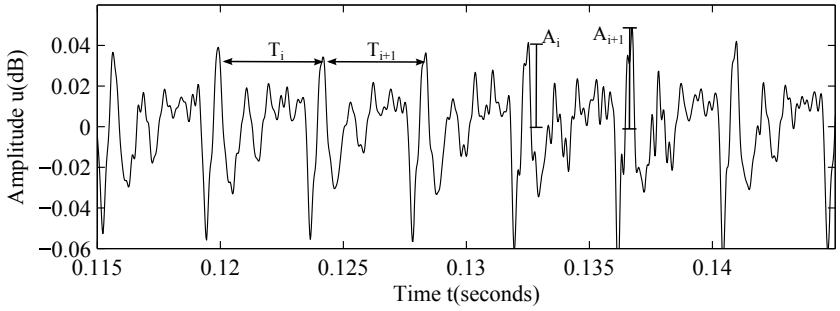

Figure 3.2: Jitter and shimmer computation from the speech signal.

The process of computing jitter and shimmer are formulated in equations 3.2 and 3.3 (Farrus and Hernando, 2009).

$$J = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \tag{3.2}$$

Where T_i is the extracted pitch period and N is the number of cycles considered.

$$Shimmer = \frac{1}{N-1} \sum_{i=1}^{N-1} |20log(\frac{A_{i+1}}{A_i})| \tag{3.3}$$

Where A_i is the extracted peak-to-peak amplitude data and N denotes the count of pitch periods.

E Formant Analysis

Formants are the resonance frequencies that can estimate the structure of vocal tract system. In this task, first four formants and formant energies have been used for dis-

criminating vocal and non-vocal regions. Formant frequency values and their energies are found to be incapable while discriminating vocal and non-vocal regions. However, there are some structural differences that are observed in the average formant structure of some vocal and non-vocal segments, shown in Figure 3.3.

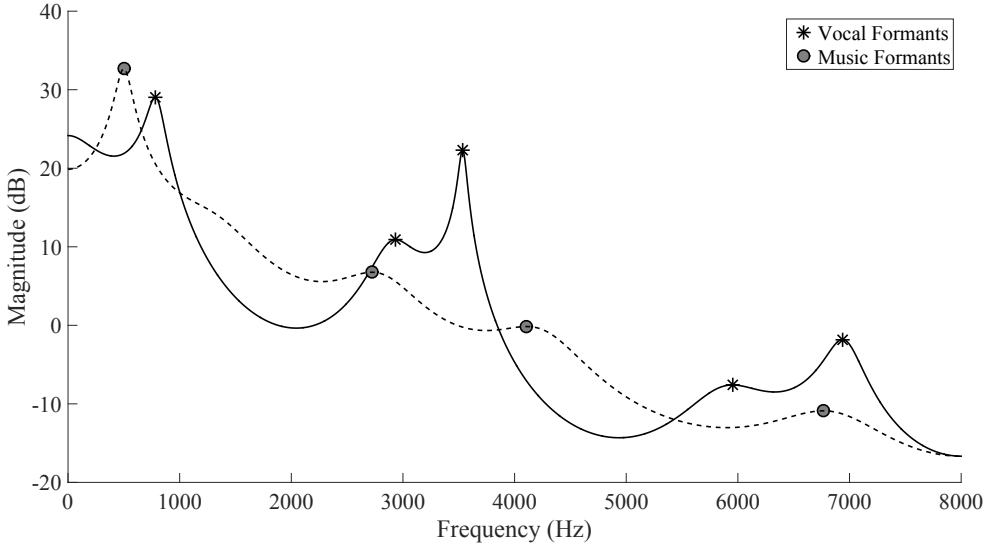

Figure 3.3: The structure of formant spectrum for vocal and non-vocal regions.

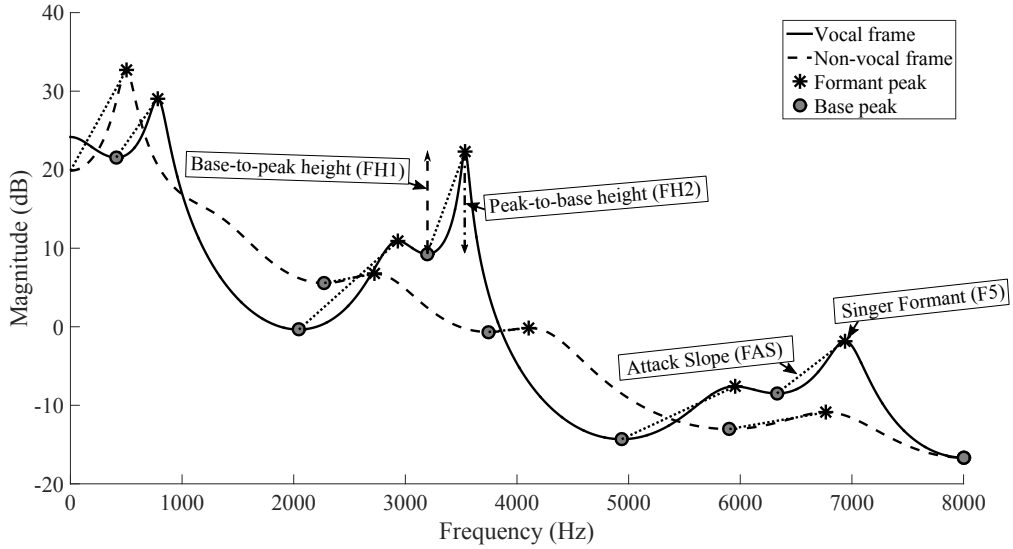

Figure 3.4: Features that are computed computed based on their discrimination for vocal and non-vocal segments. Since FH1 = FH2, only FH1 has been considered for experimentation.

Analysis has been carried out on the structure of formants and it is found that the attack slope, decay slope, and height of the peaks are quite different and found to be useful features for discrimination of vocal and non-vocal parts (shown in Figure 3.4). Hence, formant attack slope (FAS), and base-to-peak height of the formant (FH1) have

been computed for the regions of F2, F3, and F4, ignoring F1 since the values mentioned above are not much distinctive for F1. In addition to them, F5 is also considered since it is the singer formant and the value of F5 (Mendes et al., 2003) is almost null for the non-vocal segments.

Further, geometric methods are also applied on the formant structures of vocal and non-vocal regions to measure some discriminating parameters such as angle at peak (FA1) and valley (FA2). They are computed for second, third, and fourth formants as F1 is not much discriminating. The features that are extracted based on geometric methods are pictorially showin in Figure 3.5.

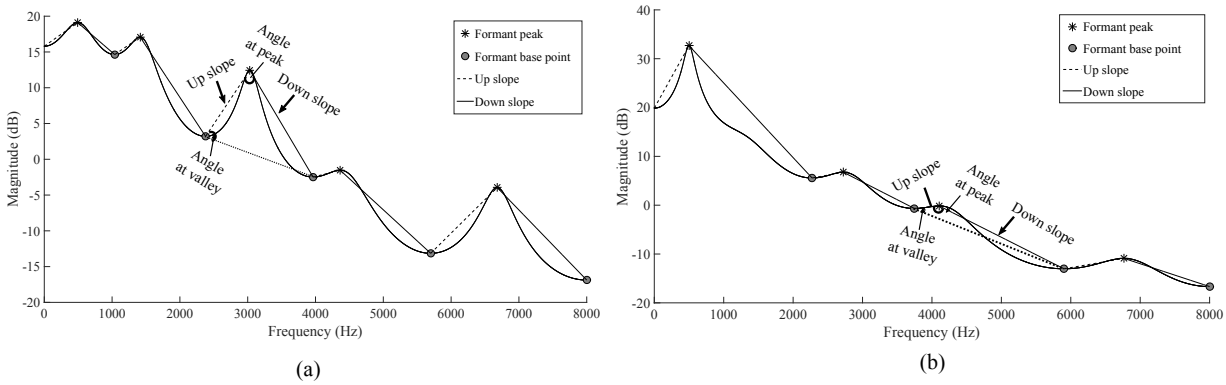

Figure 3.5: The process of computing angle values from the spectrum. (a) vocal spectrum and (b) non-vocal spectrum.

3.3.2 Classification Techniques

Four different classifiers namely support vector machines (SVM), Neuro-fuzzy classifier (NFC), random forest (RF), and neural networks (NN) have been chosen for the task of vocal and non-vocal segmentation. The configuration of each classifier and its importance are given below:

A Support Vector Machines (SVM)

The concept of statistical learning theory is the base for support vector machines (SVM). They focus on estimating the boundaries that separate the input feature space into two different classes (Vapnik, 2013). If the two classes are linearly separable, a decision boundary that separates the two classes at maximum will be chosen. A margin can be defined as the sum of the distances to the hyperplane from the nearest points of the two class labels (Vapnik, 2013). The technique of quadratic programming (QP) can be used to

solve the issue of margin maximization. The data points around the hyperplane (also called as *support vectors*) are considered to measure the margin. In the case of non-linear data, two important issues have to be addressed. One is margin maximization, and the other one is minimizing the misclassification errors. A user-defined parameter adjusts the margin and reduces the misclassification error (Cortes and Vapnik, 1995).

B Neuro-fuzzy Classifier (NFC)

Traditional pattern classification approaches involve the process of clustering the training samples and mapping test samples to the relevant cluster. Many times, the process of defining boundaries among the clusters is ineffective due to non-linearity in data. It would be difficult to have linearly separable boundary if the length of the feature vector increases. In contrast, the set of rules in fuzzy makes this task more straightforward and easy. It is easy to represent the high dimensional feature vector by mapping the relevant class using the set of non-deterministic rules (known as *fuzzy*) (Sun and Jang, 1993; Do and Chen, 2013) In the Neuro-fuzzy classifier (NFC), the input feature space is distributed to various fuzzy subspaces using fuzzy if-then rules. Further, these fuzzy rules are represented through a network structure. The network contains a multilayer feed-forward neural network structure including different layers namely input layer, fuzzy layer, fuzzification, defuzzification, normalization layer, and class label. A typical block diagram of NFC for a feature vector of length n and two output classes *vocal* and *non-vocal* is shown in Figure 3.6. The details of the membership function, fuzzification, and defuzzification are given below:

– *Membership Function:* In this layer, a membership function is identified for every input. There are several membership functions available such as triangular, trapezoidal, Gaussian, and so on. Since the data is normally distributed, a Gaussian function has been utilized and is defined Eq. 3.4.

$$\mu_{mn}(x_{kn}) = exp\left(- \frac{(x_{kn} - c_{mn})^2}{2\sigma_{mn}^2} \right) \tag{3.4}$$

Where $\mu_{mn}(x_{kn})$ represents the membership grade of the m^{th} rule and nth feature; x_{kn} represents the k^{th} sample and nth feature; c_{mn} and σ_{mn} represent the center and width of the Gaussian process,respectively.

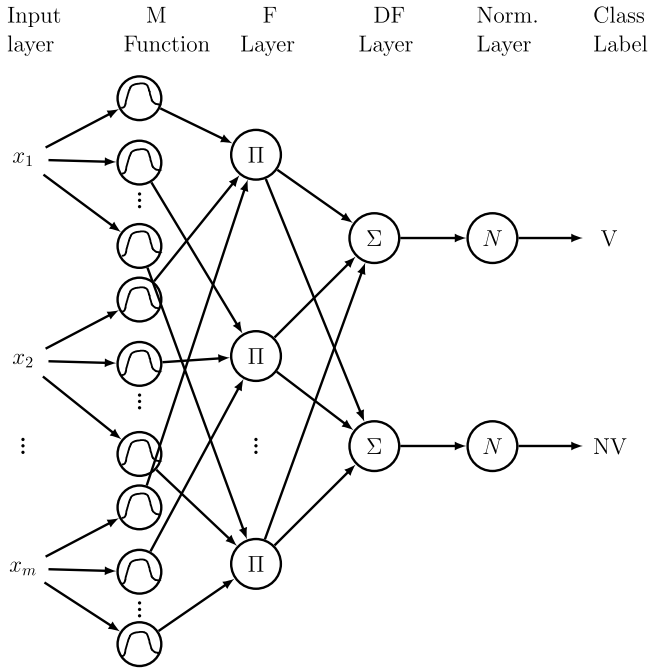

Figure 3.6: The structure of Neuro Fuzzy Classifier. *Note:* M Function → membership function, F Layer → Fuzzification layer, DF Layer → Defuzzification layer, Norm. Layer → Normalization Layer, V → Vocal, and NV → Non-vocal.

– *Fuzzification:* If the input sample satisfies the fuzzy rules, then a signal that resembles the degree of fulfillment will be generated by each node of this layer. In simple terms, a strength to the fuzzy rule is given by this layer and on the k^{th} rule:

$$\nu_{mk} = \prod_{n=1}^{N} \mu_{mn}(x_{kn}) \tag{3.5}$$

Where N represents the length of feature vector.

– *Defuzzification:* The weighted outputs will be calculated in this layer. If a rule of output is capable of controlling a particular class, then the weight is larger for that class when compared to the other classes. The process of computing weighted output for j^{th} class is given below:

$$\beta_{kj} = \sum_{m=1}^{M} \nu_{mk}\delta_{mh} \tag{3.6}$$

Where δ_{mh} represents the degree of the h^{th} class that is controlled by m^{th} rule and M represents the number of rules

Further, the normalization layer normalizes the weights as there is a chance of getting a value which is more than '1' due to summation. Finally, all the weights are compared

at the output class to identify the class label which is associated with maximum weight obtained.

C Random Forest Classifier (RF)

Random forest (RF) classifier is the composition of several decision tree classifiers. In which, each tree classifier has been built for a randomly selected feature subspace from the original input feature vector. The class which gains more votes from the set of trees is helpful in deciding the class label (Breiman, 2001). One important step in the process of RF classification is the attribute selection. There are two frequent methods used to select the attributes for decision tree induction namely information gain ratio criterion (IGRC) (Loh, 2011) and Gini index (Quinlan, 2014). Each combination of the feature vector constructed from the training data that generates a tree to the maximum depth. Since the RF classifiers are considering the attribute selection and not considering the pruning approach, their performance surely outperforms the traditional decision tree methods. However, the tree without pruning may not give overfitting issues because of Strong Law of Large Numbers (Feller, 2008) and capable of converging the error at some point. Hence, the effectiveness of the RF algorithm depends on the number of trees chosen.

D Artificial Neural Networks (ANNs)

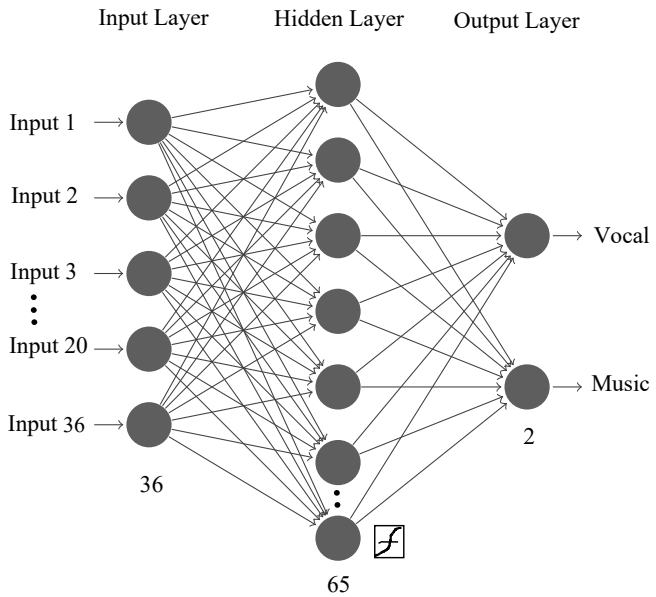

Figure 3.7: The structure of simple artificial neural network.

The process of selecting a classifier is always based on the kind of data chosen. It is also true that the features considered for this work are non-linearly related. It is hard to

discriminate them using a linear classifier. The efficient classifier that can handle highly non-linear data is artificial neural networks (ANNs) (Phillips et al., 2015; Wei et al., 2019). With this understanding, ANNs have been considered for segmenting vocal and non-vocal regions. The basics of ANNs and their variations can be found in many research articles (Jain et al., 1996). The structure of the three-layered ANN that suits the present task is shown in Figure 3.7. It contains input, hidden and output layers. It is empirically tested and found that the single hidden layer is enough to achieve better results as well as the model is computationally economical. However, the number of neurons in hidden layer plays a major role in better classification.

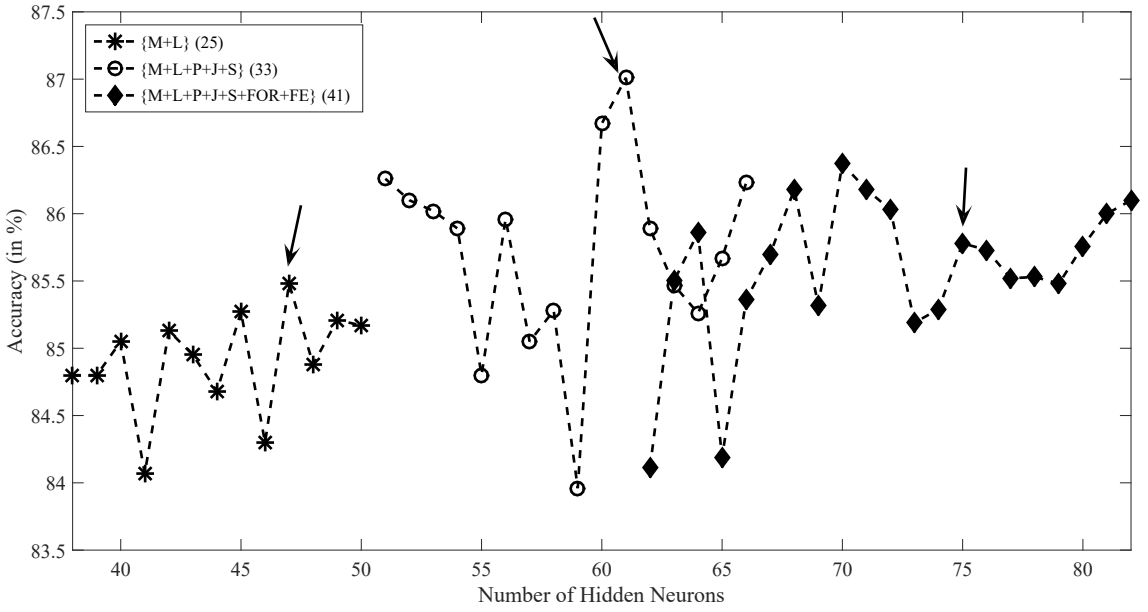

Figure 3.8: The accuracy obtained for varying number of hidden neurons. *Note:* Arrow indicates the best accuracy obtained for $N_h = \lceil (1.85 * I_n) \rceil$.

It has been stated that if the number of hidden neurons is in the range of 1.5 to 2 times of the input neurons, then they are able to classify more precisely (Boger and Guterman, 1997). Based on this, empirically the number of the hidden neurons is decided to be equal to 1.85 times the input neurons i.e. $N_h = \lceil (1.85 * I_n) \rceil$ where N_h is the number hidden neurons and I_n represents the number of input neurons. Performance of neural network model with different number of hidden layer neurons and feature sets is shown in Figure. 3.8. Arrow marks pointing to highest peaks in each category of features indicate the highest performance of the neural network model for that category of features.

3.4 Genetic Algorithm based Feature Selection (GAFS)

One of the revolutionary developments to solve and optimize the issues in computing world is evolutionary computation (Goldberg, 2006). A genetic algorithm (GA) is one of the evolutionary mechanisms which is flexible and works based on genealogy and natural phenomena (Kinnear, 1994). The genetic algorithm is designed with the principle of "*survival of the fittest*" proposed by Charles Darwin (Goldberg, 2006). Genetic algorithms are naturally intended to optimize the process, select the best solution, and discard the rest. They are also capable of providing efficient solutions to the problems. Similar to the other evolutionary algorithms, genetic algorithms are based on randomized operations. However, it is sufficient if one set of random values–called as the population of chromosomes–is generated. The remaining sets are produced from the primary set with the help of some primitive operations. Thus, they do not depend on the problem and are more capable when compared to conventional random & exhaustive search techniques (Kinnear, 1994). Moreover, genetic algorithms are found to be proficient and give best possible solutions even for the problems that do not have continuity, linearity, or other prominent information (Goldberg, 2006). The genetic algorithms have also several other advantages like providing the optimal solution, offering more than one solution, suitable for large search space problems where a large number of attributes present, etc. In contrast, genetic algorithms also have some drawbacks. One major drawback is that their unsuitability to solve small problems where size of the dataset is limited. Genetic algorithms are proven to be fit in many engineering applications such as structure optimization, control system optimization, and optimized filter design approaches. We also found that they are vaguely used in speech and music processing applications (Behroozmand and Almasganj, 2007). The advent of genetic algorithms has motivated us to use the same for selecting the appropriate features for vocal and non-vocal discrimination. We called this approach as the genetic algorithm based feature selection (GAFS). The ANNs are considered to evaluate the performance of every chromosome[2]. An example flow of GAFS-ANN approach is shown in Figure 3.9. Genetic algorithms have few natural operations such as *initialization, selection, crossover,* and *mutation* that are helpful to provide the optimal solutions. All the steps mentioned above are quite common to many applications except *selection* which varies depending on the problem. One complete iteration of GA process

[2]The terminology used in genetic algorithms for features can be referred in (Goldberg, 2006).

and its blocks are shown in Figure 3.9. The details of each block with relevant examples are given in the subsequent subsections.

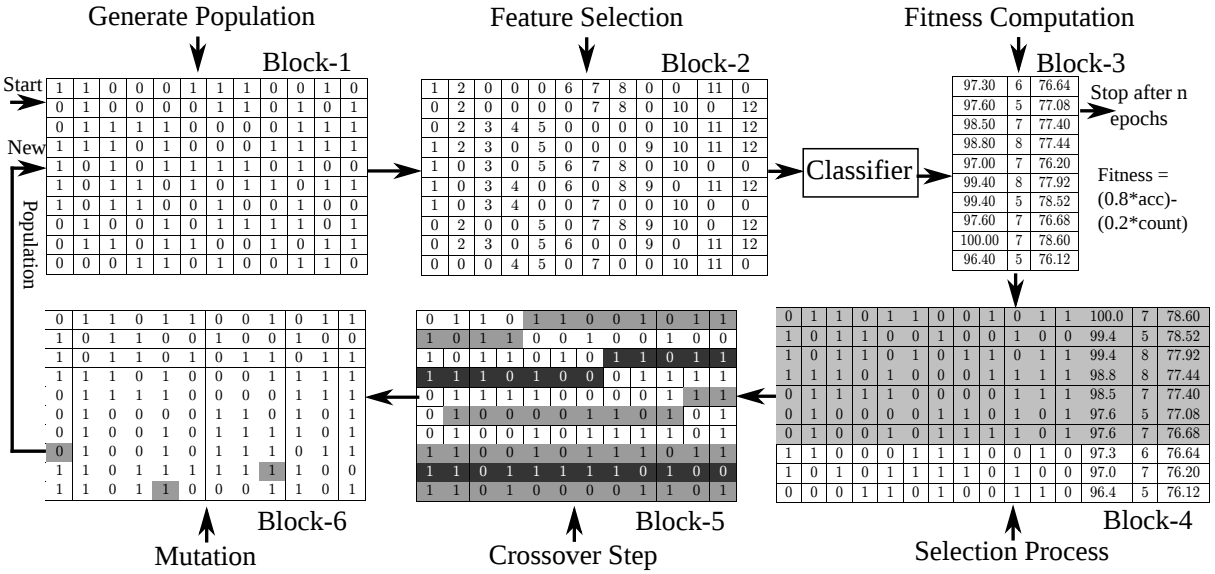

Figure 3.9: An example to illustrate the complete process involved in GAFS-ANN.

3.4.1 Initialization

The initial set of chromosomes has been generated using random number generator. A simplified example is shown in Figure 3.9 which selects the optimal set of features from the original feature vector of length 12. Each row represents 12-bit feature vector which is arranged as a 12 element tuple. The presence of $'1'$ indicates the selection of the feature and $'0'$ represents its absence in the final feature tuple. The process of choosing 12-bits is random in nature. Initially, three random numbers are generated in the range from 0 to 15. The binary encoding of a randomly generated numbers forms a chromosome of length 12. The number of random numbers depends on the size of the feature vector. Similarly, $'n'$ number of chromosomes is created to generate the initial population, the value of $'n'$ is 10 in the example shown. From the generated initial population, the features corresponding to the bit $'1'$ are selected, and the rest are ignored (shown in Block-2 of Figure 3.9). An example of generating chromosomes from the random values is shown below:

$$[\ 12, 7, 2\ ] \rightarrow [\ 1100\ 0111\ 0010\ ] \quad \rightarrow \quad \text{Chromosome A}$$
$$[\ 4, 3, 5\ ] \quad \rightarrow [\ 0100\ 0011\ 0101\ ] \quad \rightarrow \quad \text{Chromosome B}$$

3.4.2 Estimation of Fitness

Each generated feature vector by selecting features as shown in the previous step is given to the classifier. ANN is used as a classifier to test fitness of each selected feature vector. The performance accuracy of ANN classifier and the number of features in each chromosome have been considered to estimate the fitness value. The corresponding weight is given to both count and performance accuracy. The experimentation is done with different values, and better results are obtained with the fitness function proposed shown in Eq. 3.7.

$$fitness(i) = (0.8 * acc) - (0.2 * f_c) \tag{3.7}$$

Where i is the chromosome index, f_c is the count of features selected, and acc is accuracy obtained by ANN. The fitness values are shown in the 3^{rd} block of the first row of Figure 3.9. The third block contains three columns where column 1 represents the accuracy, the second column is the number of features, and the third column is the fitness values obtained.

3.4.3 Selection Process

A variety of selection techniques are available to preserve the finest chromosomes. Some of the widely used ones are tournament selection, roulette wheel selection, rank selection, proportionate selection, steady-state selection, etc. The tournament selection method is considered in the present work for the selection of features for further processing. The example explains the process of selecting the finest chromosomes based on the threshold cut-off on fitness values computed.

3.4.4 Crossover Operation

There are several forms of crossover techniques available in the literature. Some popular techniques are single-point, two-point, uniform, and arithmetic. The performance of each technique is dependent on the problem that was chosen. The process of generating child chromosome using each technique is detailed below.

A Single-point crossover

A single random number is generated for two chromosomes. Further, the values of chromosome A and B gets interchanged to create new offspring based on the random value. An example is shown in Fig. 3.10(a).

B Two-point and uniform crossover

Two or more random numbers will be considered to extract more characteristics of two parent chromosomes A and B for producing a new offspring. The examples for two-point and uniform crossovers have been illustrated in Figure 3.10(b) and 3.10(c) respectively.

C Arithmetic crossover

To examine the similar or dissimilar characteristics of two chromosomes A and B, the arithmetic crossover is popularized. The operation can be any of the AND, OR, or other operations. An example which illustrates the process of arithmetic crossover is shown in Figure 3.10(d).

3.4.5 Mutation

The mutation operation is preferred to enrich the qualities in the child. It is simply done by inverting the bit value or by interchanging the multiple bits within the chromosome. An example is shown in the last step of Figure 3.9.

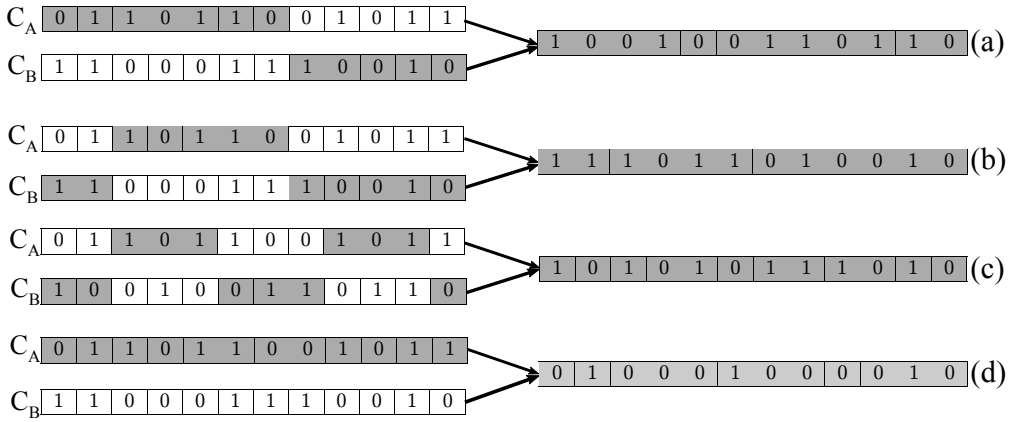

Figure 3.10: The various crossover techniques available with examples.

The above steps are performed iteratively till the optimal solution is obtained. It is decided that there is a less probability to produce better chromosome if the majority of the chromosome values are in steady state and.

3.5 Experimental Analysis

The results obtained using the proposed approach are analysed in this section. Two different datasets are used namely Tollywood and Bollywood popular songs (TBPS), and

MIR-1K are used for experimentation. The details of them are detailed in sections 2.1.1 and 2.1.2.

3.5.1 Result Analysis

A majority of the times, processing of time-domain signal alone may not be sufficient to discriminate vocal and non-vocal regions. Hence, some useful features have been considered from the frequency domain. The features such as MFCCs, LPCCs, FDLPs, statistical values of pitch, jitter, shimmer, and formants are extracted. In addition to them, a set of novel features namely formant height from peak-to-base (FH1), angle values at peak (FA1), and valley (FA2) have been computed from the formant structure after thorough analysis. These features been computed for second, third, and fourth formants since not much discrimination is observed in the case of first formant. Moreover, singer formant (F5) is also computed as it is rarely found in non-vocal segments and always available in vocal portions. Since the signal is stationary and useful to do certain analysis at lower frame lengths, a frame length of 25 ms. with an overlap of 10 ms. has been considered to compute all features mentioned above forming a 90-dimensional feature vector for a frame. The details of features are given in Table 3.1.

Table 3.1: Different features considered in this work with their acronyms and length.

Sl. No.	Feature Name	Acronym	Size
1	MFCCs	M	13
2	LPCCs	L	12
3	FDLPs	F	39
4	Stat{Pitch}	SP	4
5	Jitter	J	2
6	Shimmer	S	2
7	Formants and Formant Energies	FOR	8
7	Height from Peak-to-Base	FH1	3
8	Angle at Peak	FA1	3
9	Angle at Valley	FA2	3
10	Singer Formant	F5	1

The columns of the table represent the feature name, its acronym, and its size of the feature vector. From here onwards, in this section, the acronyms are to refer to the

features. The features all together form a feature vector of length 93. The possible feature combinations have been constructed to test their ability in segmenting the vocal and non-vocal regions and applied on both the proposed TBPS and standard MIR-1K datasets. The results obtained for the various combinations are given in Table 3.2.

Table 3.2: The accuracy of vocal and non-vocal segmentation obtained on the proposed and MIR-1K datasets using different feature combinations and classifiers. ***Note**: bold face letters indicate the best performance for that classifier and colored background represents the best accuracy for that dataset. SVM → Support vector machine, NFC → Neuro-fuzzy classifier, RF → Random forest, and NN → Neural network.*

Feature Set	**Accuracy (in %)**							
	Proposed TBPS Dataset				**MIR-1K Dataset**			
	SVM	**NFC**	**RF**	**NN**	**SVM**	**NFC**	**RF**	**NN**
{M}	53.68	64.05	72.54	70.12	65.52	63.82	73.82	72.69
{L}	52.96	63.16	68.29	69.64	63.98	65.75	72.96	70.83
{F}	48.34	53.42	53.12	59.54	52.84	61.49	68.45	61.86
{M+L}	54.25	64.95	74.92	71.56	66.12	66.72	72.83	73.66
{M+F}	56.72	62.24	71.54	72.39	56.21	59.13	70.09	63.98
{L+F}	55.02	59.39	65.07	61.82	54.25	60.29	68.45	64.82
{M+L+F}	58.66	63.56	63.14	60.47	57.57	58.37	73.58	66.56
{M+SP}	61.28	67.12	71.86	72.69	69.85	73.82	78.92	76.76
{L+SP}	59.85	65.63	69.99	72.08	72.35	75.19	77.45	75.45
{F+SP}	51.26	57.95	62.54	61.23	56.15	67.85	71.68	70.52
{M+L+SP}	62.38	67.32	74.87	71.69	76.48	76.32	76.12	77.39
{M+L+F+SP}	64.29	61.96	67.92	63.58	66.56	73.58	78.59	74.94
{M+SP+J+S}	66.72	69.52	75.84	73.54	80.36	81.69	82.98	83.67
{M+SP+J+S+FOR+FE}	59.15	57.26	62.87	61.54	70.24	69.23	72.58	71.96
{L+SP+J+S}	65.47	68.19	73.09	71.36	78.59	80.86	84.59	78.54
{F+SP+J+S}	56.12	60.04	69.57	66.21	61.28	70.23	77.25	73.96
{M+L+SP+J+S}	68.32	69.12	75.82	74.58	82.84	82.09	89.94	85.69
{M+L+F+SP+J+S}	69.69	63.46	73.23	68.48	73.96	78.96	83.58	81.23
{M+F5+SP+J+PFF}	72.98	73.87	75.98	78.96	78.35	85.29	91.65	87.23
{L+F5+SP+J+PFF}	71.59	72.45	78.35	77.58	82.56	86.56	90.28	85.96
{F+F5+SP+J+PFF}	64.18	61.73	73.36	68.33	73.69	79.15	86.68	79.54
{M+F+F5+SP+J+S+PFF}	73.54	**75.82**	**78.36**	**79.15**	**83.63**	**88.05**	**93.92**	**91.58**
{M+L+F+F5+SP+J+S+PFF}	**74.69**	73.93	76.19	75.78	79.56	82.19	89.27	83.62

From the table, it is observed that the feature combination {M+SP+J} is sufficient to classify the vocal and non-vocal segments of MIR-1K dataset. The reason could be

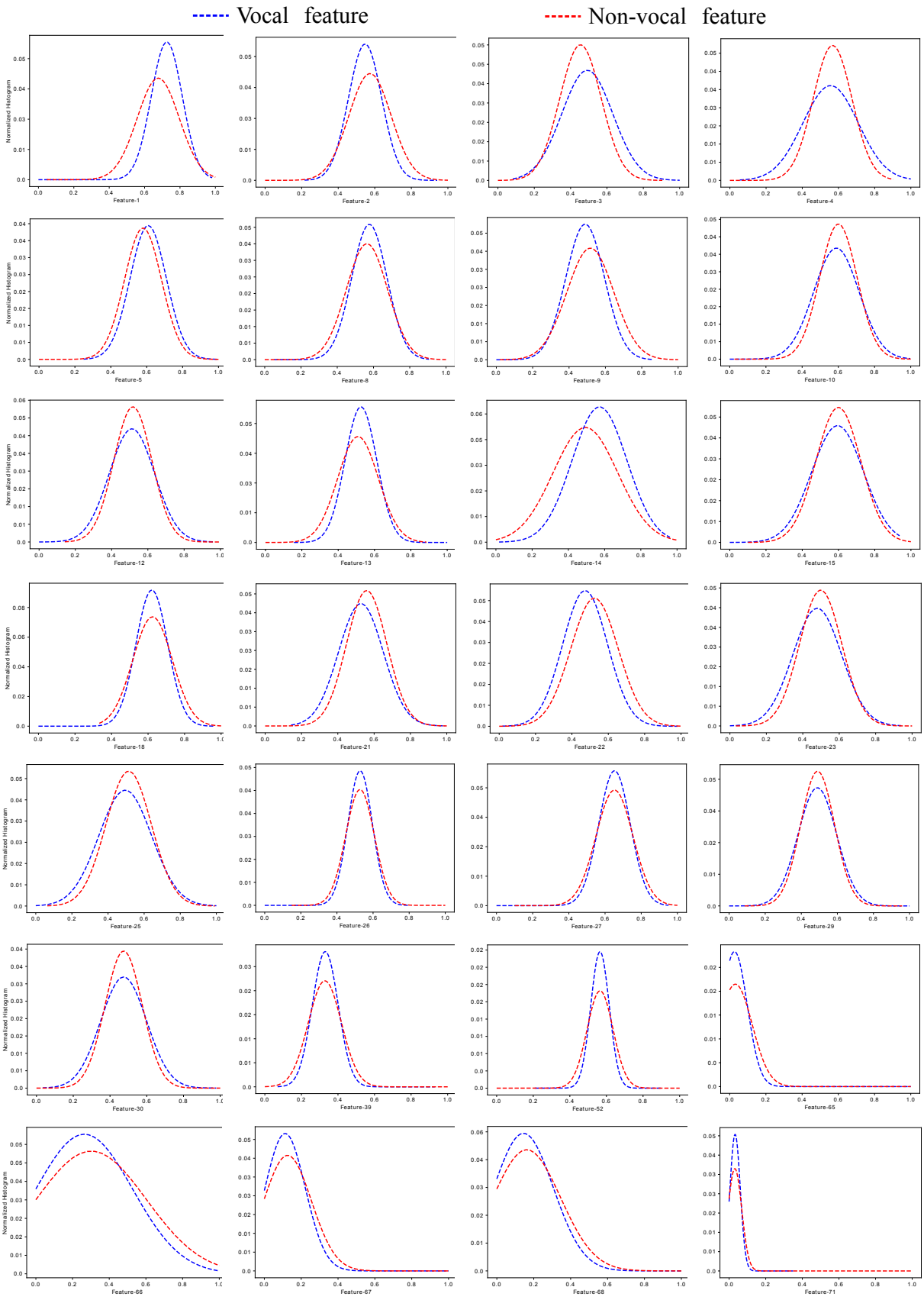

Figure 3.11: The Correlation values obtained with some features for vocal and non-vocal segments. *Note:* Only few are selected based on their discrimination.

the suppression of background accompaniment using REPET algorithm. Since the clips of MIR-1K are having less and rule-based background support, the source information can be easily separated. However, the accompaniment structure of the proposed TBPS dataset is quite different. Moreover, the composers always improvise the *staff*[3] in the case of Indian music (Samsekai Manjabhat et al., 2017; Fung, 1993). Hence, it is difficult to suppress the level of source information when compared to MIR-1K. The accuracy obtained with some feature combinations of 74 features is {M+F+F5+SP+J+S+PFF} is appreciable though the length is little high. However, use of such high dimensional feature vector leads to the computational complexity and is not suitable for real-time applications. Moreover, all features in the vector may not be relevant to the selected task leading to overfitting and underfitting issues. An approach to chose only relevant features from a set of features would help to overcome the problems specified above.

Various analysis have been done to sort out the specific features using correlation and feature selection algorithms. The correlation analysis has given the information about how far a specific dimension is distinct concerning the class. Two approaches have been considered to select the features based on correlation analysis. One is visual analysis using normalized histograms and scatter plots. The normalized histograms after curve fitting have been drawn for the two classes of each dimension and shown in Figure 3.11. Finally, the features that give at least a minimal discrimination are considered in the final list. From Figure 3.11, it is evident that considering all features reduces the performance of the system. However, some dimensions of features are found to greatly aid in improving the performance after a thorough analysis using correlation techniques.

With this motivation, different feature selection techniques have been chosen to reduce the dimensionality. The clips of MIR-1K can be well segmented with the minimal feature vector. Since Indian music is the focus of this work, extended experimental analysis is done on Indian clips. Some experiments are done on the combined dataset (TBPS+MIR-1K). However, the accuracy is not up to the mark hence, ignored the same for experimentation. The correlation values obtained for an individual feature and a set of feature category have been computed as shown in Table 3.3 and Figure 3.12 respectively. The table is evident to convey that all the dimensions are not relevant for the specified task. As the correlation values obtained for a category of features are high, it drives us to work on individual feature analysis. From Figure 3.12, it is possible to say that some of the dimensions of

[3] *Staff* is a piece of paper which contains various music symbols.

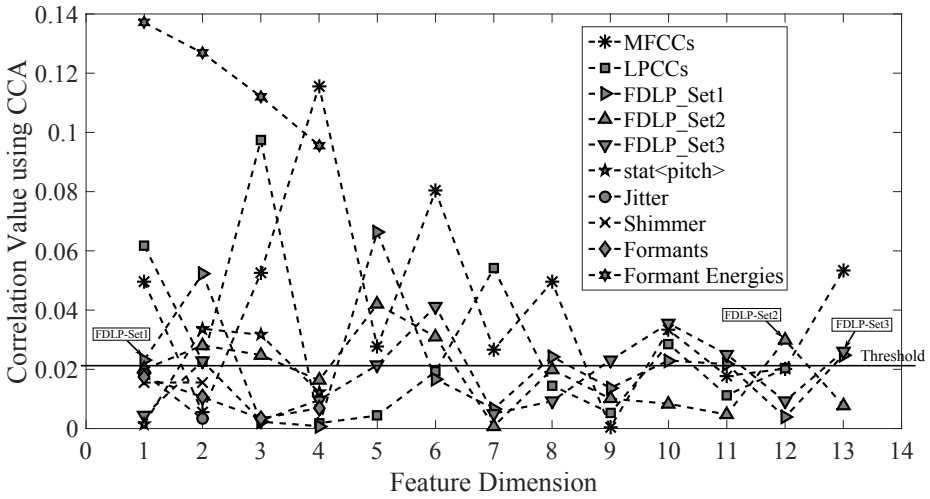

Figure 3.12: Correlation values obtained for individual feature using CCA. **Note:** FDLP is divided into three parts as its dimension is 39.

all the feature categories are very important to segment the vocal and non-vocal regions as they are less correlated. Of these, formant energies are found to be least significant for the same. Various threshold values are applied to select the suitable features based on the correlation values obtained. Out of all feature combinations, a feature vector with 43-dimensions obtained at a threshold value 0.021 is giving an accuracy of 84.32%.

Table 3.3: The correlation found between the vocal and non-vocal regions using the set of features using CCA.

Feature Vector	**Correlation Value**
MFCCs	0.1462
LPCCs	0.1859
FDLPs	0.1010
Pitch	0.2381
stat{pitch}	0.0249
Jitter	0.0140
Shimmer	0.0138
Formants	0.2229
Formant Energies	0.3329

However, it is uncertain to finalize the feature vector based on only the correlation analysis since there is a chance of unexplored combination that may give better accuracy. It is very important to observe the possible subsets of features to determine the best one.

Experimenting all the subsets, i.e., $2^n - 1$ (assuming n is the dimension of feature vector) is practically not feasible for real-time applications when n is large. There could be a chance of high computational issues in such cases. Hence, different feature selection algorithms namely correlation-based feature selection (CFS) (Hall, 1999), principal component analysis (PCA) (Abdi and Williams, 2010), gain ratio attribute evaluation (GRAE) (Robnik-Šikonja, 2004), Symmetric uncertainty attribute evaluation (SUAE) (Hall and Holmes, 2003), and Relief (Liu and Motoda, 2007) are considered. In addition, one more feature selection algorithm has been proposed for this work with the support of an evolutionary approach, called "genetic algorithm based feature selection" (GAFS). Four different classifiers namely support vector machines (SVM), Neuro-fuzzy classifier (NFC), random forest (RF), and artificial neural networks (ANN) are chosen based on their ability in handling the non-linear data. Various performance measurement techniques like precision, recall, accuracy, receiver operating characteristic (ROC) curve value, mean absolute error (MAE), and root mean squared error (RMSE) are used to evaluate the classifier efficiency. The performance values obtained for different feature selection algorithms with various classifiers are given in Table 3.4. It is found that all the feature selection algorithms are equally important. However, GAFS is selecting the best combination in optimal time. As the evolutionary algorithms are randomized, they always outperform many conventional approaches when the feature dimension is large. The comparative analysis about the feature vector lengths obtained using various feature selection algorithms versus the accuracy achieved has been plotted in Figure. 3.13. Though GAFS is little high when compared to PCA, a better accuracy has been obtained.

Out of four classifiers, ANNs are observed to be more suitable for classifying vocal and non-vocal segments as they are designed to handle highly non-linear data. Different measurement techniques mentioned above are considered to compare the classifier performance. Though the RFs are giving better accuracy many times (see Table 3.4), their error values are found to be more, and ROC is less when compared to ANNs. ROC is another factor to decide the suitability of classifier for the specific task. The classifier with more ROC sometimes gives less accuracy. Moreover, MAE and RMSE are also computed to justify the appropriateness. The values of MAE and RMSE are less in the case of ANNs.

Further, it is observed that there are some intermittent misclassifications in continuous vocal and non-vocal segments. Hence, a concept of $n-$ point moving window has been introduced to retain the continuity in both the segments. The concept of $n-$ point moving

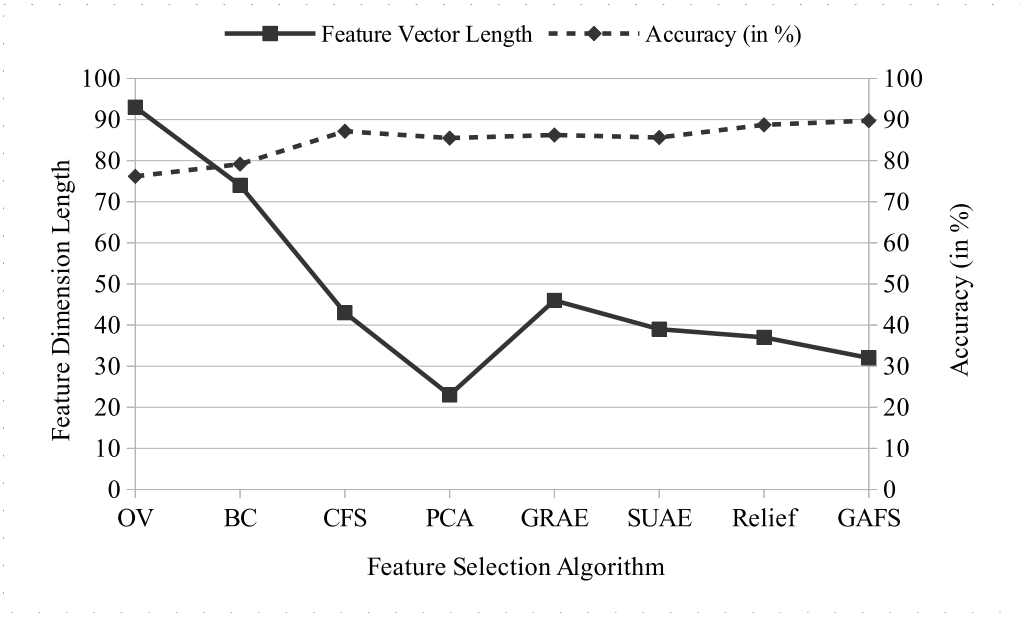

Figure 3.13: The outcome feature vector lengths of various feature selection algorithms and best accuracy values obtained with them. *Note:* Two new terms: OV: original feature vector (93-dimensional with an inclusion of base-to-peak formant values FH2 (first, second, and third formants))) and BC: Best combination (74 dimensional) obtained from original vector.

Table 3.4: The comparison of different feature selection algorithms with the proposed GAFS for four different classifiers. *Note:* Bold faced numbers indicate best performance obtained.

Performance	**Cfs Subset Eval. (43)**				**PCA (23)**				**GainRatio Attr. Eval. (39)**			
Measurement	**SVM**	**NFC**	**RF**	**NN**	**SVM**	**NFC**	**RF**	**NN**	**SVM**	**NFC**	**RF**	**NN**
Precision	0.823	0.871	0.854	0.847	0.832	0.831	0.854	0.836	0.829	0.837	0.862	0.859
Recall	0.823	0.871	0.854	0.847	0.832	0.831	0.854	0.836	0.829	0.838	0.862	0.859
Accuracy (in %)	82.34	87.15	85.49	84.76	83.25	83.19	85.49	83.64	82.93	83.84	86.24	85.99
ROC	0.848	0.832	0.879	0.892	0.837	0.896	0.932	0.937	0.874	0.896	0.932	0.917
MAE	0.352	0.349	0.316	0.279	0.296	0.279	0.234	0.186	0.348	0.324	0.301	0.223
RMSE	0.348	0.331	0.324	0.295	0.267	0.263	0.316	0.267	0.374	0.337	0.324	0.286
Performance	**SUAE (46)**				**Relief (37)**				**Proposed GAFS (32)**			
Measurement	**SVM**	**NFC**	**RF**	**NN**	**SVM**	**NFC**	**RF**	**NN**	**SVM**	**NFC**	**RF**	**NN**
Precision	0.818	0.829	0.849	0.856	0.793	0.832	0.876	0.887	0.819	0.833	**0.897**	0.892
Recall	0.818	0.829	0.849	0.856	0.793	0.832	0.876	0.887	0.819	0.833	**0.897**	0.892
Accuracy (in %)	81.86	82.93	84.95	85.63	79.36	83.25	87.63	88.71	81.92	83.28	**89.74**	89.23
ROC	0.834	0.879	0.924	0.931	0.769	0.876	0.893	0.947	0.917	0.892	0.945	**0.972**
MAE	0.324	0.316	0.287	0.268	0.369	0.324	0.287	0.278	0.296	0.287	0.267	**0.182**
RMSE	0.335	0.325	0.254	0.293	0.389	0.357	0.324	0.296	0.283	0.312	0.292	**0.259**

window is helpful to improve the classification accuracy and to locate the vocal onsets and offsets accurately. A thorough analysis has been done to decide the value of n. A human ear can perceive the information if it is played for the minimum of $1/4^{th}$ of a

Algorithm 1: Algorithm for $n-$ point moving window.

Input: An array of 1s and 0s representing vocals and non-vocals for all the frames of length 25 ms. (Outcome of classifier)

Output: An array of 1s and 0s after moving window.

```
// The array contains 1s and 0s representing vocals and non-vocals;
array[] ← classifierOutput();
// Each value of array is equal to frame length.;
length ← array.length();
// Start with first 11 frames by making center frame as pivot;
for i in 6 ... (length − 5) do
    count_1 ← 0;
    count_0 ← 1;
    // Check all the points of that frame and increase 1s count if the point value is 1.
      Else, increase 0s count;
    for j in i − 5 ... i + 5 do
        if array(j) == 1 then
            count_1 ← count_1 + 1;
        end
        else
            count_0 ← count_0 + 1;
        end
    end
    // Check the 1s and 0s count and update the pivot accordingly;
    if count_1 > count_0 then
        array(i) ← 1;
    end
    else
        array(i) ← 0;
    end
end
// Iterate the loop for all the points of an array.;
```

second (Hughes, 1946). Thus, the value of n is considered to be 11. In which, each point represents a frame length of 25 ms. Therefore, the total window length would be 275 ms.

An example illustrating the concept of an 11-point moving window is detailed in Algorithm 1. Travelling to the details, the output of classifier has been labelled with '1' for vocal and '0' in the case of non-vocal which is done for each frame of a clip. An 11-point moving window is placed on the first 11 frames and the midpoint is considered as a pivot. If the number of $1s$ in that window is greater than the number $0s$, then pivot element is replaced with '1' representing the frame is expected to be vocal. Else, pivot is replaced with '0' indicating that it is non-vocal. A frameshift of one is considered to iterate the above steps for all the array of frames. The improved accuracy values obtained before and after moving window are depicted in Figure 3.14. An accuracy of 95.16% is obtained with neural networks after windowing. In the case of singing voice detection, the system is able to give around 98% accuracy. The frame error rates of onset and offsets

are ignored since a part of singing voice can be selected from the longer portions of vocal regions.

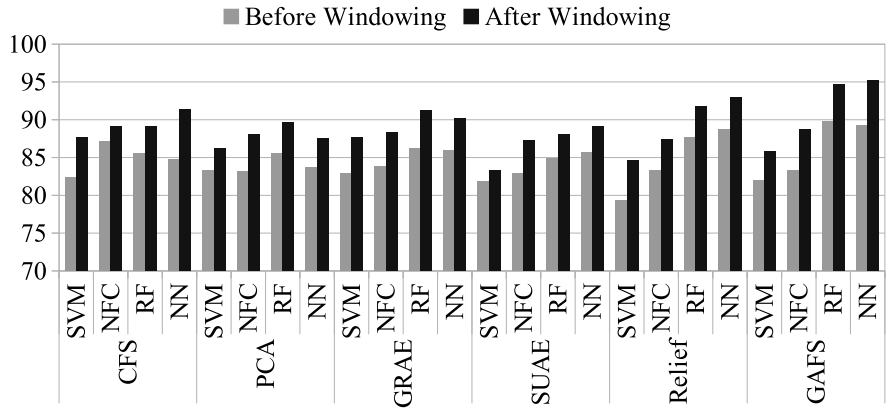

Figure 3.14: Comparison of accuracy values before and after windowing.

3.6 Summary

This chapter gives the possible solution to locate vocal onsets and offsets in a given audio clip. In particular, the focus is on Indian popular songs. Audio songs of two popular cine industries called *Tollywood* and *Bollywood* are considered. Various acoustical features and their variations have been computed along with frequency domain linear prediction (FDLP) values. Further, analysis has been done on formant structure to obtain the features that discriminate vocal and non-vocal regions. Since the length of feature vector with all these features is large, a feature selection algorithm based on genetics is also proposed in this work. Experiments have been conducted on both TBPS and MIR-1K datasets. The frame-level accuracy obtained at initial stage is less and hence, 11-point moving window algorithm is proposed to avoid intermittent misclassification among continuous vocal and non-vocal regions. The accuracy values obtained before and after 11-point moving window are given. Chapter 4 gives the implementation details of singer identification by locating the singing voice segments automatically using the approach proposed in this chapter.

Chapter 4

Singer Identification

> *The greatest respect an artist can pay to music is to give it life.*
>
> — Pablo Casals

4.1 Introduction

4.1.1 Motivation

The oldest musical instrument with which everyone is familiar is the singing voice (Sundberg, 1977; Sundberg and Rossing, 1990). One can easily discriminate the singer, once the person is familiar with the singers' singing voice. To identify a singer, a small portion of the vocal region is enough for the humans. Their perceptual apparatus and neural training mechanism help to do this efficiently. On the other hand, there is a huge growth in digital media due to the advances in technology. Thus, millions of tracks available to the listeners that create more confusion while selecting a song based on singer information and some good songs remain unnoticed. It is also observed that a majority of the song labels do not contain any information other than title of the track. Some tracks are found even without a title. Basic meta-information for a track includes album name, singer(s), genre, composer, mood, instruments, lyrics, etc. Of which, singer information is most widely used in music cataloguing and categorization. Many times users choose a song based on their favourite singers (Fujihara et al., 2010). The process of manual labelling of meta-information may take several human years which is practically impossible (Sturm, 2014). Moreover, a lot of inconsistencies may arise if manual labelling is not done by a music expert. A survey has been done to analyze the requirement of a number musicologists for labelling the meta-information and it is reported that at least 30 musicologists

are needed for a complete year to label the essential meta-information for all the tracks (Scaringella et al., 2006). It is quite difficult to even think about it. Hence, there is a need to automate the tasks of music information retrieval (MIR). The task of singer identification is further categorized into subtasks like target singer detection (TSD) in a song, target singer tracking (TST), identification of a song category (*solo or duet*), singer gender recognition, signer emotion recognition and so on. In this work, the task of singer identification is addressed.

4.1.2 Applications

The process of automating the task of singer identification has many commercial applications. It helps to identify the songs of a particular singer from the millions of tracks for a personalized collections. The song categorization based on a singer helps in extracting and recommending the songs to the listeners depending on their favourite singer. Further, it helps to create an album with a songs of particular singer. It is also possible to identify the singers who have similar characteristics based on the singers' timbre for grouping them into a cluster based on their similarities (Pachet and Aucouturier, 2004). Further, it plays a significant role in improving the rating of recommender systems.

4.1.3 Challenges

The task of automating singer identification is highly complex when compared to speaker recognition. During speech, one can observe some discriminating patterns like pauses, phonemes, unit separation, etc. Whereas, singing is a continuous speech with intentional change in pitch and vocal tract behaviour (Ratanpara and Patel, 2015). Moreover, the continuous background accompaniment for singing voice increases the complexity of singer identification (Helen and Virtanen, 2005; Comon and Jutten, 2010). The complexity in estimating the timbral information under background accompaniment is the major issue for the unavailability of an accurate singer identification system. The similarities among two singers may also confuse the system while categorizing the singers.

4.2 Research Gaps

1. The systems that are previously developed for singer identification are mostly done on studio recorded datasets that may not suitable for real-time applications.

2. The process of locating singing voice segments is done manually which needs to be automated.

3. A very less focus has been done towards popular singers of many cultural regions where their contribution to digital cloud is high.

4. The use of feature selection and the deep networks are not much considered in the literature.

4.3 Proposed Methodology

The framework of the proposed methodology has been given in Fig. 4.1. It starts with the task of database collection and locating the singing voice segments. Further, different features related timbre and temporal aspects that are suitable for singer identification are extracted. They are fed to two different classifiers for training and testing. Further, spectrograms and chromagrams are computed and are fed to CNN for estimating the performance of CNN over traditional feature-based approaches.

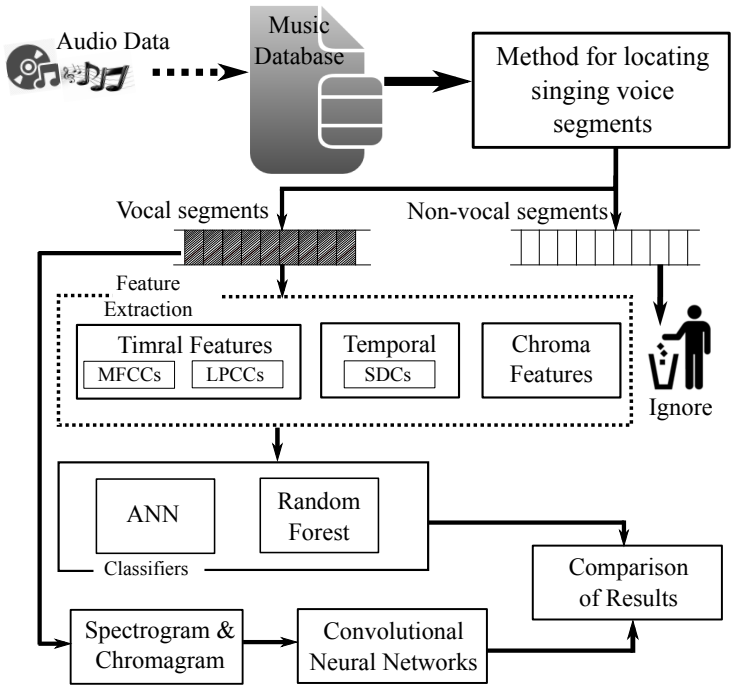

Figure 4.1: The proposed framework for Singer Identification System.

4.3.1 Feature Extraction

In this work, timbre and temporal features are exploited for singer identification. As timbre is the attribute that decides the quality of singer and also categorizes them, they are

considered with some derived features. Two different categories of features are extracted. Of these, one category is timbre, and the other is derived features of timbre values that captures temporal variations. Features like MFCCs & LPCCs are considered as timbral features, and shifted delta cepstral (SDC) coefficients have been derived from the cepstral coefficients that are used to capture temporal information.

A Mel-frequency Cepstral Coefficients (MFCCs)

MFCCs are the standard features that are fairly used in many speech processing applications and even in modelling music. As they are capable of correlating human auditory system (Murthy and Koolagudi, 2015), basic performance can be expected using them. Hence, they are extracted to discriminate singers of the audio songs.

B Linear Prediction Cepstral Coefficients (LPCCs)

LPCCs are another kind of standard spectral features that are designed based on the vocal tract system. These are the coefficients of the all-pole filter that are believed to represent the shape of a vocal tract. The process of computing LPCCs is taken from (Wong and Sridharan, 2001) and (Murthy and Koolagudi, 2018b).

C Shifted Delta Cepstral Coefficients (SDCs)

One variant of MFCC stacks and delta cepstral feature is named as SDCs (Kumar et al., 2011). They are computed from the adjacent frames of the audio clip and depend on four parameters (n, p, d, k). The value n indicates the number of cepstral coefficients, p & d indicate time shift & advances, and k denotes the span of the feature. For a given utterance with n_v number of cepstral features, $[n_v-(k-1)*p-d]$ number of SDC features can be extracted. The process of SDC feature extraction is shown in Figure 4.2. SDC feature (λ) at the time t is given as:

$$\Delta\lambda(t) = \lambda(t + jp + d) - \lambda(t + jp - d) \tag{4.1}$$

Where j lies between 0 to $k - 1$.

D Chroma Features

In a musical octave, the intensity associated with each of the 12 semitones is recorded by projecting them on 12 bins known as chroma features. Based on these 12 semitones, a 12-dimensional chroma feature vector is formed. Several routines are available to map

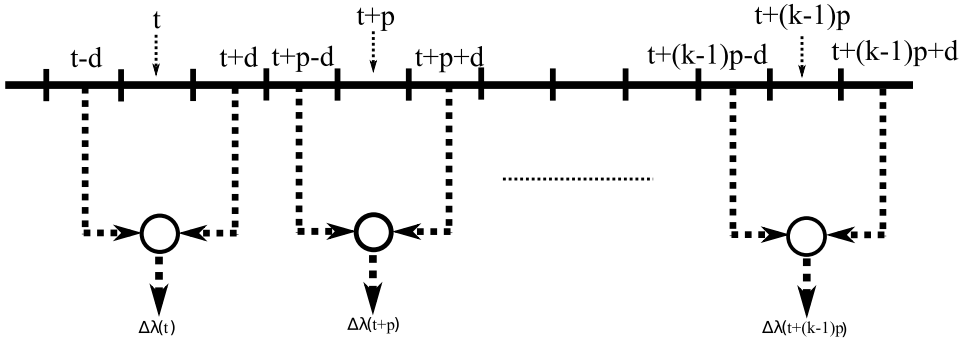

Figure 4.2: SDC feature extraction with parameters (*n-p-d-k*).

the spectra to chroma with some permissible loss. In this work, in order to improve the resolution of the underlying frequency and estimate the strong tonal components of the spectrum, in this work, phase-derivative has been considered for each FFT bin [1,4]. As they are prominent for singer identification (Ellis, 2007).

A total of 46-dimensional feature vector is formed by appending each group features that are mentioned above in the order i.e. MFCCs (13), LPCCs (12), SDCs (9), Chroma (12).

4.3.2 Feature Selection

More the number of features higher the computational complexity. There are chances for performance reduction too due to irrelevancy in some feature dimensions over the output classes. It is always necessary to develop a system with an optimized feature vector to obtain better results. Hence, feature selection is an important aspect that filters out the odd ones. Feature selection based on genetic approach, known as genetic algorithm based feature selection is introduced in this work (Murthy and Koolagudi, 2018b). The detailed process of selecting features based on GAFS is clearly explained in 3.4. This process gives a feature vector of length 28 with 41% reduction in the size of the resultant feature vector.

4.3.3 Classification Models

Two different classification models such as artificial neural networks (ANNs) and random forest (RF) are used based on their capability of capturing non-linear patterns though the number of output classes are high (Ryo and Rillig, 2017; Wei et al., 2019). Details of the classification models are provided below:

A Artificial Neural Networks (ANNs)

ANNs are found to be more suitable for many classification problems. The structure of ANNs is almost correlating with the human brain (Güçlü and van Gerven, 2017). Due to this, they are highly capable of handling non-linear patterns (Wei et al., 2019). In this work, three-layered ANN architecture has been considered containing input, hidden and output layers. The number of neurons in the input layer is equal to the length of the feature vector considered i.e. either 46 (Before features selection) or 28 (after feature selection). The number of neurons in the hidden layer is fixed to 1.8 times to the number of neurons in the input layer. The count is obtained after an empirical analysis performed based on the thumb rule (Boger and Guterman, 1997). The number of output classes decides the count of neurons in the output layer. In this case, the number of classes considered is 20. Feed-forward back propagation neural network (BPNN) has been applied to learn the singer specific properties through the feature vector. The neurons of each layer are fully connected to the neurons of adjacent layers.

B Random Forest (RF)

This classifier divides the entire feature space into several subspaces. A separate tree has been built for each subspace. These subspaces are randomly constructed from the original space (Breiman, 2001). The majority of the votes obtained for each tree are helpful in deciding the output class label. Attribute selection is one crucial step of a random forest classifier. Gini index (Quinlan, 2014) and information gain ratio criterion (Loh, 2011) are the two popular feature selection algorithms that are highly used in the literature. As the process of implementing random forest does not include pruning, it is found to be effective when compared to the traditional decision tree approaches (Ho, 1995). There may be a chance of over fitting if the tree is not pruned due to the Strong Law of Large Numbers (Feller, 2008). One advantage is that the tree gets converged at some point if not pruned. Based on this, it can be said that the effectiveness of RF classifier depends mainly on the number of trees constructed.

4.3.4 Convolutional Neural Networks (CNNs)

In CNN, convolutional layers that are designed to extract the relevant local features from all locations of raw input images. They are sometimes called as ConvNets. Their architecture is very similar to the traditional feed forward back propagation neural networks

(BPNN) with some additional blocks known as convolution layers, rectified linear unit (ReLU), pooling, flatten and *Softmax*. A simple architecture with [INPUT-CONV-RELU-POOL-FC] is given in Fig. 4.3 (LeCun, 2015; Karpathy et al., 2014). The description of each layer has been given below.

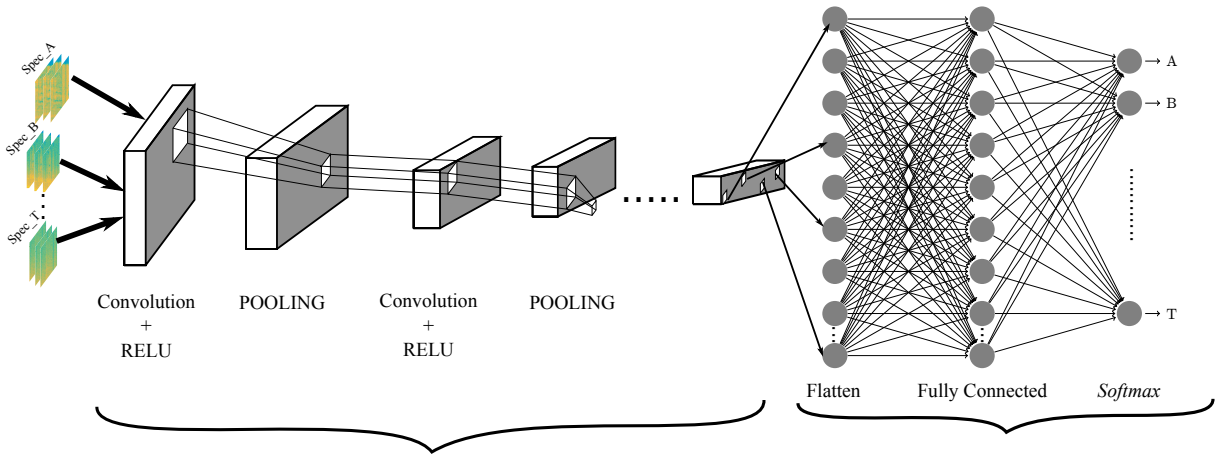

Figure 4.3: The proposed CNN framework for Singer Identification System.

A Convolutional Layer (CONV)

In this step, 2D convolution is carried out on the input images. Further, a small subset of spatially connected neurons from input images has been taken to generate local features. A weight matrix has been generated to perform the dot product with the chosen subset. The same weights are shared among all remaining nodes of convolution layers to retain the search for the same local features.

B Rectified Linear Unit (ReLU)

Increase in non-linear properties of decision function happens in ReLU layer. There are many functions available that increase non-linear properties of which, commonly used ones are hyperbolic tangent activation function ($f(x) = tanh(x)$), sigmoid function ($f(x) = (1 + e^{-x})^{-1}$), and non-staturing activation function ($f(x) = max(0, x)$).

C Pooling

Another important function of CNN architecture is pooling which reduces the size of feature maps based on the maximum feature response. Many non-linear functions have been introduced for pooling. Of which, *maxpooling* is the most commonly used. It divides the input image into a collection of rectangles that are non-overlapping with each other. The

same operation is done for each sub-region that outputs the maximum. The relativeness of the feature with other features has been estimated in this process. The number of parameters and computation time are generally optimized using pooling layer which fits in between two successive convolutions (Krizhevsky et al., 2012).

D Fully Connected Layer (FC)

After a set of convolution and pooling operations, the obtained feature space is passed on to a fully connected network. The structure of the FC layer is similar to the standard neural network architecture where each node is connected with every node of the subsequent layer. Activation function has been applied at every node to process the input value of the node.

E Softmax Layer

Similar to the sigmoidal function, the *softmax* function sets down the information in the range [0,1]. Moreover, the division happens in such a way that the sum of output values is equal to 1. Hence, it can be said that the output of *softmax* layer is in the form of probability distributions which decide the output class label. The mathematical equation for the *softmax* layer is given in Eq. 4.2.

$$\sigma(s)_i = \frac{e^{s_i}}{\sum_{i=1}^{T} e^{s_i}} \tag{4.2}$$

Where s is the vector of inputs to the output layer and i indexes the output classes whic varies from $i = 1, 2, ...T$ (Wang et al., 2016).

4.4 Experimental Analysis

In this section, the analysis of the proposed singer identification system with the results obtained using different feature combinations and classifiers is given. The explanation about the proposed dataset for this work is given in 2.1.3. In addition to that, a standard *artist20* dataset is also considered for comparative analysis and the details about *artist20* are given in 2.1.4. Further, the singing voice segments have been located by using the approach proposed in 3.3. Further, the proposed singer identification system has been implemented and the details are as given below:

The focus of experiments that have been conducted here is based on two aspects. The first one is for understanding the effectiveness of selected features and the second

Table 4.1: The perforamnce of the various feature combinations over different classifiers and the affect with CNNs on IPSD and *artist20*.

Sl. No.	**Feature Vector/ Classification Model**	**Accuracy (in %) for a clip of length (in seconds)**																
		Indian Singers Database									***artist20 dataset***							
		NN				**RF**				**CNN**	**NN**				**RF**				**CNN**
		60s	**30s**	**10s**	**5s**	**60s**	**30s**	**10s**	**5s**	**5s**	**60s**	**30s**	**10s**	**5s**	**60s**	**30s**	**10s**	**5s**	**5s**
1	MFCCs	29.44	25.12	24.06	22.39	27.86	25.48	20.69	17.12		29.58	23.32	21.15	18.36	31.25	21.05	17.54	14.53	
2	LPCCs	29.98	26.28	24.14	21.32	24.96	20.45	20.18	18.02		26.13	24.96	2138	19.58	33.59	25.78	23.63	21.43	
3	Chroma	23.59	21.25	20.47	18.54	29.18	26.89	23.85	22.72	–	31.29	26.47	24.18	22.78	36.52	31.45	26.97	26.33	–
4	MFCCs + Chroma	56.57	54.68	53.15	51.78	59.14	57.35	53.08	51.36	**NA**	45.18	42.49	40.35	39.33	56.87	42.73	40.68	38.19	**NA**
5	LPCCs + Chroma	54.15	52.66	53.94	50.12	59.58	56.80	53.99	52.18	–	44.72	42.16	40.39	39.22	59.72	53.28	50.77	49.64	–
6	MFCCs + LPCCs + SDCs + Chroma	57.88	56.21	56.02	54.95	62.41	58.29	56.28	54.36		52.15	51.03	49.39	48.25	61.18	54.29	52.18	49.96	
7	Selected <6>	61.86	59.25	58.33	58.06	**63.17**	61.78	58.47	56.15		58.23	56.94	56.12	55.75	**61.69**	60.82	57.45	53.12	
8	Spectrogram Image	**NOT APPLICABLE**								**75.50**	**NOT APPLICABLE**								42.13
9	Chromagram Image									35.23									23.12

is to analyse the use of trending CNNs in singer identification. Two different classification models namely artificial neural networks (ANNs) and random forest (RF) have been used to understand the importance of feature-based approach for singer identification. Different timbral and temporal features that have proved their worth in speech applications are used in modelling music information. The list includes Cepstral coefficients such as Mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) that are capable of correlating human perceptual mechanism. The features known as shifted delta cepstral (SDC) coefficients have been computed and used as temporal variations instead of conventional velocity (Δ) and acceleration ($\Delta\Delta$) features. Further, chroma vector has been computed from chromagrams forming a 46-dimensional feature vector with all the features that have been mentioned above.

The test data is partitioned into the clips of different lengths of 60, 30, 10, and 5 seconds. However in the dataset, the number of clips of 10 and 5 is considerably high when compared to those of 30 and 60 seconds. The number of clips that are considered in the experiments is 100 for each in the case of 5 and 10 seconds, to understand the contribution of each category of features towards singer specific information. The classifiers are trained and tested using individual and combination of features separately. The results obtained are tabulated in Table 4.1. In the table, rows of serial numbers 1 to 6 show either individual or combinations of some of the timbre, temporal or chroma features. The row with sl.no. 7 show the singer identification performance using reduced number of features by applying genetic algorithm based feature selection technique. The results obtained using neural networks and random forests have been displayed for both IPSD and *artist20* datasets. The last two rows of the table show the results obtained using convolutional neural networks by considering the spectrograms and chromagrams. For each classifier except CNN, the results that are reported in the table are for the audio clip lengths of 60, 30, 10, and 5 seconds. For CNNs, the spectrograms and chromagrams are given as inputs since CNNs are capable of extracting the features directly from the images (Vieira and Ribeiro, 2018; Liu et al., 2018). Rows with serial numbers 8 & 9 of Table 4.1 shows the accuracy values obtained using CNNs with spectrograms and chromagrams. The measurement which is considered to represent a value in each corresponding cell is accuracy, and computed using the formulae shown in Eq. (4.3).

$$SIA = \frac{\sum_{i=1}^{N} \frac{Correct_{s_i}}{Total_{s_i}}}{N} \tag{4.3}$$

Where SIA is singer identification accuracy, N is the total number of singers (in this case, $N = 20$), $Correct_{s_i}$ is the number of correctly identified clips of the singer s_i, $Total_{s_i}$ is the total number clips for the same singer.

The results of the table gives a clue that the MFCCs and LPCCs can be considered as baseline features for identifying a singer to any kind of singer database. Addition of MFCCs to other features is found to be advantageous and they do not degrade the performance. In addition to cepstral coefficients, some of their statistical values are also contributing little to the performance improvement. Though they are not contributing much, they have been included to specify their importance for singer identification. Based on this, SDCs have been directly added to the final feature vector instead of experimenting with them separately. In few works (Ellis, 2007), chroma features have been considered for singer identification. Based on this, experiments are also conducted using chroma features. They have been experimented individually and also as a group. Chroma as a combination yields to better results than that of individually. Better results are obtained with chroma as a combination rather than alone (see row numbers: 3,4,&5). The combination of all these features is giving better performance, however, computational complexity issues due to the high-dimensional feature vector. As the feature selection techniques select only relevant and useful features the results obtained using the selected feature vector are better than the results obtained using all the dimensions. One can also observe that the performance using random forest is better when compared to that of neural networks. The reason could be the feature subset selection that is involved in the process of random forest classification algorithm. At the same time, the performance is degrading if there is a reduction in the length of the input clip. The classifier may not be able to classify all the frames accurately and hence, reduction in performance.

It has been found that, all the dimensions of feature vector do not contribute much to the classification performance of certain task most of the time. Moreover, there can be a chance in performance degradation due to some irrelevant dimensions. Some visual notations have been given in Fig. 4 which depicts the distinguishable and non-distinguishable features for singer identification. In this figure, *x-axis* represents normalized feature values and *y-axis* shows their normalized frequency score of twenty different classes (in this case, singers). In the cases of MFCC-3, LPCC-1, and Chroma-12; one can observe clear visual distinction for 20 singers. Where as, distinction is difficult for MFCC-2, LPCC-3, and Chroma-10. See figures 4.4(a) and 4.4(b) respectively. This observation is called the

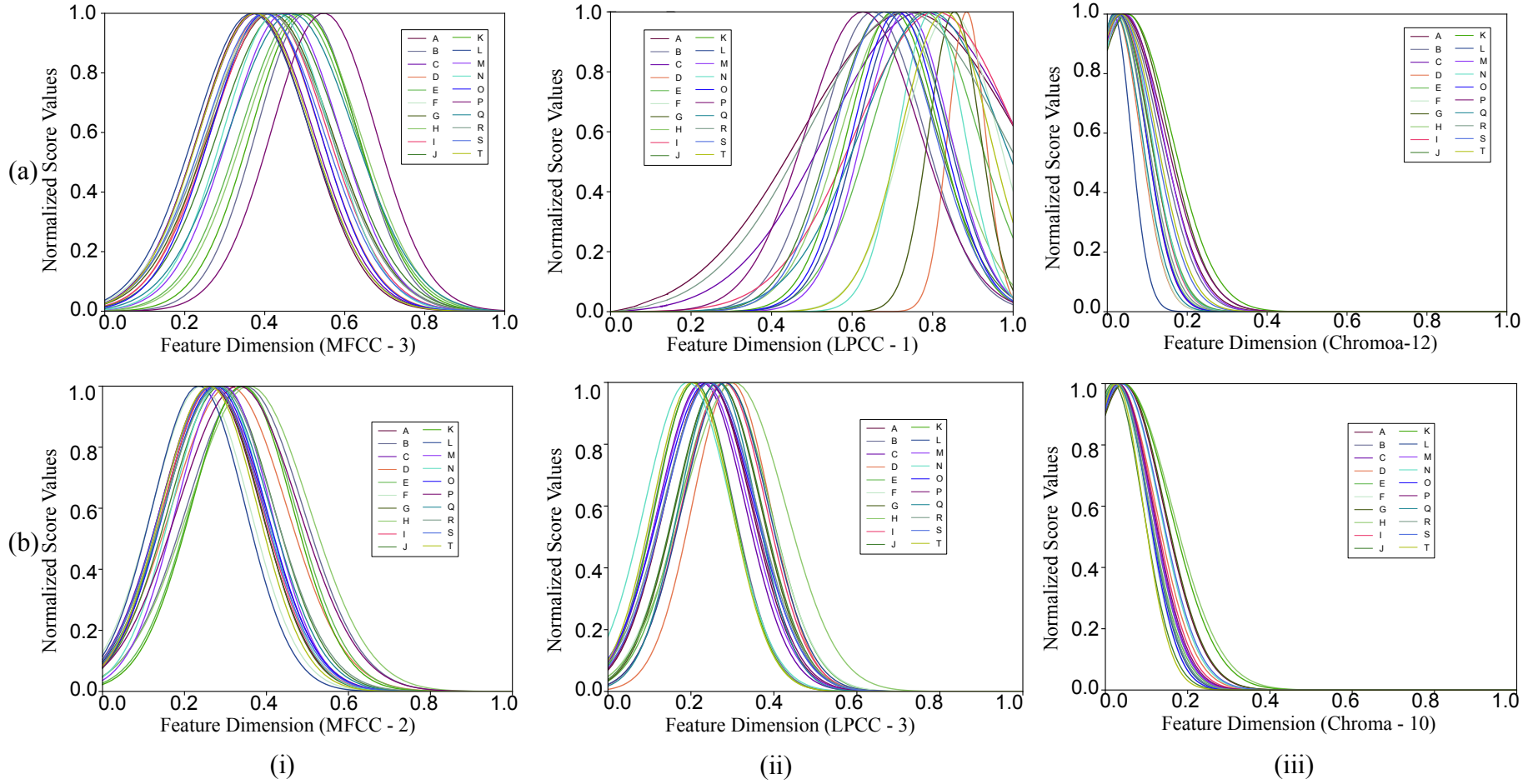

Figure 4.4: Visual correlation score histograms obtained for 20 singers which justifies the usability of a chosen feature dimension. Rows: (a) Features that are capable of discriminating singers, and (b) represent features that are not suitable for discrimination. Columns: (i) MFCC feature, (ii) LPCC feature, and (iii) Chroma feature.

visual correlation analysis.

The results obtained using the selected feature vector of size 28 have been compared with few recent singer identification works that have also been done for *artist20* dataset. The results of the proposed work with random forest classifier are found to have improved when compared to the existing works shown in Table 4.2. The table contains the information about the feature vector that is considered in previous works along with its length, classifier used, and accuracy obtained. Acoustical features such as MFCCs are used as baseline features in almost all the works that have been mentioned in the table. The design of MFCCs which correlates with the human auditory system may be the reason for that. Hence, the same MFCCs along with LPCCs have been considered as baseline features in this work as well. The support vector machine (SVM) is a highly used classifier, due to its kernel capability of handling non-linear data. However, neural networks are found to be highly accurate when compared to SVM. There is a chance to improve results with the random forest (RF) classifier as they consider majority voting to decide the best combination of feature dimensions. The selected feature vector obtained using GA is giving better performance with a random forest classifier. The reason could be the two-stage feature selection to select the best suitable subset. Of which, one selection is by GA and the other is by random forest. In this case, though ANN is highly suitable for multi-class classification problems, they are unable to provide equivalent results when compared with RF. Two approaches (Su and Yang, 2013; Sarkar and Saha, 2015) are able to give better performance when compared to proposed work. However, the feature vector length proposed in their work is substantial. They are computationally expensive and cannot be considered for real-time applications since the present trend is looking for optimized methods.

Apart from this, conventional neural networks (CNN) are also used in this work for singer identification as they are presently trending now a days in many applications. Short-term spectrograms and chromagrams from the frame of 25 ms are computed from the clips of length 5 seconds. These are directly fed to CNN during the training phase. Different configurations of CNNs are tested with different parameters such as number of filters, hidden layers, neurons in each hidden layer, activation function and so on. The results of singer identification that are obtained on chromagram and spectrogram image inputs are given in the last two rows of Table 4.1. They have been quoted them for both IPSD and *artist20* datasets. Though it is claimed that chromagrams mainly represent

Table 4.2: Comparison of Proposed Results with the existing works done for *artist20* dataset. *Note:* * It is not given in the article. However, MFCCs and their statistical variations length is more than 28.

[REF]	**Features**	**Feature Length**	**Classification**	**Accuracy (in %)**
(Ellis, 2007)	MFCC	13	SVM	54.00
(Ellis, 2007)	MFCCs with Chroma	25	SVM	57.00
(Langlois and Marques, 2009)	MFCC and stat<MFCCs>	>28*	HMM	59.14
(Sarkar and Saha, 2015)	STE+ZCR+MFCCs+ stat<MFCCs>	616	NN	62.77
(Eghbal-Zadeh et al., 2015)	MFCCs	20	kNN LDA	55.23
PROPOSED	Selected<MFCCs+LPCCs+ SDCs+Chroma> using GA	28	RF	61.69

musical patterns (Harte and Sandler, 2005), the performance obtained is incomparable with the other results. The reason can be the resolution issues.

Table 4.3: Hyperparameters considered for designing the CNN for the task of Singer Identification.

Sl.No.	**Parameter**	**Value**
1.	Batch size	8
2.	No. of channels	3 channels (RGB)
3.	Filter size	3*3
4.	Image size	256*256
5.	No. of convolution layers	4
6.	No. of flatten layers	2
7.	Softmax layer	1
8.	No.of output classes	20
9.	Activation function	*tanh* & Relu
10.	No. of epochs	Around 50-60

The CNN has given some baseline results with default parameters. Further, hyperparameter tuning has been applied to identify the right set for this dataset and the improved the performance of singer identification system. A few noticeable parameters that affect the performance have been identified and listed in Table 4.3. They include batch size, image size, number of convolutional layers, number of channels, filter size, number of epochs,

activation function, etc. The parameters known as batch size, and image size represents the number of processed inputs and the size of each image respectively. The values set to these two parameters directly affect the computational complexity. Though there is an increase in the size of image resolution which gives detailed information about the pixels, there is a limit to this as it is not possible to get more than the captured information. The experimentation has been done with different image sizes, and better accuracy is found with the image size 256. After that, the accuracy is observed to be constant.

Similarly, the parameter values of batch size, number of convolutional layers, number of flattening layers, filter size, input channels, number of output classes, number of epochs, and activation function considered for this work are detailed in Table 4.3. All these values have been set based on certain experimentation. Better performance has been obtained with the values mentioned in the table. The number of epochs is initially set to 1000, and the weights are converged at around 60 epochs.

An empirical based analysis has been used to fix the values of each parameter. The same parameters that are considered for Indian popular singers database (IPSD) are used to classify the singers of *artist20* dataset as well. However, the results obtained using these parameters are found to be worse. An average accuracy values of around 42% and 23% are obtained using spectrogram and chromagram respectively. Moreover, the performance of the chromagrams is very less even in the case of IPSD. From this, it can be seen that the features obtained from chromagram are performing well and a complete chromagram image is not suitable for singer identification. The reason can be the missing information related to singers in complete chromagram and CNNs are unable to capture the same from local windows. The accuracy obtained with CNN can motivate the future researchers to explore the required components that are to be changed in the structure of CNN.

4.5 Summary

This chapter gives the implementation details of singer identification. Two different datasets such as Indian popular singers database (IPSD) and *artist20* are used in this work for singer identification. Two different cine industries such as *Tollywood* and *Bollywood* are considered to construct IPSD. Features such as MFCCs, LPCCs, SDCs and Chroma are computed for constructing a feature vector. An accuracy values of 63.17% for IPSD and 61.69% for *artist20* are obtained with the selected features. The feature selec-

tion is done based on visual correlation and genetics based feature selection algorithms. Further, the convolutional neural networks (CNNs) are considered by feeding spectrograms as input images to them. Better accuracy of 75% is obtained for Indian dataset. However, the accuracy is not satisfactory in the case of *artist20*. CNN parameters that are considered for Indian dataset may not be capable of classifying the artists of *artist20* due to cultural changes.

Chapter 5

Music Mood Estimation using Acoustical Features and CNNs

> *Music can bring us tears, console us when we are in grief and drive us to love.*
>
> — Yi Hsuan Yang

5.1 Introduction

The Emotion expressed by a person determines their present mental state. People tend to behave abnormally when they are mentally stressed. The Emotion of a person can be recognized based on certain parameters such as voice, speech, facial expressions, heart rate, blood pressure, sweating and so on. Of these, speech and facial expression are two prominent parameters that have gained high attention from the researchers in the past two decades. Since they are convenient to analyze and compare, several research works have been proposed in the fields of speech and image processing to recognize the emotions, while the rest of the parameters still need an efficient automated system. In addition to the parameters mentioned above, music is also one such prominent factor which also helps in estimating present mental state of a person.

Every music piece is capable of expressing some emotion (Feng et al., 2003b; Huron, 2000). It is not possible to compose, perform, or listen to a music signal without the involvement of related emotion (Juslin and Sloboda, 2001; Yang and Chen, 2012). Hence, emotion is a prominent attribute that helps in music organization, cataloguing, indexing,

and recommendation. The survey of *Last.fm* says that emotion is the third most demanding attribute which is tagged by online users (Lamere, 2008). In 2004, a survey has been conducted, to know the important attributes of music clips, for implementing an effective Music Information Retrieval (MIR) system and 28.2% people mentioned that the emotion is an important attribute for music file organization (Yang and Chen, 2011). Now, the task of music emotion recognition has gained more importance and several works have been proposed in the literature with different approaches.

Music emotions are broadly categorized into three types namely *expressed*, *perceived*, and *evoked* (or *felt*) emotions (Gabrielsson, 2001; Hallam et al., 2011; Huron, 2006). Expressed emotions belong to the performer and the emotional information is completely dependent on the performer's expertise. In this case, the performer can be a singer, or a composer. The remaining two depend on the listener's view points. Some may listen to the songs and are capable of perceive emotion and a few listeners may feel the emotion. These kind of listeners are generally useful in labelling the song clips based on their inherent capabilities of recognizing emotions. The process of automatic emotion recognition is highly dependant on the representation of emotional patterns inside the song clip (Knox et al., 2011; Valstar et al., 2016). The performers express emotions with some acoustical cues into the song clips. The representation of emotions is based on the *perception meter*, *harmony*, *tonality*, *melody*, *timbre*, *rhythm*, *style*, and so on. Hence, it is possible to recognize the emotions using acoustical features (Juslin and Sloboda, 2001; Lu et al., 2006). However, certain analysis may be required to identify the suitable features from the categories mentioned above. In this work, an effort has been made to classify the six important music emotions using acoustical features and trending Convolutional Neural Networks (CNNs).

5.1.1 Applications

The process of music mood estimation has several useful applications in commercial, social, and pathological aspects. It is possible to categorize and index the music clips based on the emotional patterns in it. Moreover, the efficiency of recommender system can also be improved by tagging emotions to music clips. Since a majority of the listeners show their interest in listening to the songs depending on their emotional states, it is possible to estimate their mental behaviour, based on the songs chosen by them and would also help in developing a multi-modal system for recognizing the present environment of a

listener and provide songs accordingly. For instance, if a person is in a birthday party, then energetic songs can be played, ignoring sad songs in the playlist. The system developed for music recommendation based on mood of a song can also keep a driver awake.

5.1.2 Challenges

Though the mood information is highly needed for several applications, development of a sophisticated system is still under progress due to three major challenges: (i) lack of proper databases, (ii) ambiguities in mood labels, and (iii) cultural differences.

i. *Lack of proper databases:* There are many standard databases available in the literature for task of speech emotion recognition. Speech is a continuous task and since the system which is developed for one language may give similar performance with another language, the same is not true in the case of music emotion recognition due to prosodic similarities. There are many factors that influence emotions of one region when compared with the other region song clips. Hence, it is essential to consider these factors while constructing a database for Music Emotion Recognition (MER). Some efforts have been done to develop a database for music emotion recognition; however, there are many limitations found in them. One important issue identified are low number of emotional classes considered without proper analysis. Another important issue is copyright protection which restricts researchers from distributing the database publicly. Due to this, most of the implemented works are unable to share their datasets. Moreover, the data which is shared publicly has only a few number of clips, emotional classes, and improper emotional tags (Aljanaki et al., 2017). In addition, the emotion in a song changes over time resulting in more than one emotion for a given song clip.

   The online websites such as *moodfuse.com*, *allmusic.com*, and *last.fm* are facing the same issue of less number of tags and emotional classes. Of late, MediaEval Database for Emotional Analysis in Music (DEAM) is one combinational dataset which is designed for western moods and contain 1802 songs (online: *http://cvml.unige.ch/databases/DEAM/*) (Yang et al., 2018). However, the number of emotional classes are less in their database. By keeping in mind, the above issues, a database with six emotional classes are considered in this work for Indian *Tollywood* and *Bollywood* music.

ii. *Ambiguities in mood labels:* The task of musical mood estimation remains challenging due to the inherent ambiguities in the specified mood labels. It happens due to the disagreement, while perceiving and interpreting the music clip, by two different people (Kim et al., 2010; Schmidt and Kim, 2010). This leads to the complications in assigning a mood label for a song clip. Some times, there may be a chance of having more than one mood label for a song clip. In this thesis, Mean Opinion Score (MOS) has been taken from different listeners, music professionals and experts, while labelling the mood for a song clip. Care has been taken to consider only single label for each audio clip.

iii. *Cultural differences:* Emotions are biological and socio-logical in nature (Argstatter, 2016). There exists cultural differences in emotions with respect to prevalent, modal, and normative responses (Mesquita et al., 1997). There are two kinds of differences namely, *excess* and *deficits* in emotions. Due to these differences, there exists a lack of coherence in the emotional components of two regions. However, research has taken place estimating universal and biological differences in the emotions. There is less focus on socio-logical aspects in the literature. However, these aspects are also highly helpful in estimating the cultural differences ((Mesquita and Walker, 2003)). There are several other factors that exhibit differences in emotions for different regions. The appraisal, experience in expressing emotions, behaviour of automatic nervous system, and perceiving nature are the few notable ones (Barrett and Russell, 1999). Hence, in this thesis, care has been taken to include the emotional clips that are related to the same culture.

5.1.3 Proposed Emotional Classes

Normally *Russell's* and *Thayer's* emotional models are used to represent emotions on an X-Y plane, where X-axis represents valence and Y-axis represents arousal . Recently, many works have been reported to combine them into a single class as shown in Figure 5.1(a) (Saari and Eerola, 2014). There are 101 unique emotional classes available in the combined model. However, it is difficult to identify the song clips for all the emotional classes. Hence, six different moods are identified based on their energy levels namely, *Angry*, *Devotional*, *Energetic*, *Happy*, *Romantic*, and *Sad*. The position of the identified emotional classes is located in the proposed mood model, shown in Figure 5.1(b).

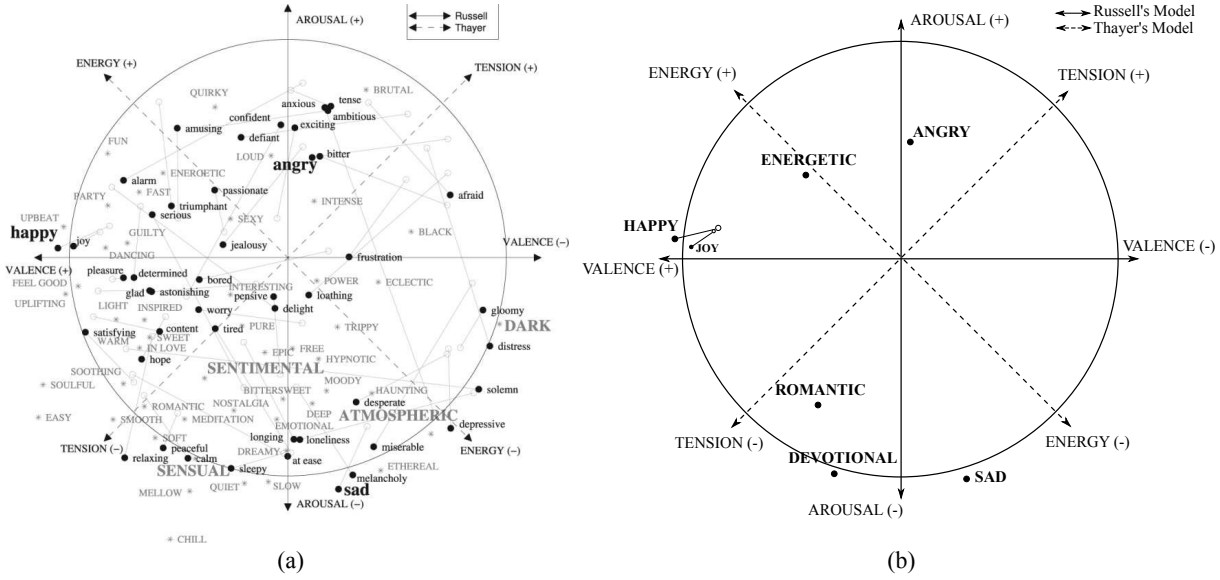

Figure 5.1: Identified emotional classes for the song clips from the combined *Russell's* and *Thayer's* models. (a) Recognized 101 unique emotional terms (PC: (Saari and Eerola, 2014)), and (b) Proposed emotional classes.

5.2 Proposed Methodology

The proposed method for music mood estimation has been depicted in Figure. 5.2. It is implemented in two levels. At the first level, the given input signal is categorized into either *energetic* or *non-energetic* class. In level-II, actual mood has been recognized, from respective *energetic* or *non-energetic* classes, using convolutional neural networks. The terminology and approach considered in each level classification have been detailed in the subsequent subsections.

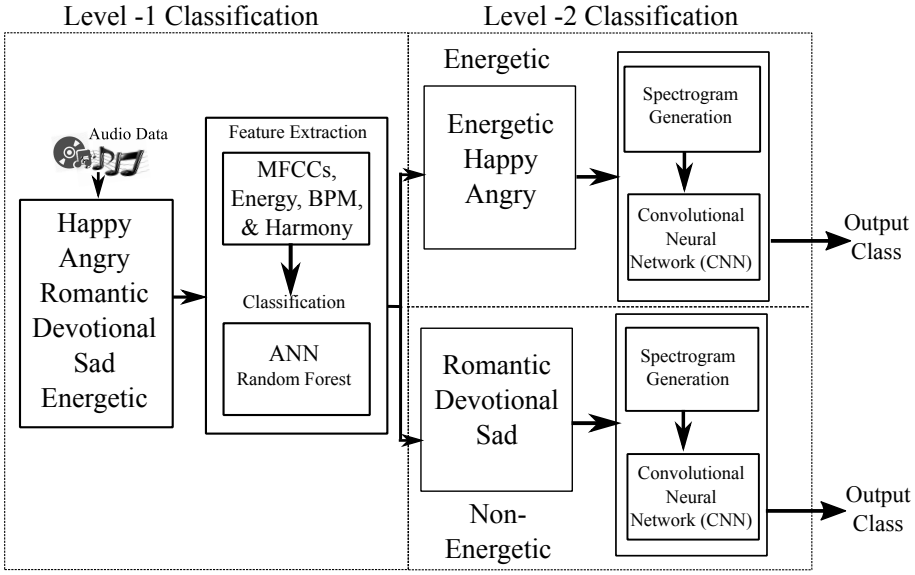

Figure 5.2: Further categorization of six moods (*Angry*, *Devotional*, *Energetic*, *Happy*, *Romantic*, and *Sad*) into *energetic* and *non-energetic* classes.

5.2.1 Level-1 Classification

In the first level of classification, the system decides the broad category of a given audio clip. Two categories have been identified. One is *energetic* class and the second one is *non-energetic* class. The set of features considered for categorizing the audio clips into two broad classes is described below. Moreover, the details of ANN classifier are also provided.

A Features considered for level-1 classification

Four different types of acoustic features namely Mel-Frequency Cepstral Coefficients (MFCCs), Short-Time Energy (STE), Beats Per Minute (BPM), and harmony are considered for categorizing the input audio clip into either *energetic* or *non-energetic* classes. Since the description about MFCCs has already been provided in Section 3.3.1(A), details of the same are not given in this chapter. The process of computing other features is detailed below:

1. *Short-Time Energy (STE):* The term energy is directly related to the loudness of the signal. The instrument which has louder outputs, enforces higher energy in the signal than that of the one that has torpid outputs (Fu et al., 2008). As the total energy of complete music clip does not give any information w.r.t. emotions, short time energy is computed from each frame, as shown in Eq. 5.1 (Anagnostopoulos et al., 2012).

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \tag{5.1}$$

   Where E_n is the energy value, x is the input signal, N represents the length of the frame, $w()$ represents analysis window which can be rectangular or hamming, and n is the sample where the analysis window is focused.

   It is observed that the energy of the songs, that belong to energetic category, is normally high when compared to the non-energetic song clips (Scherer, 2003). Hence, energy could be one of the useful feature for categorizing the songs into two classes.

2. *Beats Per Minute (BPM):* The music can be perceived in an organized manner due to the involvement of *tempo*. It provides a platform to build melodic-harmonic lines (Weikart, 2003; Norris, 2009). Tempo, rhythm, and harmony are the three

important attributes that influence the effectiveness in music emotion recognition (Fernández-Sotos et al., 2016). Based on this, two features, tempo and harmony, are considered in this work. They are chosen to estimate the mood category (either *energetic* or *non-energetic*) of a given audio clip. The tempo of a music clip has been measured using Beats Per Minute (BPM) and the process of computing BPM is given below.

Initially, the given input signal is converted into time-frequency representation using Short-Time Frequency Transformation (STFT), which is applied on shorter segments known as frames. Further, spectral energy flux is computed using the equation given in 5.2.

$$E_{sf}(f,p) = \sum_{k} h(k-p)G(f,k) \tag{5.2}$$

Where E_{sf} is the spectral energy flux, $h(k)$ approximates the differentiator filter, i.e. $H(e^{j2\pi f}) \approx j2\pi f$, and the transformation $G(f,k)$ is obtained based on low-pass filter which is applied on frequency representation $|\tilde{X}(f,p)|$ of given signal $x(n)$. Along with low-pass filter, a *non-linear compression* has also been applied by masking rapid modulations, shown in 5.3.

$$G(f,k) = \mathcal{F}|\tilde{X}(f,k)| \tag{5.3}$$

An empirical study has been carried out to decide the order of differentiator and the filter of order 8 is found more suitable w.r.t. complexity and efficiency. In connection to the parameters considered for computing spectral energy flux, an N point Fast Fourier Transformation (FFT) has been considered to evaluate STFT. The set of frequency channels obtained with this process are $\frac{N}{2}$ and they are considered for time-frequency representation. All these values are filtered using $h(k)$ to obtain spectral energy flux. Later, a temporal waveform, called $w(k)$ has been produced by summing all the positive contributions. It gives the locations, where the energy flux is large. These locations are generally the onsets of beats (*also called* beat onsets).

Further, median filter is applied to estimate the true beats. A dynamic threshold is used to suppress the unwanted beats (Alonso et al., 2004). After this process,

chosen beats are processed further periodicity among beats is ignored as the music clips considered for experimentation are polyphonic in nature.

3. *Harmony:* Harmony is another important feature for categorizing the moods. Chord sequences are a set of features that have already proven their ability in categorizing the genre classes and music classification (Cheng et al., 2008). The process of recognizing chord sequence is easier with monophonic clips. Though some works have been done on automatic chord extraction from polyphonic clips (Zenz, 2007), accurate results have not been obtained in the case of polyphonic clips. Moreover, one important observation made from the literature is that, the chromagram is more suitable for extracting chord sequence information (Oudre et al., 2011; Jiang et al., 2002; Rolland, 2014). Based on this, chroma features have been computed from chromagram instead of chord sequences. They are assumed to be better for classifying the moods in the case of polyphonic clips since they belong to super class of chord sequences.

With the above mentioned features, a 27-dimensional feature vector is formed and includes 13-MFCCs, 12-chroma, 1-STE, and 1-BPM. All these features are computed based on the acoustic cues identified among two different emotional classes called *energetic* and *non-energetic*. Further, the behaviour of acoustic features of six emotional classes is also mentioned. It says that, the behaviour of the above mentioned features is same with the classes of *energetic* and *non-energetic*. Table 5.1 shows the behaviour of each feature w.r.t. to the emotional categories. The rows of the table represent the feature name and the columns contain the value corresponding to the emotion.

Table 5.1: Acoustic cues observed among different emotions for different features.

Feature/ Emotion	**Energetic**			**Non-energetic**		
	Angry	**Energetic**	**Happy**	**Devotional**	**Romantic**	**Sad**
Tempo (BPM)	Fast	Fast	Fast	Slow	Slow	Very Slow
Sound Level (Energy)	Very High	Very High	High	Low	Regular	Very Low
Spectrum	High Energy (at HF)	High Energy (at HF)	Energy fluctuation	staccato	Regularity	Low energy Irregularity
Spectral flux	Bright	Little Bright	Bright	staccato	Little Bright with regularity	Dull

B Artificial Neural Networks (ANNs)

The Artificial Neural Networks (ANNs) are found to be better in handling non-linear data even though the size of training data is small and consists of few number of classes. Hence, ANNs are used for classifying the input audio clips into either *energetic* or *non-energetic* classes. Three layer ANN with single input, hidden and output layers is used. The number of neurons in input layer is equal to the length of the feature vector (i.e. *27*). The structure of ANN used for level-1 classification is shown in 5.3.

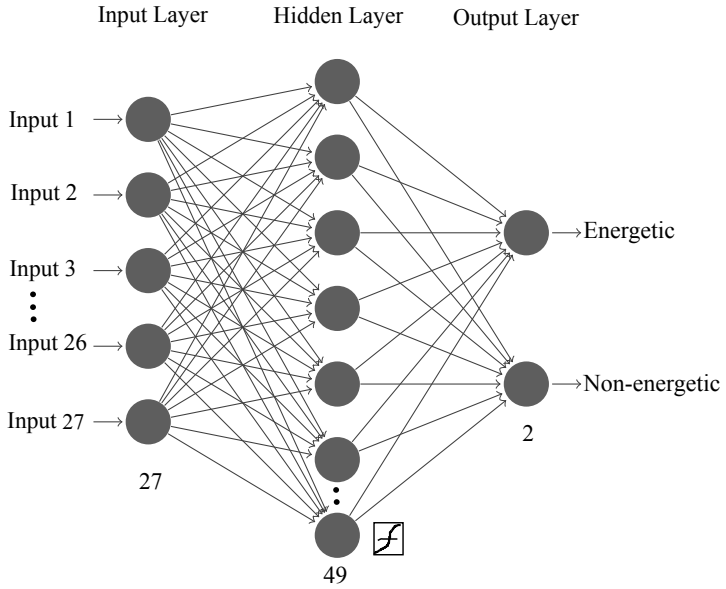

Figure 5.3: The structure of ANN considered for level-1 classification.

5.2.2 Level-2 Classification

In the second level classification, the spectrograms w.r.t. energetic classes namely, *Angry*, *Energetic & Happy*, and non-energetic classes namely, *Devotional*, *Romantic*, and *Sad* have been constructed. These spectrograms are fed to the Convolutional Neural Networks (CNNs) to estimate the actual mood class of a given clip. Details of CNN are already provided in Section 4.3.4, hence they are ignored here to avoid the unnecessary repetitions.

5.3 Experimental Analysis

The details of the dataset considered for experimentation on mood estimation is presented in Section 2.9.5. It contains the clips of six different emotions namely *Angry*, *Devotional*, *Energetic*, *Happy*, *Romantic*, and *Sad*. Since the process of labelling mood information of a song is difficult, help from music professionals has been taken. Further, Mean Opinion

Score (MOS) is considered to label the audio clips. A Majority voting has been considered for labelling the mood of an audio clip. Initially, the dataset has been divided into training and testing sets, where the training set contains 100 clips of length 3 to 5 seconds each. The length of test clip varies from 30 to 60 seconds. The background accompaniment has been suppressed using Infinite Impulse Response (IIR) filtering techniques. All the audio clips considered for experimentation are collected from high-quality audio CD's recorded at 44,100 Hz.

The first phase of the system is developed using six different emotion classes. The statistical values of pitch (4), jitter (1), shimmer (1), Short-Term Energy (STE) (1), harmony (1), and BPM (1) values have been computed as features. They are added to the baseline MFCCs (13) forming a feature vector of length 23. The features have been computed from the frames of length 25 ms. The reason for choosing the above-mentioned features is their performance in recognizing speech emotions (Koolagudi and Rao, 2012; Jacob, 2016; Koolagudi et al., 2018). Moreover, the distinction among the features for different emotions has been given in Table 5.1. Further, Artificial Neural Networks (ANNs) are trained to classify the moods. The results obtained using this approach are given in the form of confusion matrix as shown in the Figure 5.4. The values of confusion matrix have been represented pictorially here. The intensity values of black colour indicate the classification strength. From the figure, one can see that this approach may not be able to accurately recognize the moods of music. There are many misclassifications between *happy* & *energetic*, *happy* & *anger*, and *romantic* & *sad*.

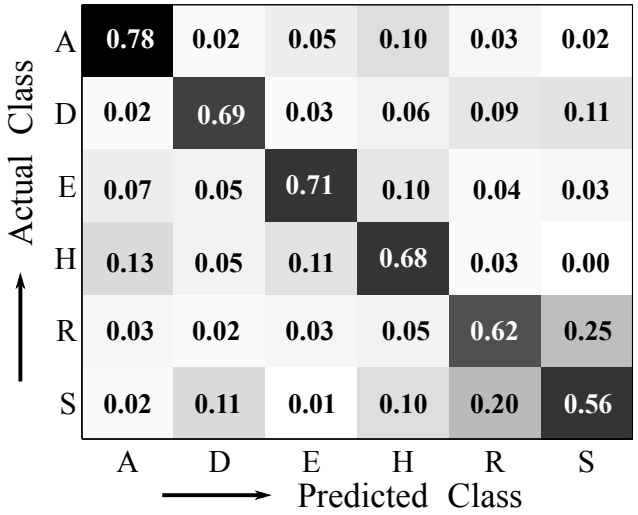

Figure 5.4: Pictorial representation of confusion matrix obtained while classifying six moods using NN classifier. The classification accuracy of Actual Vs Predicted classes.

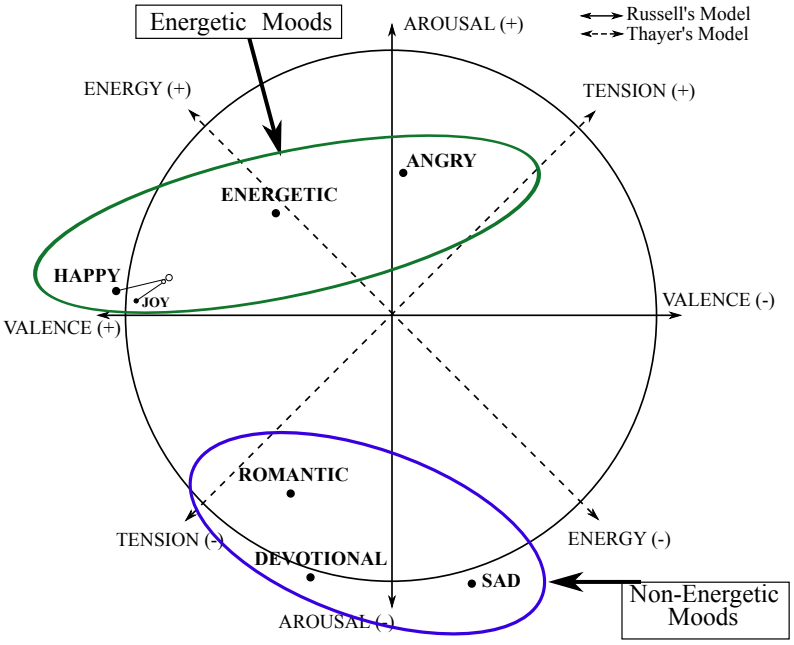

Figure 5.5: Broad categorization of six moods (*Angry*, *Devotional*, *Energetic*, *Happy*, *Romantic*, and *Sad*) into *energetic* and *non-energetic* classes.

The average accuracy obtained with this approach is around 67.20%, which is quite low and may not be suitable for real-time applications. To improve the mood recognition accuracy, a two phase system classification approach is proposed; for the first phase the songs clips are classified into broad classes such as *energetic* and *non-energetic*. These groups are formed based on the position of moods in the valence-arousal plot; shown in Figure 5.5. The *energetic* category contains the moods of *Angry*, *Energetic*, and *Happy*. The remaining moods, *Devotional*, *Romantic*, and *Sad*, belong to *non-energetic* category. A two-level classification model has been proposed, where the first-level classification categorizes the given input audio clip into either *energetic* or *non-energetic* class. Features such as MFCCs, Short-Term Energy (STE), Beats Per Minute (BPM), and chroma values are computed forming a feature vector of length 27. ANN classification model has been used, with one hidden layer, to classify the given feature vector into one of the two classes. The accuracy obtained, to classify the input into any of the two categories, is 92%. Once the first-level classification is complete, the actual mood is predicted using trending Convolutional Neural Networks (CNN).

CNNs have proved effective over traditional neural networks in the case of image processing. They perform well for audio classification as well (Hershey et al., 2017). However, based on their expertise in image classification, due to the involvement of convolution layers, spectrograms of the audio clips have been fed as an input to the CNNs. The reason

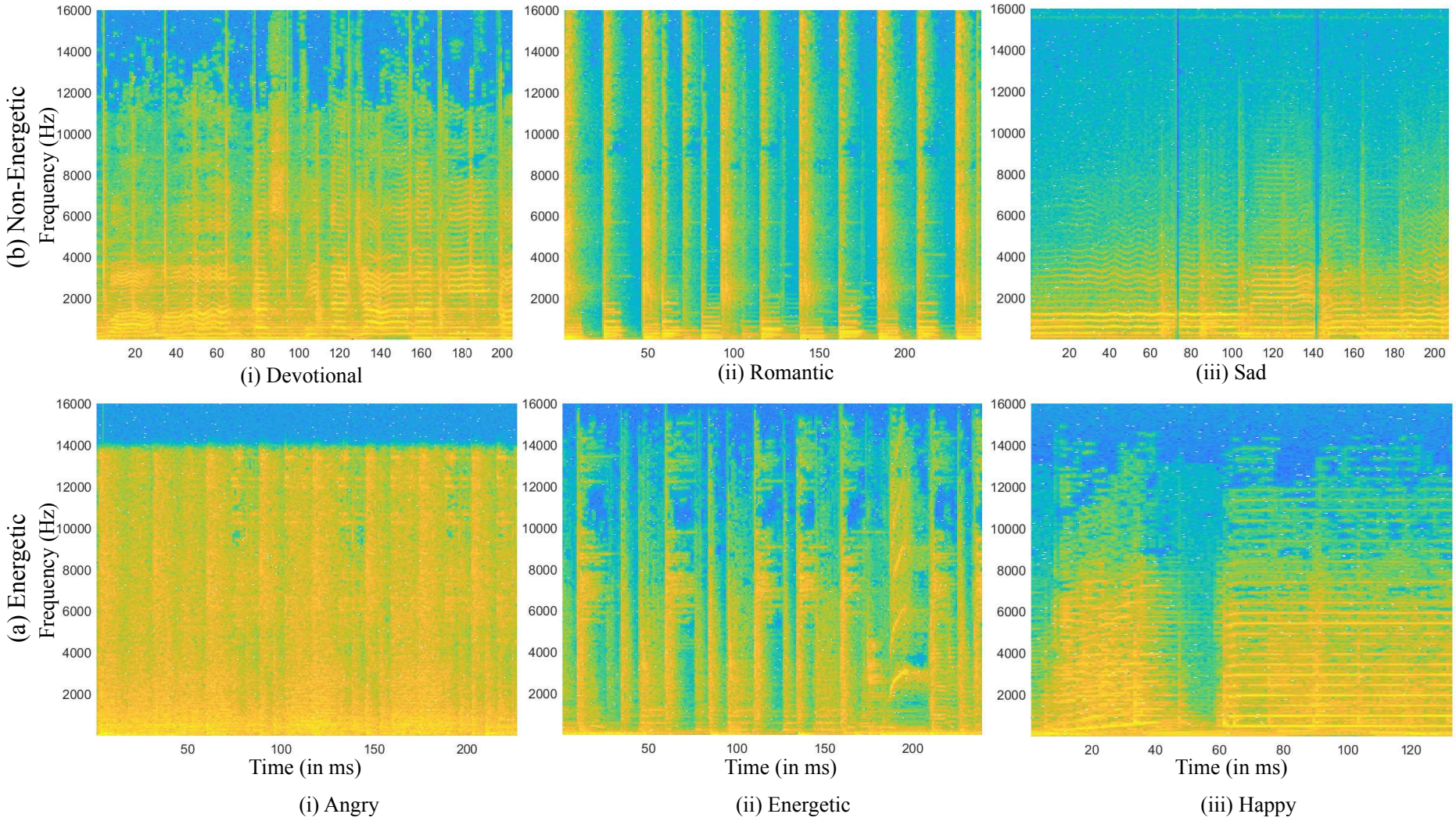

Figure 5.6: The structural differences in the spectrograms observed during the analysis of *energetic* and *non-energetic* moods.

for constructing spectrograms is their energy variations at different frequency ranges for different moods. The sample spectrogram with energies, for a range of frequency values over the time, is given in Figure 5.6. In which, the spectrograms for six moods of the two categories are given in the form of a table. Each row of the table represents category of moods whereas each column is the actual mood which is corresponding with their category. One can clearly observe the discrimination among the spectrograms of each mood class of a particular category. The spectral energy values are found to be more for some moods. Whereas, they are low in case of other moods. Hence, spectrograms have been considered, to train and test the CNNs, for better performance.

Table 5.2: Hyperparameters considered for designing the CNN for the task of mood classification.

Sl.No.	**Parameter**	**Value**
1.	Batch size	8
2.	No. of channels	3 channels (RGB)
3.	Filter size	3*3
4.	Image size	256*256
5.	No. of hidden layers	4
6.	No. of flatten layers	2
7.	Softmax layer	1
8.	No.of output classes	3
9.	Activation function	*tanh* & RELU
10.	No. of epochs	Around 50-60

To obtain a spectrogram, each audio clip is divided into 25 ms frames with an overlap of 10 ms. Short-time Fourier transform is applied to obtain frequency spectrum. They are placed on a time scale together to form a spectrogram and the resultant spectrogram is integrated into 64 Mel-spaced frequency bins. Each of these magnitudes is transformed into a logarithmic domain. During this process, the numerical issues are bypassed, using offsets. For training, 70% of the dataset has been chosen randomly and remaining 30% is considered for testing. For implementing training, CNN, TensorFlow (Abadi et al., 2016) have been used and trained asynchronously on multiple GPUs. Grid searching has been performed on batch sizes, parameters, number of GPUs, and learning rates.

However, there is no need to consider multiple GPUs for this dataset since the size is very low. Softlayer classification has been chosen since each output class is corresponding to a single mood. The same CNN architecture has been considered for classifying the mood classes of both *energetic* and *non-energetic* mood categories. The complete details about the hyperparameters of CNN are given in Table 5.2.

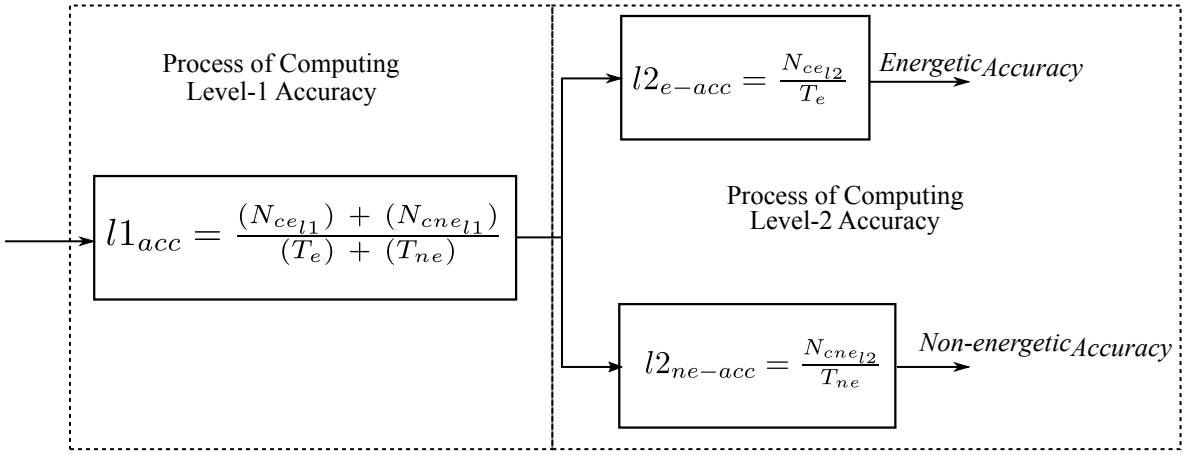

Figure 5.7: The process of computing accuracy values for *energetic* and *non-energetic* categories of moods.

Mood estimation and accuracy are computed as a performance metric to evaluate the system's performance. The process of computing accuracy values at the first level and the second level is depicted in Figure 5.7. In level-I all the test clips are initially divided into two classes namely total clips of *energetic* (T_e) and *non-energetic* (T_{ne}) categories. The total number of correctly identified *energetic* clips ($N_{ce_{l1}}$) and *non-energetic* clips ($N_{cne_{l1}}$) are summed together and divided by $((T_e) + (T_{ne}))$ to obtain the accuracy ($l1_{acc}$) at level-1 stage. At the second level, the accuracy values are obtained separately for *energetic* and *non-energetic* classes. The terms that are used in the level-2 classification, are $l2_{e-acc}$ which indicates the level-2 *energetic* accuracy; $l2_{ne-acc}$ is the level-2 *non-energetic* accuracy, $N_{ce_{l2}}$ represents the number of *energetic* mood classes that is correctly identified at level-2, and $N_{cne_{l2}}$ is the number of *non-energetic* mood classes correctly identified.

The accuracy is computed by considering the misclassifications that have been taken place at the first level. The misclassified clips are manually removed and given to CNN for obtaining the accuracy which suits for real time. The average accuracy values of 87.2% and 83.4% are obtained for the moods of *energetic* and *non-energetic* categories respectively. If misclassifications are considered, then the accuracy values drop down to 80.22% and 76.72%.

5.4 Summary

In this chapter, the task of mood estimation from the music clips is addressed using two-level classification model. Since emotion is the third most important aspect of music information retrieval, an effort has been made to recognize the moods in Indian songs. Six different moods namely *Angry*, *Devotional*, *Energetic*, *Happy*, *Romantic*, and *Sad* are proposed, based on the analysis done on *Russell's* and *Thayer's* model. Initially, these moods have been broadly categorized into *energetic* and *non-energetic* classes. In the first level, the given audio clip is classified into either of these two classes. Further, actual mood label is predicted at the second level classification. Acoustic features such as MFCCs, Chroma, Tempo, and Harmony are considered to develop the first level classification model using ANNs. For the second level, spectrograms have been constructed and are fed to CNN, to help recognise the music samples for their mood classes of *energetic* as well as *non-energetic* categories.

Chapter 6

Music Recommender System using Graph Structures

> *Music Recommender Systems (MRS) are important drivers in music industry to estimate the listener's behaviour, to identify the people who are having similar nature.*
>
> — Christine Bauer

6.1 Introduction

Music is a powerful communication tool that keeps people relaxed. It is said that music is highly an engaging activity for humans when compared to other activities like watching movies, reading books, playing games, watching TV and so on. (Song et al., 2012). A huge number of tracks has been introduced in online and offline stores during the past few decades. This enormous inclusion introduces two latent problems: (i) difficulty in music organization, (ii) difficulty in recommendation.

The problem in music organization is mainly due to the unavailability of complete meta-information for millions of tracks. The process of manual labelling can take several years and is a never ending process, as new songs are added to the repository at a higher rate than before. Hence, the task of Music Information Retrieval (MIR) has gained importance as it extracts useful meta-information about every song such as artist identification, genre classification, mood estimation, raga identification, instrument identification, music annotation, and so on. Several research works have been reported with the coordination of MIREX to build a robust MIR systems. The other issue is selection of relevant songs,

as per the need, once meta-information is obtained. It is very difficult to identify the songs that better suit the listener, from the enormous list of a variety of songs. In such cases, recommender systems can help in identifying the next song to be chosen by the listener based on his/her previous statistics.

Recommender system is a software tool that can select or suggest the items which suit the user's needs (Ricci et al., 2015). The suggestion may be; what product to buy, what kind of books to read, what movies to watch, what news to read, what sports news to read, what TV programs to watch, what music to listen to, what places to visit, who can be in the friends list, and so on (Mahmood and Ricci, 2009; Resnick and Varian, 1997; Burke, 2007). The process of developing a recommender system is different for each different applications. For instance, the system which is developed for recommending books or movies may not suit the music recommendation. Hence, different analysis and approaches have to be proposed for developing a music recommender system.

The methods used for music recommendation have been named broadly as demographic filtering, collaborative filtering, content-based filtering, context-based filtering, and hybrid filtering. Each approach has its own limitations (Lu et al., 2015). There is one more important factor, called type of listener, which is to be considered while implementing a recommender system. There are four types of listeners reported in the literature namely the savants, the enthusiasts, the casuals, and the indifferents. In literature (Herrada, 2009), it has been said that on an average there are 7% savants, 21% enthusiasts, 32% casuals, and 40% indifferents are there in the whole world. While developing music recommender systems, one of the three different approaches, namely; user-centric, system-centric, or network-centric approach may be used.

After reading through the literature, an approach has been proposed using graph structures. This approach estimates the next song to be chosen by the $listener_x$[1] by obtaining the similarity scores with other $x-1$ users. The songs chosen by the $listener_y$ ($\forall\ y = 1, 2, \ldots x-1$) with high similarity score, are more suitable for recommending them to $user_x$. Cosine similarity is used to obtain similarities among listeners. Further, listener feed back has been taken to improve the system performance.

[1]The term *listener* and *user* are interchangeably used in this thesis.

6.1.1 Applications

Music processing has many useful applications in day to day lives with different dimensions. Based on listener's past behaviour, it is possible to analyze their mental state. For instance, if the listener prefers to listen to mostly sad songs, then the listener may be in a negative state of mind. Sometimes it is possible to recommend songs based on listener's present location. For instance, if the listener is driving a vehicle, it is better to skip all slow songs or lullabies. This work has tried to categorize listener's, based on the similarities in their interests. Many a times, songs can be recommended based on the other listeners of similar interests. This approach also ensures to overcome the issue of losing many popular songs.

6.1.2 Challenges:

Developing a music recommender system has many challenges since it is a distinct task compared to other recommender systems. Generally song selection criteria of a listener depends on their mood, it is difficult to predict their next selection. The interests of an individual also change over time. For instance, one may show interest in listening to energetic songs when young. However, as they become old, they may prefer slow music or devotional music. Hence, the recommender system of one age group may not be suitable to that of another age group. It is also very difficult to implement recommender system for every individual. In this work, a graph structure based recommendation scheme has been used that mainly behaves more objectively to gender,age and other similar issues; more often irrespective of their age, gender etc.

6.2 Factors and Issues that are to be Considered while Developing an MRS

There are certain factors which are to be considered while developing a recommender system. Few important prerequisites are datasets, creation of user profile, meta-information for recommendation (e.g., based on artist, title, mood, etc.), recommendation methods, and performance evaluation strategies. The details of each of these are given below along with the considerations made in this study.

Database is an important factor for any recommender system. While implementing music recommender system for developing MRS, *Last.fm* is the popular website which

Behaviour Table{*index*, *listener_id*, *song_id*, date, time, frequency}

index	*l_id*	*song_id*	*date*	*time*	*freq.*
001	102	1523	23/05/18	02:30	15
002	136	1329	24/06/18	08:10	24
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
m*n	152	1649	10/07/18	23:24	49

Song Table{*song_id*, song_name, album_name,length, *singer_id*, mood}

l_id	*l_name*	*age*	*gender*
101	listener_1	25	M
102	listener_2	52	F
.	.	.	.
.	.	.	.
.	.	.	.
200	listener_m	35	M

song_id	*song_name*	*album*	*len(s)*	*s_id*	*mood*
1001	song_1	movie_1	180	001	Happy
1002	song_2	movie_2	315	015	Angry
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
1700	song_n	movie_n	224	007	Sad

Listener Table{*listener_id*, listener_name, age, gender)

Singer Table{*singer_id*, singer_name, gender, language}

s_id	*s_name*	*gender*	*lang*
001	singer_1	M	T
002	singer_2	F	H
.	.	.	.
.	.	.	.
.	.	.	.
020	singer_k	F	T

Figure 6.1: Different tables constructed to implement music recommender system. *Note:* In table acronyms are provided due to alignment issues. *l_id* → *listener_id*, *l_name* → *listener_name*, *s_id* → *signer_id*, *s_name* → *singer_name*, *len(s)* → *length (in seconds)*, and *date*, *time* & *freq.* → *date, time, & frequency of download* respectively.

is used by many researchers. However, it contains only the meta-information, limited to artist name, song title, album name, and download time stamp. In this work, a dataset has been constructed with 100 listeners and 700 songs. The details of the database are given in schema (shown in Figure 6.1).Four different tables are constructed named **SINGER**{*singer_id*, singer_name, gender, language}, **LISTENER**{*listener_id*, listener_name, age, gender }, **SONG**{*song_id*, song_name, album_name, length, *singer_id*, mood }, and **BEHAVIOUR**{index,*listener_id*, *song_id*, date, time frequency}. Initially,

the tables for singers, listeners, and songs have been constructed with 20-30 singers, 100 listeners, and 700 songs. Based on the information collected from the above three tables, another table called **BEHAVIOUR** has been constructed to know the listeners behaviour and their interests. The attribute underlined in the table indicate either primary[2] or foreign key[3]. In Figure 6.1, primary and foreign keys are indicated with thick and a dash respectively.

While creating the user profile, it is also important to decide the meta-information which is highly important for further processing. The listener may listen to the songs based on a particular singer, and his/her mental state. Some listeners listen to the songs casually, *called* casual listeners. Many other factors influencing recommendation include interest in melodies, particular genre, instruments, language, singer's gender, signer's age, title of the song, popularity of the song, song rating and so on effecting the performance of recommender system. Of these, singer information and mood of a song clip are two important attributes that are often considered by a listener to change a song. (Ghatak, 2012). This work has also constructed a **SONG** table with *title of the song, album_ name, singer name, mood, and length (in seconds).* Moreover, *Gender*, and *age* information is also collected for each singer in the **SINGER** table, which is a parent table for **SONG**.

Another important factor which is to be considered while implementing recommender system is the method chosen for developing it. There are few recommendation methods namely collaborative filtering, content-based, context-based, hybrid recommender systems that have gained much importance in recent years (Wang and Wang, 2014; Horsburgh et al., 2015; Chiliguano and Fazekas, 2016). However, each technique has its own limitations. Collaborative filtering is suffering from a "*cold-start* problem", content-based filtering is with computational complexity issues, and context based recommendation from issues of social tagging and Gray-sheep problems. A major problem of the hybrid system is to decide trade of between issues. For instance, there are three users $\{u_1, u_2, \& u_3\}$ interested in the items $\{\{i_1, i_2, \&\ i_3\}, \{i_2, i_3, \&\ i_4\}, \&\{i_5, \&\ i_6\}\}$ respectively. In such a case, it is possible to recommend the items of u_1 to u_2 and vice-versa. Since $u_3 s$ interests are not matching with any other, the recommendation system fails to recommend any item to u_3. This problem is generally called as "*Gray sheep* problem" (Ghazanfar and Prügel-Bennett, 2014).

[2]It ensures the uniqueness and not null properties for the column.
[3]It keeps track information of other tables based on the unique value given.

In this work, a graph based recommendation method is proposed as a base work for future research. It can push the future researchers to focus on listening path, number of times the same song is listened to, other factors affecting while selecting the next song, whether the song has been listened to completely or skipped in between, etc. The present recommender systems have considerably failed to obtain better performance since, they are depending on the user ratings (Isinkaye et al., 2015). In graph based recommender system, the analysis of a user listening graph can automatically give a hint on the performance of recommender system. However, the work done in this thesis is limited to the basic recommender system implemented using graph structures. The details of the proposed graph model for recommender system is given in Section 6.4.

6.3 Basic Terminology

A graph $G\{V, E\}$ is a data structure which is used to represent the information having set of vertices $V\{1, 2, \ldots n\}$ and edges $E\{e_1, e_2, \ldots e_t\}$. In general, adjacency matrices are used to represent the graphs where the the number of rows and columns of the matrix is equal to the number of vertices (n). The value of corresponding row i and j is set to '1' if there is an edge, e_{ij} exists between V_i and V_j. In this work, the number of songs of the cloud are considered as nodes[4] of the graph. The adjacency matrices are produced for each $listener_x$ ($\forall x = 1, 2, \ldots, m$) and the value e_ij is set to '1' if the listener is listening to j^{th} song after i. Further, sparse matrix has been constructed since the number of zero elements are found to be more in the adjacency matrix as the listener is incapable to listen all the songs of database in a single life.

6.3.1 Sparse Matrix

A sparse matrix is a compressed version of matrix if there are more number of zeros in the original matrix. The representation of sparse matrix is $SP(nz, 3)$, where, nz is the number of non-zero elements. The first and second columns hold the row and column indices respectively. The third column has a non-zero value. For instance, the original matrix is of size (25*25) and has 20 non-zero elements. Assuming that each element needs 4 bytes of memory space, the total size needed by the original matrix is equal to 25*25*4 = 2500 bytes ($\approx 2.44KB$). The same can be represented in (20*3) size with the help of sparse matrix. The size needed for sparse representation is 20*3*4 = 240 bytes ($\approx 0.234KB$),

[4]The terms *vertex* and *node* are interchangeably used in this thesis.

which is approximately consumes 90% less size. The space complexity issues are resolved using sparse matrices in this work. An algorithm for the sparse construction is given in Algorithm 3.

6.4 Proposed Methodology

The complete process of proposed music recommendation system is given in Figure 6.2. Initially, adjacency matrices are constructed for listeners $\{1, 2, \ldots, m\}$, in which, the number of rows and columns are equal to the number of songs (n) considered to construct a graph. Further, each listener's ($listener_i$)$\{\forall i = 1, 2, \ldots, m\}$ adjacency matrix is compared with all adjacency matrix of $listener_j$, $\forall j = 1, 2, \ldots, m$ such that $i \neq j$. Row vector similarity technique has been used to obtain the cosine similarity among two rows of an adjacency matrix. Since a majority of cells have *zeros (0s)* in the adjacency matrix, sparse matrix has been constructed to reduce the memory space as well as the issues of computational complexity.

Each sparse matrix represents a listener's behaviour in terms of songs heard. Further, two sparse matrices are analysed to obtain the similarity among the listening patterns of two listeners i and j. If two listening behaviours of the listeners i and j are similar, beyond certain level, then the songs heard by the listener may be recommended to the others. Based on the high similarity values obtained among different listeners, the songs heard by these listeners have been chosen to fill the playlist of others in the different groups. To evaluate the statistically of recommendation, rating from listener has been taken out of 5 for each song of the generated playlist. An average rating of 3.5 has been obtained for the proposed recommendation approach on a list of 35 listeners. The details of obtaining similarity metric along with the analysis of the proposed algorithm are given in the subsequent subsections.

6.5 Similarity Metric

Consider a dataset consisting of n songs and m listeners. This work mainly focuses on building a recommender system using the inherent similarities found in listening patterns of the users. A nXn matrix is formed by using the songs in the database. Rows and columns of the matrix indicate the songs numbered 1 to n. The contents of the matrix are either '1' *or* '0' based on the listeners choice of songs. If the listener prefers to listen

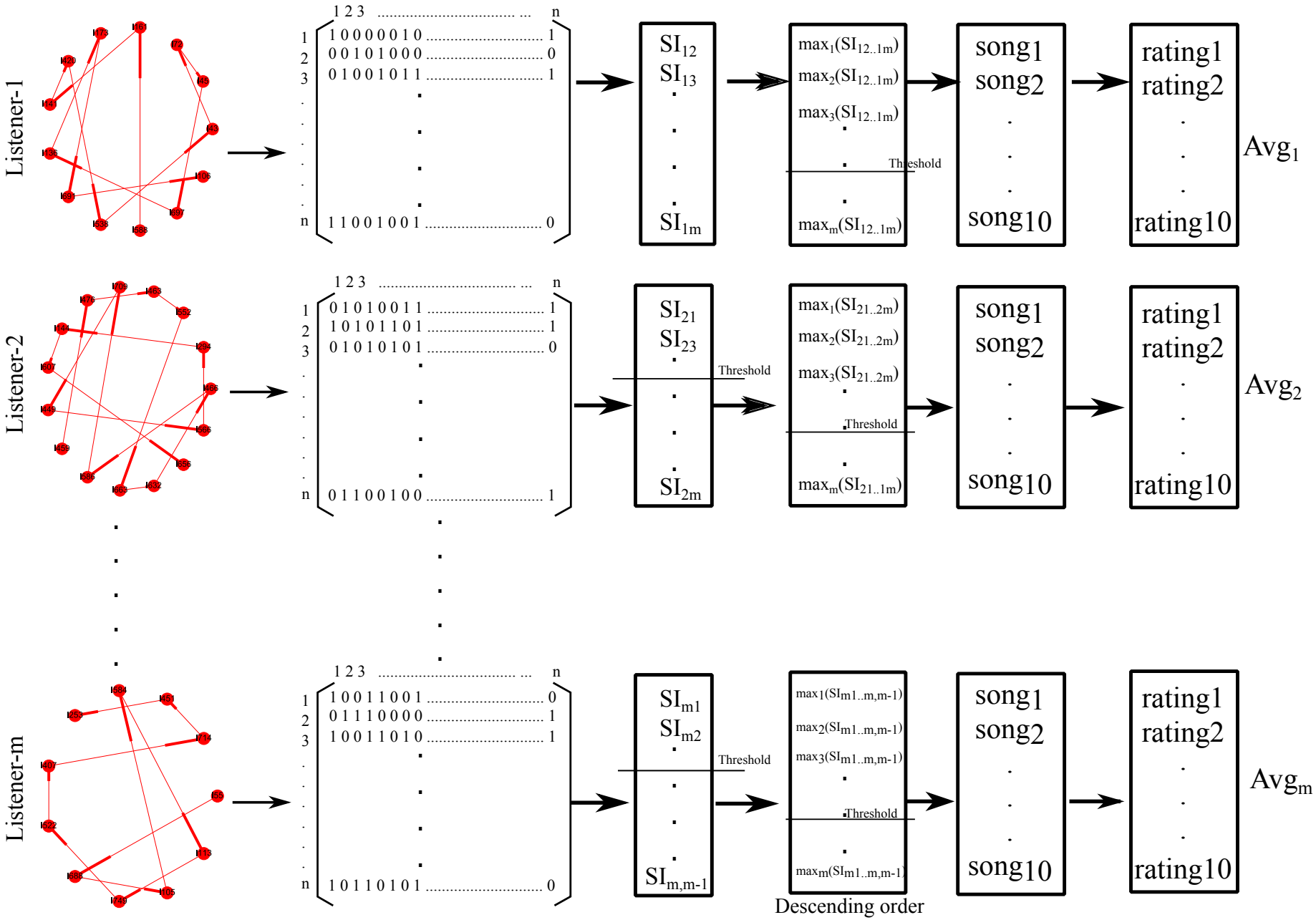

Figure 6.2: Proposed flow diagram for generating the recommended playlist for listeners. Darkness in line indicates an arrow.

to j^{th} song $\{\forall j = 1, 2, \ldots n\}$ after i^{th} song $\{\forall i = 1, 2, \ldots, n\}$, then the i^{th} row and j^{th} column in the matrix is made '1'. An adjacency matrix has been obtained and they are formed for every user.

The comparison of such matrices, for similarity, can help in recognising similar listeners recommends their songs to each other. The matrices are tallied using *cosine similarity* which gives similarity in the listening patterns of two persons. This indeed helps in recommending the next song for the users of similar patterns.

Consider two listeners A and B and two adjacency matrices X and Y assigned to them. Similarity in their pattern can be recognised by using *cosine similarity*. This is obtained by extracting the i^{th} row of X and i^{th} row of Y matrix and computing the dot product on the vectors. Later, the value is normalised by dividing with individual magnitudes of the vectors. The formula used to obtain the similarity metric among two given rows of matrices is given in Eq. 6.1.

$$Similarity(X_i, Y_j) = \frac{\overrightarrow{X_i}.\overrightarrow{Y_j}}{||X_i||.||Y_j||} \tag{6.1}$$

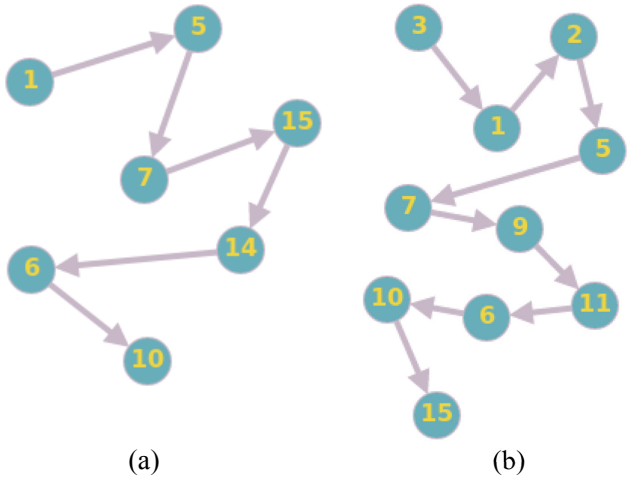

Figure 6.3: Graphs considered for example given in Figure 6.4.

Figure 6.4 pictorially explains the process of similarity metric computation. Two graphs given in Figure 6.3 are used to compute the listening pattern. Only vertices having edges are considered to form Figure 6.3, as the complete graph with all nodes may confuse the perception of the reader. Graphs 6.3(a) indicates the $listener_1$ behaviour and 6.3(b) is for $listener_2$. Initially, the adjacency matrices are constructed to compute the similarity between them, as shown in Figure 6.4(a) & (b). Later, sparse matrices have been computed using Algorithm 3, resulting in the matrices shown in Figure 6.4(c) & (d).

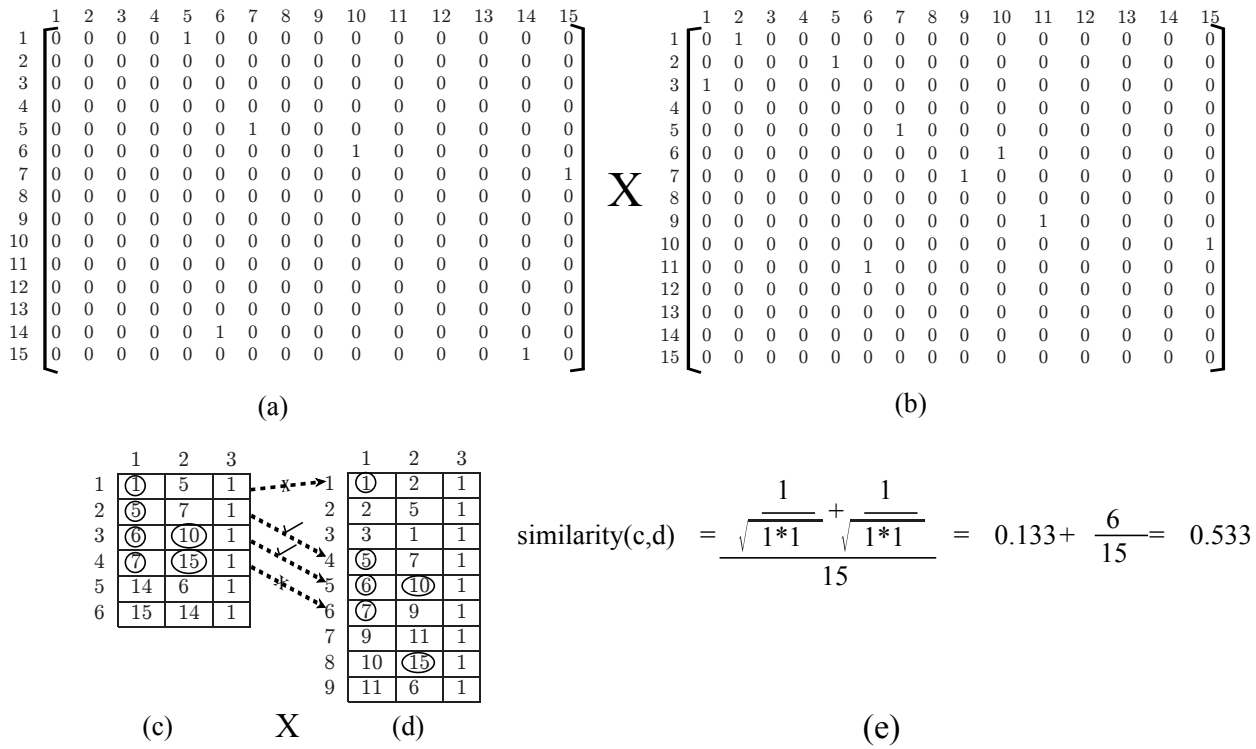

Figure 6.4: An example illustrating the process of computing similarity between two adjacency matrices using sparse matrices. *Note:* This example is given to explain the process of all three 2, 3, 4 algorithms.

Further, similarity has been computed using Algorithm 4. In this example, no values is similar hence, the similarity between two graphs is zero, as shown in Figure 6.4(e).

6.6 Algorithm Analysis

Three different algorithms are developed to perform different tasks of song recommender system in order to recommend songs to the listeners. *mainFunction()* takes adjacency matrices of m listeners as input. The size of each adjacency matrix is $n * n$. The adjacency matrix of each $listener_i$ $\{\forall i = 1, 2, \ldots, m\}$ is compared with adjacency matrices of $listener_j \{\forall j = 1, 2, \ldots, m\}$ and $i \neq j$ (see line 1 and 5 in Algorithm 2). For each listener, sparse matrix has been constructed as the number of non-zero elements in adjacency matrix. The sparse represents the same adjacency matrix in less space. For instance, if there are 9 non-zero elements in Figure 6.4(b), to represent the same using adjacency matrix, a memory space of $15*15*4 = 900$ bytes is needed (Assuming each element needs 4 bytes of memory). The same can be represented in $9*3*4 = 108$ bytes using sparse representation (shown in Figure 6.4(d)). If n value is in lakhs, then the reduced sparse matrix is capable of occupying very less space when compared to original adjacency matrix. The process of constructing a sparse matrix is explained in Algorithm 3.

Algorithm 2: Main() :: Music Recommender System

Input: Adjacency_matrix[n][n] for m listeners.
Output: Recommended playlists for m listeners.

```
1  for i ← 1 to m do
2      sim[m − 1] ← 0;
3      sim_j ← 1;
4      sp1 ← constructSparse(adjMatrix[i]);
5      for j ← 1 to m − 1 do
6          if i ≠ j then
7              sp2 ← constructSparse(adMatrix[j]);
8              sim[sim_j + +] = checkSimilarity(sp1, sp2);
9      findAvg(sim);
       /* Apply threshold to find most similar listeners                    */
10     generatePlaylist[i](songs(similar));
```

Algorithm 3: Sparse Matrix Construction

Input: Adjacency matrix[n][n]
Output: Sparse matrix[nz][3]

```
;                        // Stores only positions of non-zero (nz) elements
1  Function constructSparse(adjMatrix[n][n]):
2      nz ← countNonZero(adjMatrix[n][]n);
3      sparse[nz][3] ← 0;
4      spi = 1;
       // Sparse Index
5      for i ← 1 to n do
6          for j ← 1 to n do
7              if adjMatrix[i][j] ≠ 0 then
8                  sparse[spi][1] = i;
9                  sparse[spi][2] = j;
10                 sparse[spi + +][3] ← adjMatrix[i][j];
11     return sparse[nz][3];
12 Function countNonZero(adjMatrix[n][n]):
13     nz ← 0;
14     for i ← 1 to n do
15         for j ← 1 to n do
16             if adjMatrix[i][j] ≠ 0 then
17                 nz + +;
18     return nz;
```

A sparse matrix is equal to the size of $(\#nz * 3)$, where, $\#nz$ is the number of non-zero elements. Each row of the sparse matrix has three columns: row index, column index, and $value(row, column)$ of a non-zero element. To construct the sparse matrix,

[5]The term nz is equal to the number of rows in sparse matrix *or* number of non-zero elements of adjacency matrix.

Algorithm 4: Similarity checking b/w two sparse matrices.

```
1  Function checkSimilarity(sp1[r][3], sp2[s][3]):
2      r1 ← max(unique(sp1[][1]));
3      s1 ← max(unique(sp2[][1]));
4      i ← 1, j ← 1, sim_i ← 0;
5      sim[min(r1, s1)] ← 0;
6      while (i < r1) and (j < s1) do
7          dotProd ← 0, sq1 ← 0, sq2 ← 0;
8          temp1 ← i, temp2 ← j;
9          val1 ← sp1[i][1], val2 ← sp2[j][1];
10         while val1 == sp1[temp1][1] do
11             temp1 + +;
12         while val2 == sparse2[temp2][1] do
13             temp2 + +;
14         min_t ← min(temp1, temp2);
15         if min_t == temp1 then
16             k ← i, l ← j;
17             while k ≤ temp1 do
18                 m ← col_match(sp1[k], sp2[l...temp2]);
19                 if m > 0 then
20                     dotProd+ = (sp1[k][3] ∗ sp2[m][3]);
21                     sq1+ = (sp1[k][3] ∗ sp2[k][3]);
22                     sq2+ = (sp2[m][3] ∗ sp2[m][3]);
23                     k + +;
24                     if m < temp2 then
25                         l = m + 1;
26                     else
27                         break;
28         else
29             k ← j, l ← i;
30             while k ≤ temp2 do
31                 m = col_match(sp1[l...temp1], sp2[k]);
32                 if m > 0 then
33                     dotProd+ = (sp1[m][3] ∗ sp2[k][3]);
34                     sq1+ = (sp1[m][3] ∗ sp2[m][3]);
35                     sq2+ = (sp2[k][3] ∗ sp2[k][3]);
36                     k + +;
37                     if m < temp1 then
38                         l = m + 1;
39                     else
40                         break;
41         sim[sim_i + +] = (dotProd/sqrt(sq1 ∗ sq2));
42         i = i + (temp1 − i), j = j + (temp2 − j);
43     return avg(sim);
```

Table 6.1: Time complexities in terms of *Big Oh* notation that are consumed by processor for evaluating algorithms.

Sl. No.	**Algorithm**	**#Max. operations**	**Time Complexity**
1.	constructSparse() Algorithm - 3	Two times n^2 which is equal to $2n^2$.	$O(n^2)$
2.	checkSimilarity() Algorithm -4	The number of multiplications in sparse matrix is equal to $(nz^2)^5$ in the worst case. Where, adjacency matrix needs n^2 operations.	$O(nz^2)$
3.	mainFunction() Algorihtm - 2	Each listener's adjacency matrix is compared with $(m-1)$ listeners. It happens for m number of listeners. So, the number of iterations is equal to (m^2-m). Moreover, *constructSpace()* and *checkSimilarity()* are called inside the main function.	$O(m^2.n^2)$

entire adjacency matrix has to be traversed. Hence, the time complexity needed for constructing a sparse matrix is $O(n^2)$, as shown in second row of Table 6.1.

For similarity check of two sparse matrices, the listening behaviour of *cosine similarity* is computed on the two matrices. This algorithm computes the similarity for each row and finally finds the average which is treated to be the similarity of two graphs. The cosine function operated on two adjacency matrices, has the time complexity of $O(n^2)$ and the number of multiplications computed is also equal to n^2. This can be performed in (nz^2) time on sparse matrix.

According to algorithm 4, if the first columns of two sparse matrices are the same, then, only the similarity among the subsequent next column is computed and compared. If rows and columns of first and second matrices are same, then multiplication happens. In case of direct multiplication of adjacency matrices, every value of one adjacency matrix is multiplied with the other one. This process needs $O(n^2)$ multiplications. The same operation is possible using sparse matrix in less time which is almost equal to $O(nz^2)$ in the worst case. The details of calculating time complexity for Algorithm 4 are given in 2^{nd} row of Table 6.1.

Since both the *constructSparse()* and *similarityCheck()* are called from *mainFunction()*, the total time needed to process the complete program is $O(m^2.n^2.nz^2)$. The same process may take $O(m^2.n^4)$ if adjacency matrices are considered for similarity check.

The time complexities have been computed for each algorithm and are mentioned in

Table 6.1. Initially, time complexity values are estimated for the functions called inside the loop of *mainFunction()*, i.e, Algorithm 2. The time needed to construct a sparse matrix (Algorithm 3) is $O(n^2)$,as given in row one. The similarity checking algorithm (Algorithm 4) needs $O(nz^2)$ to check the similarity values of two given sparse matrices, shown in second row of Table 6.1. Hence, the time taken by Algorithm 2 is $O(m^2.n^2.nz^2)$ which is shown in the third row of the table.

6.7 Summary

In this work, a graph based recommendation approach has been proposed to analyse the listening pattern of music listeners. The existing recommendation approaches have some major issues like "*cold-start problem*", "Gray-sheep problem", complexity issues, etc., Four different tables called **SINGER**, **SONG**, **LISTENER**, and their **BEHAVIOUR** are designed and used. Sequence of hearing songs is represented as a adjacency matrix. *cosine similarity* among the adjacency matrices is used to estimate similarity in listening behaviour. The **SONG** table has been provided, with additional meta-information like duration (in seconds), singer, and mood information, to properly estimate the listener's behaviour. Further, the playlist of 10 songs has been generated, based on the listening similarities. User ratings are taken for every song of playlist, to evaluate the system performance. An average rating of 3.5 is obtained using this approach.

Chapter 7

Summary, Conclusions and Future Work

> *Reasoning draws a conclusion, but does not make the conclusion certain, unless the mind discovers it by the path of experience.*
>
> — Roger Bacon

This chapter concludes the work along with some possible future research directions. Since the MIR system is still under construction and is dependent on many subtasks such as singing voice detection, singer identification, composer recognition, instrument identification, genre classification, raga identification, mood estimation, music annotation, and so on, implementation of each system has its own contribution towards its development (Murthy and Koolagudi, 2018a). In this thesis, a few important tasks such as vocal and non-vocal segmentation, singer identification, and music mood estimation are considered while developing a music recommender system.

This chapter contains summary of the work addressed in this thesis, learning outcomes are given as conclusions and some issues are highlighted as directions for future research.

7.1 Summary and Conclusions

This section gives conclusions made out of each section in a detailed manner.

7.1.1 Vocal and Non-vocal Segmentation

An approach has been proposed to select the relevant and suitable features for the task of vocal and non-vocal segmentation. An evolution based genetic algorithm (GA) has

been proposed for feature selection. To develop such system for vocal and non-vocal segmentation a database with relevant tracks is essential. In this work, two datasets namely the standard MIR-1K dataset, and the other is TBPS dataset collected from *Tollywood* and *Bollywood* songs of Indian film industries. Some novel features have been computed such as formant height from base-to-peak (FH1), formant angle at peak (FA1), and valley (FA2) on top base-line features have been computed, that are forming a 93-dimensional feature vector. It is practically not possible to consider high dimensional feature vectors for real-time applications due to the issues of response time and complexity. Hence, an approach has been proposed to select the relevant features among 90-dimensional feature vector and to reduce the dimensionality of the feature set. Various feature selection techniques have been chosen to compare the results with evolutionary based feature selection which uses the concepts of genetics. It is observed that the evolutionary algorithms are highly efficient when compared to the correlation-based feature selection and wrapper methods (Murthy and Koolagudi, 2018b). Out of four different classifiers used, ANNs are found to be more suitable for segmenting vocal and non-vocal regions. The concept of point moving window has been used to avoid the problems of intermediate misclassification of vocal and non-vocal frames which further improves the performance of the system.

The task of vocal and non-vocal segmentation is found to be difficult due to the stochastic nature of their properties. However, it can be implemented using thorough analysis on the properties of music signal. One such property is repetition nature of music. Locating the repeating patterns may give a clue for locating the onsets of singing voice. The concept of Q-transformation has been used to estimate the repeating patterns at signal level. From the features extracted from Q-transformation, the system is able to recognize the repeating patterns of a signal up to 75%. Moreover, formant analysis is also giving better discrimination for discriminating the vocal and non-vocal regions.

7.1.2 Singer Identification

In this work, an approach has been proposed to recognize the singer's information using convolutional neural networks. Initially, different feature combinations have been tried to identify singer information using the random forest and artificial neural networks. However, they are unable to learn the properties of singer. Hence, CNNs have been considered as they are capable of extracting the features from the images without forming a

manual intervention. Though they are giving better performance for Indian popular songs, the same architecture has failed to achieve similar performance with *artist20* dataset.

The task of singer identification has been implemented using song clips of different lengths varying from 5 to 60 seconds. The degradation in the performance of singer recognition has been observed when the length of input clip is reduced. However, identifying a singer from smaller snippets is essential to make the MIR system real-time applicable. In such cases, feature-based approaches may not show acceptable performance. It is also observed that effective singer identification is possible with CNNs using spectrogram images of song clips as inputs. CNN configuration which is done for one set of singers may not be useful to the other set due to cultural variations and many other unknown factors such as instrumentals, genres, etc. Due to this, the CNN designed for Indian singers dataset has not shown similar performance with *artist20* dataset.

7.1.3 Music Mood Estimation

Since the classification of moods is quite difficult at single level, a two-level classification approach has been proposed in this work. At the initial stage, different acoustical features have been computed for categorizing the song clips into either *energetic* or *non-energetic* classes. Further, CNNs have been used to identify the actual class label.

In music mood estimation, there is a chance for ambiguities since the perceiving nature differs from person to person. Hence, the support of mean opinion score (MOS) and experts opinion always help in labelling the moods. It is also a difficult to develop a system for more number of classes at one phase. The reason may be the pattern similarity in the signal for certain moods. For instance, the signal characteristics of *happy* and *angry* will be same since their energy levels are similar at signal level. However, the analysis in frequency domain may give cues for better classification. The process of developing a multi-level classification system is able to efficiently classify the moods since the single level classification is suffering with either over-fitting or under-fitting issues. There could be a problem of improper information for some moods due to ambiguities in them. Though the spectrogram has been constructed for analysing the mood characters at frequency level, it is quite difficult to identify discriminative properties unless the complete image is processed in the learning phase. Hence, deep neural networks always do better in such cases as they deal with the whole image instead of features obtained by us. They are capable of self extracting features from the image based on their discriminative

characteristics for different classes. The multi-lingual database is also to be analysed to develop a mood classification system for different cultures.

7.1.4 Music Recommender System

As seen in the literature, music recommender system is developed using collaborative and content-based recommendation approaches. Apart from them, graphs are the data structures that are capable of holding information as well as they are capable of estimating the listener's behaviour. Hence, in this work, a graph based recommendation has been proposed which can be enhanced by the future researchers who have willing to work on recommender systems. The edges of the graph can give information about the movement of listener from song to song. It can be the information about singer, mood, composer, album, etc. Moreover, it is possible to estimate the performance of recommender system without the intervention of listener using graph structures.

7.2 Future Work

The work and ideas presented in this thesis may be further extended and improved as follows

→ The performance of vocal and non-vocal segmentation can be enhanced by using larger database and more test clips. Features based on Q-transform, Wavelets etc., may give better performance due to their capability to retain and enhance the resolution compared to traditional spectrogram techniques. These approaches may also help in singing voice detection.

→ The task of singing voice detection can be effectively addressed based on estimating repeated patterns as majority of vocals start immediately after a recurring pattern.

→ The process of singer identification may be poor when the number of singers is more. Providing multi-level classification approach may improve the efficiency of singer identification. The problem can be extended to identify the similarity in singers through certain features.

→ Characterising a timbre with respect to the specific singer is also an essential step that has to be taken care of while developing the singer identification system.

→ As CNNs are capable of discriminating the moods effectively, high resolution frequency spectrogram having more and better information may be used with CNNs for better performance.

→ Since the task of developing MRS has been considered as a prime objective in this thesis, a method based on graph structures has been proposed. Enhancing the database, and implementation of complete recommender system based on graph structures by using the information like time of listening to a song, whether completely listened to, listened to repeatedly etc., is the immediate task which may recommend more relevant songs.

7.3 Future Directions

→ Lack of benchmark datasets in eastern countries, especially the music of Indian subcontinent is a major concern to the music research community. The song categories of eastern world contribute to a major portion of the digital audio domain. The provision of standard datasets to cover different aspects of MIR highly motivates the researchers to work on this area.

→ The features that are computed for various speech and audio processing techniques have been directly used for a majority of MIR tasks without thorough analysis. Some standard correlation analysis may help in deciding the feature set. In this process, it is also possible to reduce the dimensionality which further minimizes the computational complexity issues. The task of identifying requirement related features for different issues of MIR system is still a major problem which needs an immediate concern.

→ At present, the task of vocal and non-vocal segmentation has been considered just as a subtask of a singer identification, for which, extracting a small portion would be sufficient. However, there are many other applications for a complete vocal and non-vocal segmentation task. The process of separating source information may simplify the task of locating vocal onset and offset points. This separation is also helpful in developing an efficient *Karaoke* system without foreground voice. Hence, a special focus is essential on vocal and non-vocal segmentation.

→ Majority of singer identification systems that are accessible at present, consider

audio clips with solo singer, and minimal or monophonic musical accompaniment. This can be the major hindrance in considering them for real-time applications. Hence, there is an immense need to develop singer identification systems that can handle the clips with multiple singers, overlapping singers, and variety of background instruments. Singer tracking in duets is also an important step towards a complete solution. The process of detecting gender of a singer could be one possible solution which simplifies the task of singer identification.

→ The taxonomy of genres is not well-defined in the music industry. Many a times it is found that the same track appears in the category of more than one genre. The taxonomy creation is surely a motivation for the researchers to design a proper genre classification system.

→ The task of raga identification can be improved only with the support of tonic identification. Though significant amount of work has already been reported on tonic (frequency) estimation, the approaches reported are not adequately developed for live concerts. As tonic frequency is an essential component in estimating the singer of a song, it is essential to develop a complete and effective tonic identification system. Existing tonic identification systems have not considered, all the 72 Melakarta ragas. The task of raga identification and note transcription can help in designing an automated tutor for those who are interested in teaching learning Indian classic music. A system can also be developed to judge the singers in live performances through objective evaluation.

→ The task of instrument identification is highly focused on monophonic clips, and the overlapping issues are less addressed. The approach of independent component analysis may help in estimating the instruments though they are recorded in a polyphonic environment. The timbre of an instrument is also highly helpful for the task of instrument identification. The histogram analysis for different features of various instruments can be another possible solution in order to distinguish them. Multi-pitch detection is helpful to recognize the multiple instruments in a selected clip. More fine-tuned automated systems are essential to detect the instruments in polyphonic environment.

→ The area of speech emotion recognition has been highly addressed, and many ap-

proaches have been proposed. The difficulty in deciding the mood of a song clip is the main reason for not having standard models and benchmark datasets for music mood estimation. An effort has to be made for a general dataset and standard baseline model of mood estimation in the context of songs. Moreover, the existing works have concentrated only on the song portions, or on instrumentals separately. However many a times, in real world scenarios they are available together in songs. It is also true that the vocals carry much information related to moods. This hints towards extensive efforts on mood estimation from a song based on vocals in it.

→ Query By Humming (QBH) is another important task of MIR which has been implemented mostly on MIDI files which may not be suitable for real-time scenarios. Recently, some systems have been proposed for query-by-singing (QBS) which is also useful to extract or search clips based on either lyrics or human voice. The accuracy of Top-1 rank is also not up to the appreciable mark for generalization of the performance. From this background, it is said that there is a huge research scope in QBH and QBS.

→ The process of annotating each portion of the song is the ultimate solution to MIR which gives complete information. The task is certainly dependent on the other tasks of MIR. At present, the works have focused on labelling the songs based on the lyrics, instruments, and solo singer. There are many other important tasks of MIR that are to be concentrated on, to provide complete annotation such as gender, multiple singers, raga, and so on. It is also useful that the portion of a song can be labelled with more than one tags based on the information present.

Inclusion of all the above aspects while developing the tasks of MIR may result in a sophisticated MIR system which can be used for real-time applications.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, volume 16, pages 265–283.

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Agarwal, G., Maheshkar, V., Maheshkar, S., and Gupta, S. (2018). Recognition of emotions of speech and mood of music: A review. In *Proceedings of the International Conference on Wireless Intelligent and Distributed Environment for Communication*, pages 181–197. Springer.

Ahmed, N., Natarajan, T., and Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, 100(1):90–93.

Aljanaki, A., Yang, Y. H., and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *Journal of Public Library of Science (PLOS) ONE*, 12(3):e0173392.

Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, T., and Cremer, M. (2001). Content-based identification of audio material using mpeg-7 low level description. In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR)*.

Alonso, M. A., Richard, G., and David, B. (2004). Tempo and beat estimation of musical signals. In *Proceedings of the 5^{th} International Symposium on Music Information Retrieval (ISMIR)*, pages 158–163.

Anagnostopoulos, C. N., Iliou, T., and Giannoukos, I. (2012). Features and classifiers for

emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, pages 1–23.

áOscar Celma (2010). *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer.

Argstatter, H. (2016). Perception of basic emotions in music: Culture-specific or multicultural? *Psychology of Music*, 44(4):674–690.

Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., Lüke, K.-H., and Schwaiger, R. (2011). Incarmusic: Context-aware music recommendations in a car. In *Proceedings of the International Conference on Electronic Commerce and Web Technologies*, pages 89–100. Springer.

Bang, S. W., Kim, J., and Lee, J. H. (2013). An approach of genetic programming for music emotion classification. *International Journal of Control, Automation and Systems*, 11(6):1290–1299.

Barrett, L. F. and Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, 8(1):10–14.

Barrington, L., Yazdani, M., Turnbull, D., and Lanckriet, G. R. (2008). Combining feature kernels for semantic music retrieval. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)*, pages 614–619.

Bartsch, M. A. and Wakefield, G. H. (2004). Singing voice identification using spectral envelope estimation. *IEEE Transactions on Speech and Audio Processing*, 12(2):100–109.

Becchetti, C. and Ricotti, K. P. (2008). *Speech Recognition: Theory and C++ Implementation*. John Wiley & Sons.

Behroozmand, R. and Almasganj, F. (2007). Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis. *Computers in Biology and Medicine*, 37(4):474–485.

Bello, J. P. (2007). Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR)*, volume 7, pages 239–244.

Benetos, E., Kotti, M., and Kotropoulos, C. (2006). Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–224. IEEE.

Benzi, K., Defferrard, M., Vandergheynst, P., and Bresson, X. (2016). Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*.

Berenzweig, A. L., Ellis, D. P., and Lawrence, S. (2002). Using voice segments to improve artist classification of music. In *Proceedings of the 22nd International Conference on Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society.

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., and Kégl, B. (2006). Aggregate features and adaboost for music classification. *Journal of Machine learning*, 65(3):473–484.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596. ISMIR.

Bischoff, K., Firan, C. S., Paiu, R., Nejdl, W., Laurier, C., and Sordo, M. (2009). Music mood and theme classification-a hybrid approach. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR)*, pages 657–662.

Biswas, R., Srinivasa Murthy, Y. V., Koolagudi, S. G., and Swaroop, G. V. (2018). Objective assessment of pitch accuracy in equal-tempered vocal music using signal processing approaches. In *Proceedings of 6th International Conference on Advanced Computing, Networking and Informatics (ICACNI)*. Springer.

Björkner, E. (2006). *Why so Different?: Aspects of Voice Characteristics in Operation and Musical Theater Singing*. PhD thesis.

Bogdanov, D., Haro, M., Fuhrmann, F., Gómez, E., and Herrera, P. (2010). Content-based music recommendation based on user preference examples. In *The 4th ACM Conference on Recommender Systems. Workshop on Music Recommendation and Discovery (WOMRAD)*, Barcelona, Spain.

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., and Herrera, P. (2013). Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management*, 49(1):13–33.

Boger, Z. and Guterman, H. (1997). Knowledge extraction from artificial neural network models. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation*, volume 4, pages 3030–3035. IEEE.

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

Bruce, L. M., Koger, C. H., and Li, J. (2002). Dimensionality reduction of hyper-spectral data using discrete wavelet transform feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10):2331–2338.

Burke, R. (2007). Hybrid web recommender systems. In *Proceedings of the Adaptive Web*, pages 377–408. Springer.

Cai, W., Li, Q., and Guan, X. (2011). Automatic singer identification based on auditory features. In *Proceedings of the Seventh International Conference on Natural Computation (ICNC)*, volume 3, pages 1624–1628. IEEE.

Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., and Wack, N. (2006). Ismir 2004 audio description contest. *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep.*

Casey, M. and Slaney, M. (2007). Fast recognition of remixed music audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1425–1428. IEEE.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Celma, Ò. and Serra, X. (2008). Foafing the music: Bridging the semantic gap in music recommendation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):250–256.

Cheng, H. T., Yang, Y. H., Lin, Y. C., Liao, I. B., and Chen, H. H. (2008). Automatic chord recognition for music classification and retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1505–1508. IEEE.

Chiliguano, P. and Fazekas, G. (2016). Hybrid music recommender using content-based and social information. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2618–2622. IEEE.

Chou, W. and Gu, L. (2001). Robust singing detection in speech/music discriminator design. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, volume 2, pages 865–868. IEEE.

Chua, B. Y. (2008). *Automatic Extraction of Perceptual Features and Categorization of Music Emotional Expressions from Polyphonic Music Audio Signals*. PhD thesis, Monash University.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20(3):273–297.

Cunningham, S. J., Bainbridge, D., and Downie, J. S. (2012). The impact of mirex on scholarly research (2005-2010). In *In proceedings of the 13th International Society for Music Information Retrieval (ISMIR)*, pages 259–264. ISMIR.

Dannenberg, R. B., Birmingham, W. P., Pardo, B., Hu, N., Meek, C., and Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the musart testbed. *Journal of the Association for Information Science and Technology*, 58(5):687–701.

Dittmar, C., Bastuck, C., and Gruhne, M. (2007). Novel mid-level audio features for music similarity. In *Proceedings of the International Conference on Music Communication Science*, pages 38–41.

Dittmar, C., Lehner, B., and Prätzlich, T. (2015). Cross-version singing voice detection in classical opera recordings. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), Malaga, Spain*, pages 618–624. ISMIR.

Do, Q. H. and Chen, J. F. (2013). A neuro-fuzzy approach in the classification of students' academic performance. *Computational intelligence and neuroscience*, 2013:6.

Downie, J. S., Ehmann, A. F., Bay, M., and Jones, M. C. (2010). The music information retrieval evaluation exchange: Some observations and insights. In *Advances in music information retrieval*, pages 93–115. Springer.

Downie, X., Laurier, C., and Ehmann, M. (2008). The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)*, pages 462–467. ISMIR.

Eghbal-Zadeh, H., Schedl, M., and Widmer, G. (2015). Timbral modeling for music artist recognition using i-vectors. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1286–1290. IEEE.

Ellis, D. P. (2007). Classifying music audio with timbral and chroma features. In *Proceedings of the Eighth International Symposium on Music Information Retrieval (ISMIR)*, volume 7, pages 339–340.

Ellis, D. P. and Poliner, G. E. (2007). Identifying cover songs' with chroma features and dynamic programming beat tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–1429. IEEE.

Erickson, R. (1975). *Sound Structure in Music*. University of California Press.

Farrus, M. and Hernando, J. (2009). Using jitter and shimmer in speaker verification. *IET Signal Processing*, 3(4):247–257.

Fazekas, G., Raimond, Y., Jacobson, K., and Sandler, M. (2010). An overview of semantic web activities in the omras2 project. *Journal of New Music Research*, 39(4):295–311.

Feller, W. (2008). *An Introduction to Probability Theory and its Applications*, volume 2. John Wiley & Sons.

Feng, Y., Zhuang, Y., and Pan, Y. (2003a). Music information retrieval by detecting mood via computational media aesthetics. In *Proceedings IEEE International Conference on Web Intelligence (WIC)*, pages 235–241. IEEE.

Feng, Y., Zhuang, Y., and Pan, Y. (2003b). Popular music retrieval by detecting mood. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–376. ACM.

Fernández-Sotos, A., Fernández-Caballero, A., and Latorre, J. M. (2016). Influence of tempo and rhythmic unit in musical emotion regulation. *Frontiers in Computational Neuroscience*, 10:80.

Fu, L., Mao, X., and Chen, L. (2008). Relative speech emotion recognition based artificial neural network. In *Proceedings of the Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA'08)*, volume 2, pages 140–144. IEEE.

Fu, Z., Lu, G., Ting, K.-M., and Zhang, D. (2010). On feature combination for music classification. In *Proceedings of the Conference on Structural, Syntactic, and Statistical Pattern Recognition*, pages 453–462. Springer.

Fu, Z., Lu, G., Ting, K. M., and Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319.

Fujihara, H., Goto, M., Kitahara, T., and Okuno, H. G. (2010). A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):638–648.

Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G. (2005). Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR)*, pages 329–336.

Fung, C. V. (1993). A review of studies on non-western music preference. *Update: Applications of Research in Music Education*, 12(1):26–32.

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1_suppl):123–147.

Ganapathy, S. (2012). Signal analysis using autoregressive models of amplitude modulation. *Johns Hopkins University*.

Geist, K., Geist, E. A., and Kuznik, K. (2012). The patterns of music. *Young Children*, 2:74–79.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset

for audio events. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Ghatak, K. (2012). Mood based music recommendation method and system. US Patent 8,260,778.

Ghazanfar, M. A. and Prügel-Bennett, A. (2014). Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, 41(7):3261–3275.

Goldberg, D. E. (2006). *Genetic algorithms*. Pearson Education India.

Gómez, E. and Herrera, P. (2004). Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR)*. ISMIR.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2003). Rwc music database: Music genre database and musical instrument sound database.

Grosche, P., Müller, M., and Serrà, J. (2012). Audio content-based music retrieval. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Güçlü, U. and van Gerven, M. (2017). Probing human brain function with artificial neural networks. *Computational Models of Brain and Behavior*, page 413.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato.

Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering*, 15(6):1437–1447.

Hallam, S., Cross, I., and Thaut, M. (2011). *Oxford Handbook of Music Psychology*. Oxford University Press.

Hamel, P., Bengio, Y., and Eck, D. (2012). Building musically-relevant audio features through multiple timescale representations. In *Proceedings of the 13th Annual International Symposium on Music Information Retrieval (ISMIR)*, pages 553–558. ISMIR.

Han, B. j., Ho, S., Dannenberg, R. B., and Hwang, E. (2009). {SMERS}: Music emotion recognition using support vector regression. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR)*, pages 651–656. ISMIR.

Han, B. j., Rho, S., Jun, S., and Hwang, E. (2010). Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460.

Harma, A. and Laine, U. K. (2001). A comparison of warped and conventional linear predictive coding. *IEEE Transactions on Speech and Audio Processing*, 9(5):579–588.

Harte, C. and Sandler, M. (2005). Automatic chord identification using a quantised chromagram. In *Audio Engineering Society Convention 118*. Audio Engineering Society.

Helen, M. and Virtanen, T. (2005). Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)*, pages 1–4. IEEE.

Herrada, O. C. (2009). *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, pages 246–268.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE.

Homburg, H., Mierswa, I., Möller, B., Morik, K., and Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *Proceedings of the 6th Annual International Symposium on Music Information Retrieval (ISMIR)*, pages 528–531.

Horsburgh, B., Craw, S., and Massie, S. (2015). Learning pseudo-tags to augment sparse tagging in hybrid music recommender systems. *Artificial Intelligence*, 219:25–39.

Hsu, C. L. and Jang, J. S. R. (2010). On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319.

Hu, X. and Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR)*, pages 67–72. ISMIR.

Hu, Y., Chen, X., and Yang, D. (2009). Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR)*, pages 123–128. ISMIR.

Hughes, J. (1946). The threshold of audition for short periods of stimulation. *Proceedings of the Royal Society of London, Series B-Biological Sciences*, 133(873):486–490.

Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. *Perception*, 10(1):83–92.

Huron, D. B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. MIT press.

Hyung, Z., Lee, K., and Lee, K. (2014). Music recommendation using text analysis on song requests to radio stations. *Expert Systems with Applications*, 41(5):2608–2618.

Isaacson, E. J. (2002). Music ir for music theory. *The MIR/MDL Evaluation Project White paper Collection*, pages 23–26.

Isinkaye, F., Folajimi, Y., and Ojokoh, B. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273.

Jacob, A. (2016). Speech emotion recognition based on minimal voice quality features. In *Proceedings of the International Conference on Communication and Signal Processing (ICCSP)*, pages 0886–0890. IEEE.

Jain, A. K., Mao, J., and Mohiuddin, K. (1996). Artificial neural networks: A tutorial. *IEEE Computer*, 29(3):31–44.

Jawaheer, G., Szomszor, M., and Kostkova, P. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st inter-*

national workshop on information heterogeneity and fusion in recommender systems, pages 47–51. ACM.

Jensen, J. H. (2010). *Feature Extraction for Music Information Retrieval.* Multimedia Information and Signal Processing, Aalborg University.

Jensen, J. H., Christensen, M. G., Ellis, D. P., and Jensen, S. H. (2009). Quantitative analysis of a common audio similarity measure. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):693–703.

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., and Cai, L. H. (2002). Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02)*, volume 1, pages 113–116. IEEE.

Juslin, P. N. and Sloboda, J. A. (2001). *Music and Emotion: Theory and Research.* Oxford University Press.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kassler, M. (1966). Toward musical information retrieval. *Perspectives of New Music*, 4(2):59–67.

Kim, H. G. and Sikora, T. (2004). Audio spectrum projection based on several basis decomposition algorithms applied to general sound recognition and audio segmentation. In *Proceedings of the 12th European Signal Processing Conference*, pages 1047–1050. IEEE.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the 11^{th} International Symposium on Music Information Retrieval*, pages 255–266. ISMIR.

Kim, Y. E. and Whitman, B. (2002). Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR)*, pages 164–169. ISMIR.

Kim, Y. E., Williamson, D. S., and Pilli, S. (2006). Towards quantifying the "album effect" in artist identification. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR)*, pages 393–394.

Kinnear, K. E. (1994). *Advances in genetic programming*, volume 1. MIT press.

Kitahara, T. (2010). Mid-level representations of musical audio signals for music information retrieval. In *Advances in Music Information Retrieval*, pages 65–91. Springer.

Klapuri, A. P. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816.

Knees, P., Pohle, T., Schedl, M., and Widmer, G. (2007). A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of the 30th annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 447–454. ACM.

Knox, D., Beveridge, S., Mitchell, L. A., and MacDonald, R. A. (2011). Acoustic analysis and mood classification of pain-relieving music. *The Journal of the Acoustical Society of America*, 130(3):1673–1682.

Koolagudi, S. G., Murthy, Y. S., and Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. *International Journal of Speech Technology*, 21(1):167–183.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Korhonen, M. D., Clausi, D., Jernigan, M., et al. (2005). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(3):588–599.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Krumhansl, C. L. (2001). *Cognitive Foundations of musical pitch*. Oxford University Press.

Kumar, K., Kim, C., and Stern, R. M. (2011). Delta-spectral cepstral coefficients for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4787. IEEE.

Kuusi, T. (2009). Tune recognition from melody, rhythm and harmony. In *Proceedings of the 7^{th} Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM).*

Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114.

Langlois, T. and Marques, G. (2009). A music classification method based on timbral features. In *ISMIR*, pages 81–86.

LeCun, Y. (2015). *LeNet-5, convolutional neural networks.*

Lee, C. H., Shih, J. L., Yu, K. M., and Lin, H. S. (2009). Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia*, 11(4):670–682.

Lehner, B. and Widmer, G. (2015). Monaural blind source separation in the context of vocal detection. In *Proceedings of the 16th International Society for Music Information Retrieval (ISMIR)*, pages 309–315. ISMIR.

Lehner, B., Widmer, G., and Bock, S. (2015). A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks. In *Proceedings of the 23rd IEEE European Signal Processing Conference (EUSIPCO)*, pages 21–25. IEEE.

Lesaffre, M. (2006). *Music Information Retrieval: Conceptual Framework, Annotation and User Behaviour.* PhD thesis, Ghent University.

Levy, M. and Bosteels, K. (2010). Music recommendation and the long tail. In *1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain.* Citeseer.

Li, T. and Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR)*, volume 3, pages 239–240. ISMIR.

Li, T. and Ogihara, M. (2005). Music genre classification with taxonomy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, volume 5, pages v–197. IEEE.

Li, T. and Ogihara, M. (2006). Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574.

Li, T., Ogihara, M., and Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–289. ACM.

Li, Y. and Wang, D. (2006). Singing voice separation from monaural recordings. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR)*, pages 176–179. ISMIR.

Lidy, T. and Rauber, A. (2005). Mirex 2005: Combined fluctuation features for music genre classification. In *Proceedings of the 6th Annual International Symposium on Music Information Retrieval (ISMIR)*. ISMIR.

Liu, C.-C. and Huang, C.-S. (2002). A singer identification technique for content-based classification of mp3 music objects. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 438–445. ACM.

Liu, H. and Motoda, H. (2007). *Computational Methods of Feature Selection*. CRC Press.

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., and Wang, J. (2018). Applications of deep learning to mri images: a survey. *Big Data Mining and Analytics*, 1(1):1–18.

Logan, B. et al. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.

Loui, P. (2015). A dual-stream neuroanatomy of singing. *Music Perception: An Interdisciplinary Journal*, 32(3):232–241.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems*, 74:12–32.

Lu, L., Liu, D., and Zhang, H.-J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18.

Macy, L. W. (2001). *Grove Music Online*. Macmillan Publishers.

Maddage, N. C., Xu, C., and Wang, Y. (2004). Singer identification based on vocal and instrumental models. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 375–378. IEEE.

Mahmood, T. and Ricci, F. (2009). Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 73–82. ACM.

Mandel, M. and Ellis, D. (2005). Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR)*, pages 594–599.

Marchand, U. and Peeters, G. (2016). The extended ballroom dataset. In *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*.

Markel, J. D. and Gray, A. J. (2013). *Linear Prediction of Speech*, volume 12. Springer Science & Business Media.

Mauch, M., Fujihara, H., Yoshii, K., and Goto, M. (2011). Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 233–238.

McKay, C., McEnnis, D., and Fujinaga, I. (2006). A large publicly accassible prototype audio database for music research. In *Proceedings of the 7th Annual International Symposium on Music Information Retrieval (ISMIR)*, pages 160–163.

McVicar, M., Ellis, D. P., and Goto, M. (2014). Leveraging repetition for improved automatic lyric transcription in popular music. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3117–3121. IEEE.

Mellody, M. and Wakefield, G. (2000). Signal analysis of the singing voice: Low-order representations of singer identity. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 98–101.

Mendes, A. P., Rothman, H. B., Sapienza, C., and Brown Jr, W. S. (2003). Effects of vocal training on the acoustic parameters of the singing voice. *Journal of voice*, 17(4):529–543.

Mesaros, A. and Astola, J. (2005a). Inter-dependence of spectral measures for the singing voice. In *Proceedings of the International Symposium on Signals, Circuits and Systems (ISSCS).*, pages 307–310. IEEE.

Mesaros, A. and Astola, J. (2005b). The mel-frequency cepstral coefficients in the context of singer identification. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR)*, pages 610–613. ISMIR.

Mesaros, A., Virtanen, T., and Klapuri, A. (2007). Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR)*, pages 375–378.

Mesquita, B., Frijda, N. H., and Scherer, K. R. (1997). Culture and emotion. *Handbook of Cross-cultural Psychology*, 2:255–297.

Mesquita, B. and Walker, R. (2003). Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour Research and Therapy*, 41(7):777–793.

Mion, L. and Poli, G. D. (2008). Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466.

Mörchen, F., Ultsch, A., Thies, M., and Löhken, I. (2006). Modeling timbre distance with temporal statistics from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):81–90.

Müller, M., Mattes, H., and Kurth, F. (2006). An efficient multi-scale approach to audio synchronization. In *In proceedings of the seventh International Society for Music Information Retrieval (ISMIR)*, pages 192–197. ISMIR.

Murthy, Y. and Koolagudi, S. G. (2018a). Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. *ACM Computing Surveys (CSUR)*, 51(3):45.

Murthy, Y. S. and Koolagudi, S. G. (2015). Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In *Proceedings of the 28th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1271–1276. IEEE.

Murthy, Y. S. and Koolagudi, S. G. (2018b). Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs). *Expert Systems with Applications*, 106:77–91.

Ness, S. R., Theocharis, A., Tzanetakis, G., and Martins, L. G. (2009). Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 705–708. ACM.

Norris, S. (2009). Tempo, auftakt, levels of actions, and practice: Rhythm in ordinary interactions. *Journal of Applied Linguistics*, 6(3).

Nwe, T. L. and Li, H. (2007). Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):519–530.

Orio, N., Rizo, D., Miotto, R., Schedl, M., Montecchio, N., and Lartillot, O. (2011). Musiclef: a benchmark activity in multi-modal music information retrieval. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 603–608. ISMIR.

Oudre, L., Grenier, Y., and Févotte, C. (2011). Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2222–2233.

Pachet, F. and Aucouturier, J. J. (2004). Improving timbre similarity: How high is the sky. *Journal of Negative Results in Speech and Audio Sciences*, 1(1):1–13.

Pachet, F. and Cazaly, D. (2000). A taxonomy of musical genres. In *Proceedings of Content-based Multimedia Information Access Conference (RIOA)*, pages 1238–1245.

Patil, H. A., Radadia, P. G., and Basu, T. K. (2012). Combining evidences from mel cepstral features and cepstral mean subtracted features for singer identification. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 145–148. IEEE.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report.

Phillips, P., Zigan, K., Silva, M. M. S., and Schegg, R. (2015). The interactive effects of online reviews on the determinants of swiss hotel performance: A neural network analysis. *Tourism Management*, 50:130–141.

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., and Widmer, G. (2009). On rhythm and general music similarity. In *ISMIR*, pages 525–530.

Popper, A. N. and Fay, R. R. (2014). *Perspectives on Auditory Research*, volume 50. Springer.

Quinlan, J. R. (2014). *C4. 5: Programs for Machine Learning*. Elsevier.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rafii, Z. and Pardo, B. (2013). Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):73–84.

Ramona, M., Richard, G., and David, B. (2008). Vocal detection in music with support vector machines. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1885–1888. IEEE.

Ratanpara, T. and Patel, N. (2015). Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):16.

Reed, J. and Lee, C. H. (2009). On the importance of modeling temporal information in music tag annotation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1873–1876. IEEE.

Regnier, L. and Peeters, G. (2009). Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1685–1688. IEEE.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer.

Robnik-Šikonja, M. (2004). Improving random forests. In *European conference on machine learning*, pages 359–370. Springer.

Rolland, J. B. (2014). Chord detection using chromagram optimized by extracting additional features. pages 27–31. ISMIR.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Ryo, M. and Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, 8(11).

Saari, P. and Eerola, T. (2014). Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2548–2560.

Salamon, J., Serra, J., and Gómez, E. (2013). Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58.

Samsekai Manjabhat, S., Koolagudi, S. G., Rao, K., and Ramteke, P. B. (2017). Raga and tonic identification in carnatic music. *Journal of New Music Research*, 46(3):229–245.

Sarala, P. and Murthy, H. A. (2013). Inter and intra item segmentation of continuous audio recordings of carnatic music for archival. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 487–492. ISMIR.

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

Scaringella, N., Zoia, G., and Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine*, 23(2):133–141.

Schedl, M., Knees, P., and Widmer, G. (2005). Interactive poster: Using comirva for visualizing similarities between music artists.

Schedl, M., Widmer, G., Knees, P., and Pohle, T. (2011). A music information system automatically generated via web content mining techniques. *Information Processing & Management*, 47(3):426–439.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256.

Schmidt, E. M. and Kim, Y. E. (2010). Prediction of time-varying musical mood distributions using kalman filtering. In *Proceedings of the Ninth International Conference on Machine Learning and Applications (ICMLA)*, pages 655–660. IEEE.

Schörkhuber, C. and Klapuri, A. (2010). Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64.

Serra, J., Gómez, E., and Herrera, P. (2010). Audio cover song identification and similarity: Background, approaches, evaluation, and beyond. In *In proceedings of the Advances in Music Information Retrieval*, pages 307–332. Springer.

Seyerlehner, K., Widmer, G., and Pohle, T. (2010). Fusing block-level features for music similarity estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 225–232.

Shardanand, U. (1994). *Social Information Filtering for Music Recommendation.* PhD thesis, Massachusetts Institute of Technology.

Sharma, R., Murthy, Y. S., and Koolagudi, S. G. (2016). Audio songs classification based on music patterns. In *Proceedings of the Second International Conference on Computer and Communication Technologies*, pages 157–166. Springer.

Shen, J., Shepherd, J., Cui, B., and Tan, K. L. (2009). A novel framework for efficient automated singer identification in large music databases. *ACM Transactions on Information Systems (TOIS)*, 27(3):18.

Shen, J., Shepherd, J., and Ngu, A. H. (2006). Towards effective content-based music retrieval with multiple acoustic feature combination. *IEEE Transactions on Multimedia*, 8(6):1179–1189.

Shenoy, A., Wu, Y., and Wang, Y. (2005). Singing voice detection for karaoke application. In *Proceedings of the Visual Communications and Image Processing*, pages 596028–596028. International Society for Optics and Photonics.

Silla Jr, C. N., Koerich, A. L., and Kaestner, C. A. (2008). The latin music database. In *Proceedings of the 9th Annual International Symposium on Music Information Retrieval (ISMIR)*, pages 451–456.

Simpson, A. J., Roma, G., and Plumbley, M. D. (2015). Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *Proceedings of the 11th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 429–436. Springer, Springer.

Song, Y., Dixon, S., and Pearce, M. (2012). A survey of music recommendation systems and future perspectives. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*. Queen Mary University of London.

Strube, H. W. (1980). Linear prediction on a warped frequency scale. *The Journal of Acoustical Society of America*, 68(4):1071–1076.

Sturm, B. L. (2013). The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *Clinical Orthopaedics Related Research*, 1306.1461.

Sturm, B. L. (2014). A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66. Springer.

Su, J., Yeh, H., Philip, S. Y., and Tseng, V. S. (2010). Music recommendation using content and context information mining. *IEEE Intelligent Systems*, 25(1).

Su, L. and Yang, Y. H. (2013). Sparse modeling for artist identification: Exploiting phase information and vocal separation. In *ISMIR*, pages 349–354.

Sun, C. T. and Jang, J. S. (1993). A neuro-fuzzy classifier and its applications. In *Fuzzy Systems, 1993., Second IEEE International Conference on*, pages 94–98. IEEE.

Sundberg, J. (1977). The acoustics of the singing voice. *Scientific American*, 236(3):82–91.

Sundberg, J. and Rossing, T. D. (1990). The science of singing voice. *The Journal of Acoustical Society of America*, 87(1):462–463.

Tellegen, A., Watson, D., and Clark, L. A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science*, 10(4):297–303.

Thayer, R. E. (1989). *The Bio-psychology of Mood and Arousal*. Oxford University Press.

Thomas, M., Murthy, Y. S., and Koolagudi, S. G. (2016). Detection of largest possible repeated patterns in indian audio songs using spectral features. In *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5. IEEE.

Tingle, D., Kim, Y. E., and Turnbull, D. (2010). Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the 11th International Symposium on Multimedia information retrieval (ISMIR)*, pages 55–62. ISMIR.

Tolonen, T. and Karjalainen, M. (2000). A computationally efficient multi-pitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6):708–716.

Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. P. (2008). Multi-label classification of music into emotions. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)*, volume 8, pages 325–330.

Tsai, W. H. and Lee, H. C. (2011). Performance evaluation of speaker-identification systems for singing voice data. *International Journal of Computational Linguistics & Chinese Language Processing*, 16(1-2).

Tsai, W. H., Liao, S. J., and Lai, C. (2008). Automatic identification of simultaneous singers in duet recordings. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)*, pages 115–120. ISMIR.

Tsai, W. H. and Wang, H. M. (2004). Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, volume 4, pages iv–221. IEEE.

Tsai, W. H. and Wang, H. M. (2006). Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):330–341.

Tsunoo, E., Tzanetakis, G., Ono, N., and Sagayama, S. (2009). Audio genre classification using percussive pattern clustering combined with timbral features. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 382–385. IEEE.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2007). Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 439–446. ACM.

Turnbull, D. R., Barrington, L., Lanckriet, G., and Yazdani, M. (2009). Combining audio content and social context for semantic music discovery. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387–394. ACM.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302.

Tzanetakis, G., Jones, R., and McNally, K. (2007). Stereo panning features for classifying recording production style. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR)*, pages 441–444. Citeseer.

Tzanetakis, G., Martins, L. G., McNally, K., and Jones, R. (2010). Stereo panning information for music information retrieval tasks. *Journal of the Audio Engineering Society*, 58(5):409–417.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Vembu, S. and Baumann, S. (2005). Separation of vocals from polyphonic audio recordings. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 337–344. ISMIR.

Vieira, A. and Ribeiro, B. (2018). Image processing. In *Introduction to Deep Learning Business Applications for Developers*, pages 77–109. Springer.

Vincent, E. (2006). Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98.

Wang, A. et al. (2003a). An industrial strength audio search algorithm. In *In proceedings of the fourth International Symposium on Music Information Retrieval (ISMIR)*, volume 4, pages 7–13. Washington, DC.

Wang, C. K., Lyu, R. Y., and Chiang, Y. C. (2003b). An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. In *Proceedings of the Eighth European Conference on Speech Communication and Technology*. Speech communication and Technology.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. (2016). CNN-RNN: a unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294.

Wang, X. and Wang, Y. (2014). Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the* 22^{nd} *ACM International Conference on Multimedia*, pages 627–636. ACM.

Wang, Y., Kan, M. Y., Nwe, T. L., Shenoy, A., and Yin, J. (2004). Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 212–219. ACM.

Wei, J., Liu, C. H., Zhu, Z., Cain, L. R., and Velten, V. J. (2019). Vehicle engine classification using normalized tone-pitch indexing and neural computing on short remote vibration sensing data. *Expert Systems with Applications*, 115:276–286.

Weihs, C., Ligges, U., Mörchen, F., and Müllensiefen, D. (2007). Classification in music research. *Advances in Data Analysis and Classification*, 1(3):255–291.

Weikart, P. S. (2003). Value for learning and living. *Child Care Information Exchange*, 24(153):86–88.

Whitman, B., Flake, G., and Lawrence, S. (2001). Artist detection in music with minnowmatch. In *Proceedings of the IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, pages 559–568. IEEE.

Wolter, K., Bastuck, C., and Gärtner, D. (2008). Adaptive user modeling for content-based music retrieval. In *Proceedings of the 6th Workshop on Adaptive Multimedia Retrieval: Identifying, Summarizing, and Recommending Image and Music*, pages 40–52. Springer.

Wong, E. and Sridharan, S. (2001). Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pages 95–98. IEEE.

Yang, J., Liu, J., and Zhang, W. (2010). A fast query by humming system based on notes. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2898–2901.

Yang, X., Dong, Y., and Li, J. (2018). Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4):365–389.

Yang, Y. H. and Chen, H. H. (2011). *Music Emotion Recognition*. CRC Press.

Yang, Y. H. and Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40.

Yang, Y. H., Lin, Y.-C., Su, Y. F., and Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457.

Zentner, A. (2003). Measuring the effect of online music piracy on music sales. Technical report, University of Chicago working paper.

Zenz, V. (2007). Automatic chord detection in polyphonic audio data. Master's thesis, Vienna University of Technology.

Zhang, T. (2003). Automatic singer identification. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, volume 1, pages I–33. IEEE.

Zhang, T. and Packard, H. (2003). System and method for automatic singer identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 756–756. ICME.

List of Publications

Journal Publications

1. Murthy, Y. S., & Koolagudi, S. G. (2018). Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS). Expert Systems with Applications, 106, 77-91.

2. Murthy, Y. V., & Koolagudi, S. G. (2018). Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review. ACM Computing Surveys (CSUR), 51(3), 45.

3. Murthy, Y. V., Koolagudi, S. G., & Jeshventh, T.K.R. (2018). Singer Identification using Convolutional Neural Networks (CNNs), Engineering Applications and Artificial Intelligence, Elsevier. **[Communicated]**

4. Murthy, Y. V., & Koolagudi, S. G. (2018). A Two-level Classification for Music Mood Estimation using Feature-based approaches and CNNs, International Journal of Speech Technology (IJST), Springer. **[Communicated]**

Conferences

1. Murthy, Y. V., Jeshventh, T. K. R., Zoeb, M., Saumyadip, M., & Shashidhar, G. K. (2018, August). Singer Identification from Smaller Snippets of Audio Clips Using Acoustic Features and DNNs. In 2018 Eleventh International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

2. Roshni Biswas, Y. V. Srinivasa Murthy, Shashidhar G. Koolagudi, and Vishnu Swaroop G., (2018). Objective Assessment of Pitch Accuracy in Equal-Tempered Vocal Music using Signal Processing Approaches, Proceedings of the Sixth International Conference on Advanced Computing, Networking, and Informatics (ICACNI), Springer.(In Press)

3. Srinivasa Murthy, Y.V., Koolagudi, Shashidhar., & Swaroop, Vishnu G., (2017). Vocal and Non-vocal Segmentation based on the Analysis of Formant Structure, In proceedings of the 9th International Conference on Advances in Pattern Recognition (ICAPR), Indian Statistical Institute (ISI), Bangalore, IEEE.

4. Thomas, M., Jothish, M., Thomas, N., Koolagudi, S. G., & Murthy, Y. S. (2016, November). Detection of similarity in music files using signal level analysis. In Region 10 Conference (TENCON), 2016 IEEE (pp. 1650-1654). IEEE.

5. Thomas, M., Murthy, Y. S., & Koolagudi, S. G. (2016, May). Detection of largest possible repeated patterns in Indian audio songs using spectral features. In Electrical and Computer Engineering (CCECE), 2016 IEEE Canadian Conference on (pp. 1-5). IEEE.

6. Sharma, R., Murthy, Y. S., & Koolagudi, S. G. (2016). Audio songs classification based on music patterns. In Proceedings of the Second International Conference on Computer and Communication Technologies (pp. 157-166). Springer, New Delhi.

7. Murthy, Y. S., & Koolagudi, S. G. (2015, May). Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference on (pp. 1271-1276). IEEE.

Brief Bio-Data

Y V Srinivasa Murthy

Research Scholar

Department of Computer Science and Engineering

National Institute of Technology Karnataka, Surathkal

P.O. Srinivasnagar

Mangalore - 575025

Phone: +91 9848544449

Email: urvishnu@gmail.com

Permanent Address

Y V Srinivasa Murthy

S/o Srinivasa Rao Yarlagadda

D.No: 1-77, Vulavalapudi (Village), Nagavaram (Post)

Gudlavalleru (Mandal), Krishna Dt. - 521 331

Andhra Pradesh, INDIA

Qualification

M. Tech. in Computer Science and Engineering, GITAM Institute of Technology (GIT), GITAM University, Andhra Pradesh, 2010.

B. E. in Computer Science and Engineering, Gudlavalleru Engineering College (GEC), Gudlavalleru, Andhra Pradesh, 2006.

Diploma in Computer Engineering (DCME), Vemulapalli Kodanda Ramayya (V.K.R) & Vemulapalli Naga Bhushanam (V.N.B) Polytechnic, Gudivada, Andhra Pradesh, 2003.

Work Experience

At present, I am Working as an Assistant Lecturer at the CSE Department National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, India.

Worked as an Assistant Professor in the CSE Department of various prestigious institutes like GITAM University, Anil Neerukonda Institute of Technology and Sciences (ANITS), Gudlavalleru Engineering College (GEC), and Sasi Institute of Engineering and Technology (SIET) since 2006.