Generating a digital signature for singers to identify their songs

H.M.S.R. Heenkenda

2015/CS/050

This dissertation is submitted to the University of Colombo School of Computing
In partial fulfillment of the requirements for the
Degree of Bachelor of Science Honours in Computer Science

University of Colombo School of Computing
35, Reid Avenue, Colombo 07,
Sri Lanka
July, 2020

Declaration

I, H.M.S.R. Heenkenda (2015/CS/050) hereby certify that this dissertation entitled "Generating a digital signature for singers to identify their songs" is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

....

Date Signature of the Student

I, Dr. D.D. Karunarathne, certify that I supervised this dissertation entitled "Generating a digital signature for singers to identify their songs" conducted by H.M.S.R. Heenkenda in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Computer Science.

....

Date Signature of the Supervisor

Abstract

A counterfeit is an imitation of the voice of a popular artist, done with the intention of selling or passing it on as a genuine. This imitating of songs from the original artists is being done very smart and smooth, so it becomes impossible to detect it as real or fake. These wrongdoers make an income by selling songs which are imitated disguised as originals. This study proposes a solution for this problem by providing digital signatures for singers that are generated using songs sung by the artists. The songs contain vocal signals surrounded with instrumental music. In order to generate signatures for the voice of the singer, the vocals have to be isolated. This study proposes an isolation technique, which is proved against a prevailing technique. The signature is generated by using the features extracted after voice isolation. The signature of the singer is originated as a Gaussian Mixture Model. The project had been implemented using open source software. The evaluation had been performed through quantitative and qualitative approaches. The outcome of this research had been successful in generating digital signatures for singers. The singers had been identified accurately even for those who possess similar voices.

Preface

The work presented in this study has utilized libraries in python for the implementation. The equations and algorithms found within is the work of the author unless mentioned otherwise by the author. Apart from these the specific code segments mentioned under Chapter 4, the body of work mentioned herein is the work of the author of this document. The python codes are based on the work found in the librosa 0.7.2 documentation. The extended code segments can be found in the Appendices of this document. The evaluation was conducted by the author.

Acknowledgement

I would like to express my sincere gratitude to my Supervisor Dr. D.D. Karunarathne and Co-supervisor Dr. S.M.D.K. Arunatilake for the continuous support on this project, for their patience, motivation and guidance throughout this project. My gratitude is expressed then to my parents for supporting me emotionally and financially as well as for being my strength. I would like to extend my sincere gratitude to my friends from University of Colombo School of Computing (UCSC) for their cooperation and enthusiasm.

Contents

List of Figures

List of Tables

Acronyms

MIR	Music Information Retrieval
CBIV	Content Based Integrity Verification
MFCC	Mel-Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
REPET	Repeating Pattern Extraction Technique
SVM	Support Vector Machine
PLP	Perceptual Linear Prediction
RMS	Root Mean Square
STFT	Short Time Fourier Transform
STN	Sines+transients+noise
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EM	Expectation Maximization
IDE	Integrated Development Environment
HPSS	Harmonic-percussive source separation
OpenGL	Open Graphics Library
MATLAB	matrix laboratory
API	Application Programming Interface
LIBSVM	Library for Support Vector Machines
UK	United Kingdom
ICT	Information and Communication Technology
OCR	Oxford, Cambridge and RSA
DUET	Degenerate Unmixing Estimation Technique
FFT	Fast Fourier Transform

Chapter 1

Introduction

1.1 Background to the research

In contrast to real property, Intellectual Property is not tangible because it originates in the human mind involving human intellect. Therefore, it is generally described as “the fruit of human intellect”. Intellectual property rights in a country are the laws that make the provisions for the protection of the results in this human endeavor.

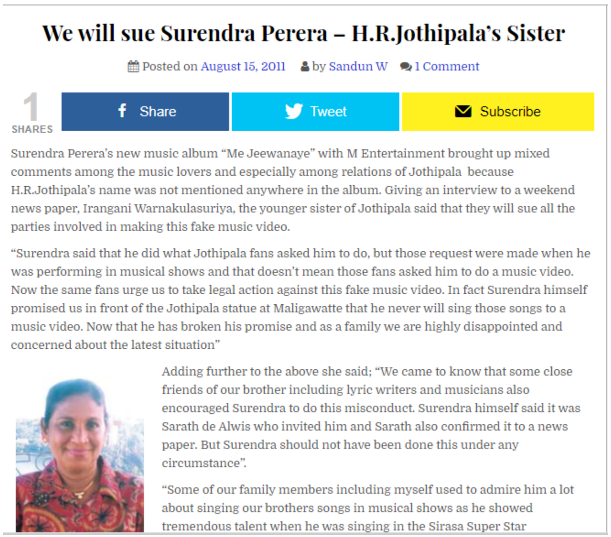
Intellectual Property of Act No. 36 of 2003 is the framework within which the Intellectual Property rights are currently protected in the Island of Sri Lanka. Intellectual Property is classified under several Headings in the Act, which is Copyright and related rights, Industrial Designs, Patents, and Trade Marks and many more to be given the protection of law (Dissanayake, 2016).

Copyright includes all creative and artistic works such as books, movies, music, paintings, sound recordings, and computer programs, etc. which need no registration under the act but protected automatically by operation of law. Intellectual property right gives the copyright owner the exclusive right to reproduce the work, prepare derivative works, perform the work and display the work (Murray, 2014).

A famous singer generally has a right to protect the use of their voice in order to sell products. These right attaches when the singer is widely known and their distinctive voice is deliberately imitated in order to sell a product (O’Neil, 2004). When someone uses the voice of a singer and sings one of that singer’s songs and earns money by that, that will be misappropriation under the laws in Sri Lanka.

This is misappropriation because the sellers of the product have appropriated what is not theirs and have committed a tort. This comes under infringement of copyrights and also a violation of the right of publicity.

The following Figure 1.1 depicts an instance where imitating the voice of a popular singer was considered a tort.



We will sue Surendra Perera – H.R.Jothipala's Sister

Posted on August 15, 2011 by Sandun W 1 Comment

Surendra Perera's new music album "Me Jeewanaye" with M Entertainment brought up mixed comments among the music lovers and especially among relations of Jothipala because H.R.Jothipala's name was not mentioned anywhere in the album. Giving an interview to a weekend news paper, Irangani Warnakulasuriya, the younger sister of Jothipala said that they will sue all the parties involved in making this fake music video.

"Surendra said that he did what Jothipala fans asked him to do, but those request were made when he was performing in musical shows and that doesn't mean those fans asked him to do a music video. Now the same fans urge us to take legal action against this fake music video. In fact Surendra himself promised us in front of the Jothipala statue at Maligawatte that he never will sing those songs to a music video. Now that he has broken his promise and as a family we are highly disappointed and concerned about the latest situation"

Adding further to the above she said; "We came to know that some close friends of our brother including lyric writers and musicians also encouraged Surendra to do this misconduct. Surendra himself said it was Sarath de Alwis who invited him and Sarath also confirmed it to a news paper. But Surendra should not have been done this under any circumstance".

"Some of our family members including myself used to admire him a lot about singing our brothers songs in musical shows as he showed tremendous talent when he was singing in the Sirasa Super Star

Figure 1.1: Imitating vocals being considered as a tort.

At present, audio fingerprinting and voice biometrics-based technologies are being used in various applications like human voice authentication systems. An instance of this would be Nuance– simple and secure voice biometric authentication which enables easy and secure biometric authentication using just a person's voice. There are also a number of audio signal processing based applications used in the music industry, for applications like content-based audio identifications (CBIV), Content-based integrity verification (CBID) and watermarking support (Froitzheim, 2017). To explain audio fingerprinting briefly, it is the process of encoding a (potentially) unlabeled piece of audio in any format into a so-called fingerprint. It is usually required for this process to work in a compact, discriminative, robust and efficient way such that resulting fingerprint can be easily stored. The prime objective of multimedia fingerprinting is an efficient mechanism to establish the perceptual equality of two multimedia objects: not by comparing the whole files, but by comparing the associated signatures (Cano et al., 2003) (Haitsma and Kalker, 2002).

As mentioned above, with the capabilities of MIR and signal processing techniques, it is worthwhile to venture into integrating these technologies into generating a digital signature for singers to recognize their songs. Although there are applications to provide audio fingerprints to songs, it was unable to find an application to give singers a unique signature to their voice by giving their songs as an input.

1.2 Research problem and Research Questions

1.2.1 Problem Statement

Counterfeiting involves duplication of both the music product and of its packaging. A counterfeit is an imitation, usually, one that is made with the intention of infrequently passing it off as genuine. Only the copyright holder of a song has the right to make commercial usage from his work. If another party tries to counterfeit the original work and make commercial usage out of it, that would come under the copyright infringement and breaking of intellectual property rights of that artist. Counterfeit product and an original work of an artist(song) will be very hard to distinguish when heard by a normal person. It requires extensive knowledge and practice in music for a person to distinguish between original singing and well-imitated singing. At present, in order to start an investigation regarding commercial usage of an imitation of a singer's work, the singer himself has to recognize and appeal to the court for justice which at the end would not even be fruitful as there is no technological help to distinguish between an original song and an imitated version. So, it would be better if the singers themselves had a digital signature to secure what is theirs. In another aspect, when a song is heard, sometimes it is hard to recognize the singer. If you are given an application which would recognize the singer of that song by just using your mobile phone, that would also be a great achievement. Generation of a digital signature for a singer by using his songs would be the first step in this purpose too.

1.2.2 Research Aim

To explore the possibility of generating a unique digital signature for an artist by using his songs.

1.2.3 Research Questions

1. How can audio signal processing and music information retrieval be used to distinguish between voice extracted from songs?
2. What gains can be made using the proposing method over usual voice recognition methods used?
3. Would this proposing signature meet the other requirements? (can be easily stored, indexed and compared)

1.2.4 Research Objectives

1. Use audio signal processing and music information retrieval to generate digital signatures for singers by using their songs.
2. Assisting relevant parties in order to distinguish between an original song and a counterfeit work.
3. Preventing listeners misjudging a singer as someone else after listening to a song.
4. Explore methods to cluster various features in voice which have not been paid attention to and come up with a unique set of features.

1.3 Delimitation of Scope

This project will generate signature per each artist in order to secure their vocal identity in their songs. The following details the delimitations of this project.

- This project will come up with a model (signature) such that there exists at least one unique feature to distinguish between voices of two singers.

- The signature will have the ability to be easily stored, indexed and compared with other signatures.

- The signature will be generated by using the songs sung by the relevant artist.

1.4 Methodology

The philosophical foundation of this research is based on the view, that existing procedures can be verified through experiments, observations and mathematical logic. Therefore this research is an experimental research, which is designed to collect and interpret data, based on experiments and observations. Finally, the validation will be conducted through experimentation and the findings of the research will be observable and quantifiable.

1.5 Contribution

This dissertation contributes to the area of Audio signal processing and music information retrieval. Specifically, it introduces a solution for the artists who had been affected by the dissenters who imitate their voice in song production and in performance. The unique features of the voice of a singer can be used to distinguish between this real and fake singing. Those unique features are extracted and accumulated into a unique digital signature which will not be similar to two different artists. Those signatures can be compared with each other and be examined if a song is classified as a certain singer's accurately. The signature generation is done by following a process of voice isolation and feature extraction. Voice isolation has been done using a combined strategy of a pre-defined voice isolation approach.

1.6 Definitions

Throughout this document, the term artist is used to refer to singers of different songs who will be the ultimate goal of the resulting system of this research. The term voice isolation is used to refer to the removal of musical components from the audio tracks. The term signature stands for the final unique model generated for a specific singer after extracting unique features from his vocals.

1.7 Outline of the Dissertation

The remainder of the dissertation is structured as explained in this section.

Literature Review included in Chapter 2 was conducted with the intention of identifying the gap in the available body of work. Also, the review helped to identify the strengths and weaknesses of the proposed method. The Design of the research and the rationale behind the design choices have been detailed in Chapter 3. The design of the proposed voice isolation method, signature generation design and the evaluation design have been discussed in this chapter.

Chapter 4 contains a discussion on the Implementation of the proposed solution. An introduction to the technological components and the methodology to realize the design given in Chapter 3 is discussed in Chapter 4. This is followed by the Results and Evaluation in Chapter 5 which contains results from the quantitative evaluation and the qualitative evaluation of the project. A discussion on the results is also contained here.

Finally, the Conclusions of this research are discussed in Chapter 6. The conclusions on the research problem and questions are contained in this chapter along with future work that can stem from this research.

1.8 Chapter Summary

This concludes the introductory chapter on the research which has laid the foundation for the dissertation. This chapter has given an introduction to the domain in which it will be applied. Background of the research with the details of foundation framework has been included in this chapter as well. The research problem to be addressed along with the aim of the research and the research questions have been introduced in this chapter. The methodology of the research has been given next followed by the contribution which is a brief summary on the major findings of the research. Given this introduction, the dissertation can proceed with detailed insight into the research.

Chapter 2

Literature Review

This review is conducted mainly to identify solutions for the research questions posed in Chapter 1. How the past researchers had tried to solve Voice Isolation in audios and speaker identification tasks separately are being discussed thoroughly in this chapter. Also, a comparison had also been made in order to identify the strengths and weaknesses of each method.

Generating digital signatures for singers to identify their songs using songs as input can be considered as a novel research. Even though there had been many approaches in order to do artist classification, no researcher had used songs as their input. Therefore, this research is primarily a novel research for what is known up to today.

2.1 Introduction

Voice, joyfully raised in a song is a complex human achievement, one that is nearly miraculous in nature. Dunn (Dunn, 2013) discusses the phenomenon of voice in his paper, reflecting how voice should be treated as same or more as instrumental contribution in a song. He presents the vantage points, the human singing voice is in several important respects, like all other musical instruments, like some other musical instruments, and like none other music instruments. He further deliberates the characteristics of voice, Vibration and resonance, variety of technique and tone, and finally pitch. The relevance of voice in songs have been evaluated in (Demetriou et al., 2018) addressing the two questions, what components of music are most salient to people's music taste, how do vocals rank relative to other components of music.

The main attribute that distinguishes musical instruments from one another is timbre. Timbre is the quality of sound that differentiates different sounds (Velankar, 2013). Acoustically all kinds of sounds are similar but they possess fundamental differences. It is separated from the expression attributes. Brightness and roughness, can also be helpful to understand the dimensions of timbre. The timbre of a sound depends on its waveform, their frequencies, and their relative intensities. The most common methodologies to extract features related to the timbre of sound are Mel Frequency Spectral coefficients and Formant Analysis (Jensen, 1999), (Bonjyotsna and Bhuyan, 2013).

2.1.1 Voice Isolation

In the singer identification from polyphonic music signals task, the major challenge to face is the negative influences caused by accompaniment sounds. Many researchers have proposed strategies to isolate vocals from accompaniment sounds. The following Figure 2.1 exhibits the flow of voice isolation.

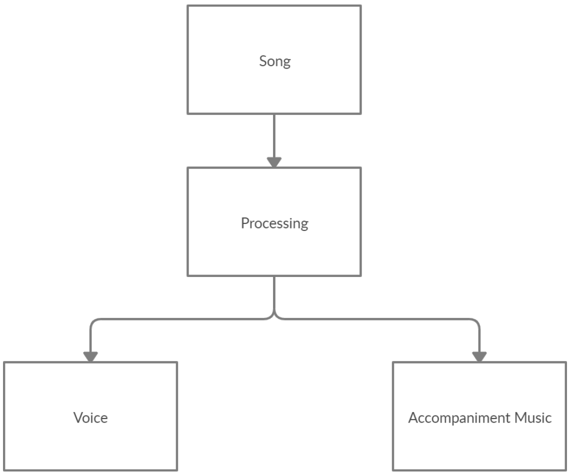


Figure 2.1: Flow of voice Isolation.

In order to isolate the voice of a song, features of the song spectrums should be extracted. There are basically two types of approaches to extract features of sound, MFCC and Formant analysis. The Mel frequency cepstral coefficients

(MFCC) can be considered as a representation of the short-term power spectrum of a sound, which is supported on a linear cosine transform of a log power spectrum on a nonlinear frequency. Mel-frequency cepstral coefficients collectively create a Mel Frequency Spectrum. They are derived from a type of cepstral representation of the audio clip (Velankar, 2014). MFCCs are commonly used as features in speech recognition systems, Genre classification and in audio similarity measures. Formants are defined because of the spectral peaks of the acoustic spectrum of the voice. They are the distinguishing or meaningful frequency components of human speech and of singing.

Li and Wang (Li and Wang, 2007) have proposed a computational auditory scene analysis system to separate voice from music accompaniment for single-channel recordings. The stages in their approach consist of singing voice detection stage, pitch detection stage using Hidden Markov Model and separation stage. The most remarkable approach of their research is that they have separated voice from monaural recordings where the channel is mono-aural. They have described that the majority of sounds generated during singing is voiced, while speech has a larger amount of unvoiced sounds. The persons who study the sound of the human voice who are also known as phonecists divide the consonants into two types, voiced and voiceless. Voiced consonants require the use of the vocal cords to produce their signature sounds; voiceless consonants do not (Kenneth, 2004).

As Li and Wang used the Hidden Markov Model,(Ozerov et al., 2007) "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs" have introduced a general method for source model adaptation which is expressed in the framework of Bayesian models. Particular cases of the proposed approach are then investigated experimentally on the matter of separating voice from music in popular songs. The obtained results show that an adaptation scheme can improve consistently and significantly the separation performance as compared with non-adapted models. The difference of spectral distribution (timbre) of voice and instruments, and modeled them by Gaussian Mixture Model. In their method, the GMM had been trained in advance in a supervised way and tuned adaptively for each input. An estimation technique to locate the singing pitch and then separating the singing voice jointly and iteratively had been done by Hsu et.al. (Hsu et al., 2012). They have used the Tandem Algorithm to detect multiple pitch contours and then separates the singer by estimating the Ideal binary mask.

Another approach had been proposed by Tachibana et al. in “Singing voice enhancement in monaural music signals supported two-stage harmonic or percussive sound separation on multiple resolution spectrograms” (Tachibana et al., 2014) where they have considered and focused on the fluctuation of the singing voice and on detecting it by using differently resolved spectrograms. It is based on the pitch estimation parameter. They have proposed percussive sound separation system on multiple resolution spectrograms.

The concept of using matrix factorization has been used by Zhu et al. in their research “Multi-stage non-negative matrix factorization for monaural singing voice separation” (Zhu et al., 2013) where they developed a new algorithm for monaural singing voice separation. The algorithm used Non-negative Matrix Factorization to decompose long window and short window mixture spectrograms and then employed a spectral discontinuity and a temporal discontinuity thresholding method to select components for the two negative matrix factorization respectively.

By focusing on the principle that musical accompaniment is an interference in singing just like background noise is an interference in the speech signal, (Umesh and Sinha, 2007) had conducted their research “A study of filter bank smoothing in MFCC features for recognition of children’s speech”. As they mention, the interference of singing is due to its harmonic’s changes and repetition in the song. They have addressed during vocal tract length normalization, the Bandwidth of the Mel Frequency Cepstral Coefficient filters should not be scaled, only the center frequencies should be scaled to get improved performance. In real-world sound, sources are usually mixed with different audio signals. The process during which individual sources are estimated from the mixture signal labeled as Sound Source Separation.

Robust Principal Component Analysis algorithm was proposed (Huang et al., 2012) for singing voice separation from monaural recordings. This method has used decompositions of the low-rank matrix and sparse matrix of the input data matrix. Singing voice separation has been divided into two main parts namely a supervised system in which training data is required and an unsupervised system in which training data is not required. (Lin et al., 2010) “The augmented Langrange multiplier method exact recovery of corrupted low-rank matrices” addressed algorithm known as the Augmented Lagrange Multiplier (Matrix recovery method) for exact recovery of corrupted low-rank matrices. This method has included optimization techniques and a fast convergence rate.

Another technique mostly used in the literature in vocal separation is Robust Component Analysis. Candes et.al. (Candes et al., 2009) has addressed this principle with a detailed derivation in their paper. The results they have obtained by recovering the principal component of the data matrix have been demonstrated in their research paper. Their discussion is highly focused on introducing an algorithm to solve optimization problems. “Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization” (Wright et al., 2009) addressed about Robust Principal Component Analysis method with the formulation. They have proved that by using their proposed technique, matrices can be recovered efficiently.

“Melody Extraction from Polyphonic Music Signals” (Salamon and Gómez, 2010) thesis addressed about melody extraction application from polyphonic music signals. They have discussed about general information of the music signal and its properties and also explained the important definitions. Their task has been limited to a single source predominant fundamental frequency estimation from musical content with a lead voice or instrument. They have described the challenges melody extraction faces from a signal processing point of view and has noted the differences between melody extraction, monophonic pitch estimation, and multi-pitch estimation. By means of a case study, they have highlighted some of the most common errors made by melody extraction algorithms and has even identified their possible causes.

In order to overcome these errors (Rafii and Pardo, 2011) had proposed a method for separation of singing voice and music. The main theory they have applied is that if there exists a frame which is similar to some other frames within that particular song’s spectrum, that frame would be replaced by a measure of central tendency. They have used the median as their central tendency measure.

This approach is specifically based on discarding the repeating musical structures from the song which would result in the vocals separately.

Another important aspect in this area is finding a desirable dataset. A set of criteria had been proposed for singing voice separation (Hsu and Jang, 2010). It states that the singing voice and the musical accompaniment should be recorded separately therefore, the performance of the separation result can be evaluated by comparing it with the premixed singing voice, the manual explanation such as lyrics, pitch range, unvoiced types, variations and repetition of music note for each clip should be as sufficient as possible for all kinds of possible evaluations for singing voice separation and lastly the dataset should be publicly available without copyright issues. They have finally stated that the MIR-1K dataset meets all these criteria.

The past research papers had stated the fact that there are many challenges in separating the vocals from a song. Salamon et al. (Salamon et al., 2014) had addressed approaches, applications, challenges and case studies for melody extraction from Polyphonic Music Signals. The musical context of different types of mixtures of recording or live concerts is available. In that context, some portion is either the voice or the traditional speech for entertainment purposes. They have finally concluded that the main parameter to isolate vocals is the pitch ranges. REPET (Rafii and Pardo, 2011) (Repeating Pattern Extraction Technique) is an application implemented by Rafii and Pardo as an improvement for their older method with a large number of new datasets. This separates the music accompaniment and singing voice from the song. By focusing on the non-stationary signals, Vembu and Baumann (Vembu and Baumann, 2005) "Separation of vocals from polyphonic audio recordings" proposed a method. They have considered that the non-stationary signals as the vocal section. However, it is said this has a poor quality of source separation.

A comparison of the isolation approaches described above are depicted in the following Table 2.1.

Table 2.1: Summarizing of voice isolation approaches

#	Author	Research approach	Principle
1	(Li and Wang, 2007)	Using a HMM to detect singing voice through pitch	Pitch Detection
2	(Ozerov et al., 2007)	Detecting timbre difference of voice and instruments using adapted Bayesian models.	Timbre Identification
3	(Hsu et al., 2012)	Locating singing pitch and then separating the voice jointly and iterating using Tandem Algorithm	Pitch Detection
4	(Tachibana et al., 2014)	Focusing on fluctuation of singing voice and detecting them by using differently resolved spectrograms.	Pitch Estimation
5	(Zhu et al., 2013)	Using non negative matrix factorization to decompose long and short frames.	Spectral discontinuity thresholding
6	(Umesh and Sinha, 2007)	Considering instrumental music as background noise in speech signals.	Harmonic changes and repetition
7	(Huang et al., 2012)	Decomposing low-rank and sparse matrices using a suprvised system.	Robust Component Analysis
8	(Lin et al., 2010)	Using augmented Lagrange Multiplier for recovery of corrupted low-rank matrices	Robust Component Analysis
9	(Salamon and Gómez, 2010)	Fundemental frequency estimation.	Pitch Estimation
10	(Rafii and Pardo, 2011)	Repetition of instrumental music.	Similarity Matrix
11	(Vembu and Baumann, 2005)	Considering non stationary signals as voice signals.	Resting signals

The summarization shows some important principles used in each and every approach. It seems that most of the researchers had tried to pave their way using pitch detection. This had been done by assuming the fundamental frequency of the song lies on the vocals of the singer. Another popular mechanism to isolate vocals had been spectrogram analysis. Spectral discontinuity thresholding and robust component analysis had been driven on detecting changes in the spectrogram. Another interesting principle in songs, that had been analyzed is repeating patterns of the song and considering that partition as background music.

2.1.2 Artist Identification

"Singer Identification in Popular Music Recordings using voice coding features" (Kim and Whitman, 2002) proposed a method to classify singers based on Linear Predictive Coding. In this approach, the voice of the song is extracted and the source signal is analyzed and resynthesized according to a source-filter model of the human voice. Two different classifiers were trained using established pattern recognition algorithms, Gaussian Mixture Model and Support Vector Machine. Three different feature sets (linear scale data, warped scale data, and both linear and warped data) have been tested while two different classifiers (GMM and SVM) being used in each case.

"Artist detection in music using Minnowmatch" (Whitman et al., 2001) has addressed a model with a series of interconnected modules with swappable glue layers. This design has allowed the system to perform various music-IR related tasks in permuted configurations. The glue layer between each module can link over networks or on disk or memory. A support Vector Machine was trained and the output was fed into a neural network. After training each individual artist an SVM, a "Metalearner" training machine is created, A meta learner has been defined as a neural net that has n number of inputs, one for each magnitude output of an artist SVM, and n number of outputs weighted 0.1 on each except the correct artist, which is weighted 0.9. This neural net is then trained with each example of the learning representation. Each Learning Representation was fed through each previously trained SVM, and meta learner input vectors were created iteratively. After creating this meta learner dataset, it had been trained and the final resulting neural network was stored. For the test set, the stored meta learner was used to compute the most probable artist for each input slice example.

"Implementation of Singer Identification Model using K-Means Clustering Algorithm" (Dharini and Revathy, 2014) as the name suggests has been addressed using the K-means clustering algorithm. The training and testing phase had been done for direct film songs (vocal with background) for 10 singers. In the training phase 15 film songs of a singer had been taken as input. The input songs had been made to undergo a set of pre-processing steps. The three stages of preprocessing are pre-emphasis, frame blocking, and windowing. The Perceptual Linear Prediction (PLP) features had been extracted from each frames of the pre-processed signal. The singer model had been developed by the K-means clustering algorithm for each singer. In the clustering method, the cluster centroids had been obtained for a cluster size of 256 and stored. One model had been created for each singer by performing training and testing on the songs considered directly. The mean of minimum distances had been computed for each model. The singer had been classified based on the selection of the model which produces a minimum of average.

"Analysis and application of audio features extraction and classification method to be used for North Indian Classical Music's singer identification problem" (Deshmukh and Bhirud, 2014) has discussed the simplest suitable audio feature descriptor and therefore the classifiers to be used for the matter of Singer identification in North Indian general music. In contrast to western music, which is harmonious in nature, north Indian general music is more complex structure and requires perceptual analysis alongside a smaller number of audio descriptors and a straightforward method of classification so in order to reduce the computational complexity of the system. Several approaches had been analyzed and then proposed and a singer identification process had been implemented that reduces the complexity and increase the efficiency of the solution to the problem of identification of a singer in North Indian general music. The efficiency achieved by combining RMS energy, Brightness, and Fundamental Frequency had been found to be 70 percent when K-means clustering has been used for classification of the singer of north Indian classical songs.

"Audio Signal Classification" (Subramanian et al., 2004) has proposed an approach to classify audio signals. He has further discussed the features which would be used in order to proceed in his approach. A number of features like pitch, timbral, rhythmic features had been discussed with regard to their ability to differentiate the various audio formats. The selection of the important features as well as the common techniques used for classification had been explained. Lastly, an approach called the confusion matrix had been studied in order to evaluate the performance of the classification system.

"F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search" (Fujihara et al., 2006) had proposed a method for estimating F0s of vocal from polyphonic audio signals. Based on the existing multiple-F0 estimation method, the vocal probabilities of the harmonic structure of each F0 candidate had been evaluated. In order to calculate the vocal probabilities of the harmonic structure, the harmonic structure had been extracted and resynthesized by using a sinusoidal model and extract feature vectors. Then the vocal probability had been evaluated by using vocal and non-vocal Gaussian mixture models (GMMs). Finally, F0 trajectories had been tracked using these probabilities based on the Viterbi search.

A comparison of the isolation approaches described above are depicted in the following table 2.2.

Table 2.2: Summarizing of artist identification approaches

#	Author	Research approach	Evaluating model used
1	(Kim and Whitman, 2002)	Linear Predictive Coding	SVM and GMM
2	(Whitman et al., 2001)	Training a neural network	SVM and a metalearner(Neural Network)
3	(Dharini and Revathy, 2014)	Perceptual Linear Prediction	K-means Clustering
4	(Deshmukh and Bhirud, 2014)	Using sundamental frequency, RMS energy and brightness as features	K-means Clustering
5	(Subramanian et al., 2004)	Using features such as pitch, timbral, rhythemic features	Confusion Matrix
6	(Fujihara et al., 2006)	Estimating fundamental frequency	A pre-trained vocal and a non-vocal GMM

Most of the past researchers had used the K-means clustering model for artist identification purposes. Most of the features extracted seem to be similar. The reason for that may be because vocals show their uniqueness through the primary features which are spectral energy, frequency, and timbral features. Some researchers had used SVM and GMM models for evaluation.

2.2 Conclusion of the literature Review

In conclusion, much research had been taken place for vocal isolation and artist identification. Though there have been many approaches used, for vocal isolation, the most common way had been pitch estimation assuming that the fundamental frequency of the song lies in the vocal partition of the song. Artist identification modeling had mostly based on K-means clustering.

Pitch estimation has advantages and disadvantages when used for voice isolation. The main reason why it is not used in this project is that the fundamental frequency of a Sri Lankan song does not lie fully on the vocal partition of a song. It may lie on the instrumental partition too. Therefore, using pitch estimation to isolate vocals had been discarded in this project. Robust component analysis had worked primarily on songs which were of the genre rock and pop. But most of the Sri Lankan songs had been of the classical genre. Therefore, using robust component analysis had also be avoided.

K-means is excellent in fine-tuning cluster borders locally but fails to relocate the centroids globally. K-means cannot either relocate centroids that are not needed or where more centroids are needed because there can be stable clusters in between. (Fränti and Sieranoja, 2019)

Therefore, using K-means for clustering in this project is avoided. The next most popular methodology had been using two pre-trained models for vocal and non-vocal sections and evaluated afterward. But due to the infeasibility to locate that many numbers of separated vocals for a singer to train the vocal model, this method had been discarded too.

The reason why the Support vector machine was not used for modeling was SVMs do not perform well on highly skewed/imbalanced data sets. These are training data sets in which the number of samples that fall in one of the classes far outnumbers those that are a member of the other class. As the dataset used here is highly skewed, SVM s were not used. (*Singing scales*, 2000)

This section had discussed the methodologies past researchers had used in both voice isolation and artist identification. This section also had described why some methods could not be adapted to this research project and the reasons why they were discarded. The next chapter would discuss the methods which were adapted efficiently to use in this research.

Chapter 3

Design

Past researching had been done separately for voice isolation and singer identification as discussed in chapter 2. The main difference with this research and their findings is that in past work, singer identification had been done with the use of audio clips containing musical octave sung by the artist. The illustration below Figure 3.1 represents the C major scale in music which had been sung as inputs for those researches. C major scale when sung gives a perfect understanding about the singer's voice. (*Singing scales*, 2000) C major scale in eastern music consists of the notes "Sa", "Re", "Ga", "Ma", "Pa", "Dha" and Ni" sounds.

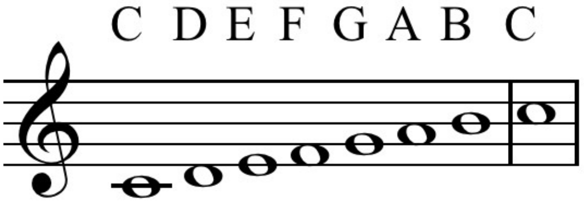


Figure 3.1: C Major scale in music

This research had focused on generating digital signatures for singers by using their songs because the differentiation of the same songs which may be hard to be identified using our hearing by who sings which. Therefore, in contrast to the past work, this research design includes singer identification using the songs of the artists as input. The flow of this research includes three main stages,

- Voice separation in songs.
- Feature Extraction.

- Signature Generation.

These three stages had been designed in doing the research, due to several rationale. The reasons are discussed comprehensively in the Section 3.2.

3.1 Conceptual overview of the Project

The flow of the research has addressed the following pattern as depicted in the Figure 3.2. The Voice of a particular singer is isolated using audio tracks of the artist. Unique features of the voice is extracted after observing the spectrum of the vocal audio. Using those features, a distinctive model for that specific vocalist is generated.

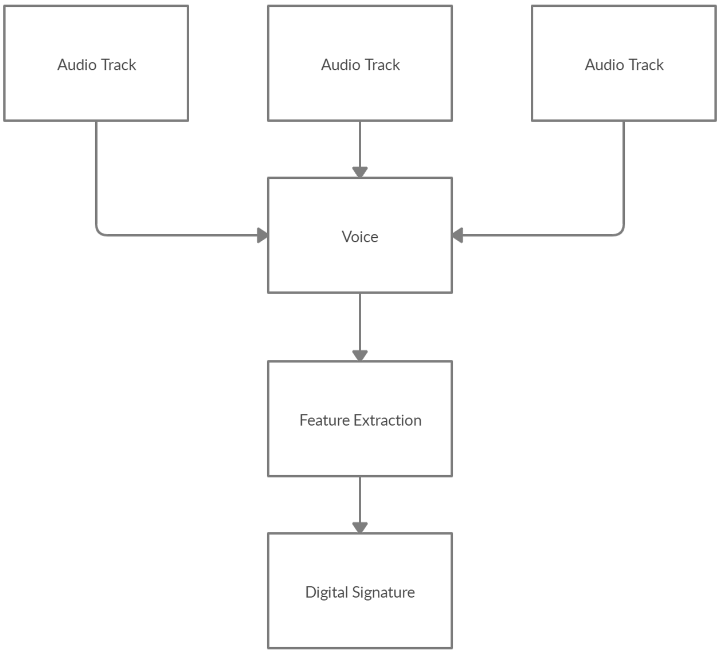


Figure 3.2: The flow of design of the research

Voice isolating process had taken up about 20 songs per artist, attenuates instrumental music in them using justified filters in order to isolate vocals. Features of the songs had been extracted using specific libraries in python and finally a signature (model) per artist is generated from those features.

3.2 Discussion of the Overview

The discussion below details the overview in Figure 2.1. The justification of the design and the methodology is discussed in detail under each subsection along with the challenges of accomplishing each of the sub tasks.

3.2.1 Voice Isolation

The vocal isolation has been done in this research in order to stop the features of background music be falsely considered as features of the vocalist. As this research is done in order to identify the singer and not the song, this step is critical and necessary.

When considering some genre classification tasks, it is visible that the features had been extracted from the raw input, which is here the song track.(Pampalk et al., 2005) But for tasks like genre classification, the features which get extracted from instrumental background is necessary. For an example, the instrument "Banjo" is used in "Folk" genre of music. (*What are the different styles of music played on the banjo?*, 2010). Which means the usage of a specific instrument in a particular song might even decide of what kind of music that song belongs to. But when in the task of artist identification, it is possible by two different artists to use the same instrument in their songs. For instance, two artists using a Banjo for instrumental music in the two of their songs does not mean both the songs are sung by one specific artist. But it might classify both the songs' genre as folk. Therefore it is crucial to remove the instrumental background music from the songs. How the evaluation accuracy had risen up when isolating vocals will be discussed in results and evaluation section in order to supply evidence for this logic.

When examining a Sri Lankan song, a typical and a common structure can be observed. The structure of a typical Sri Lankan song can be seen in the following Figure 3.3.

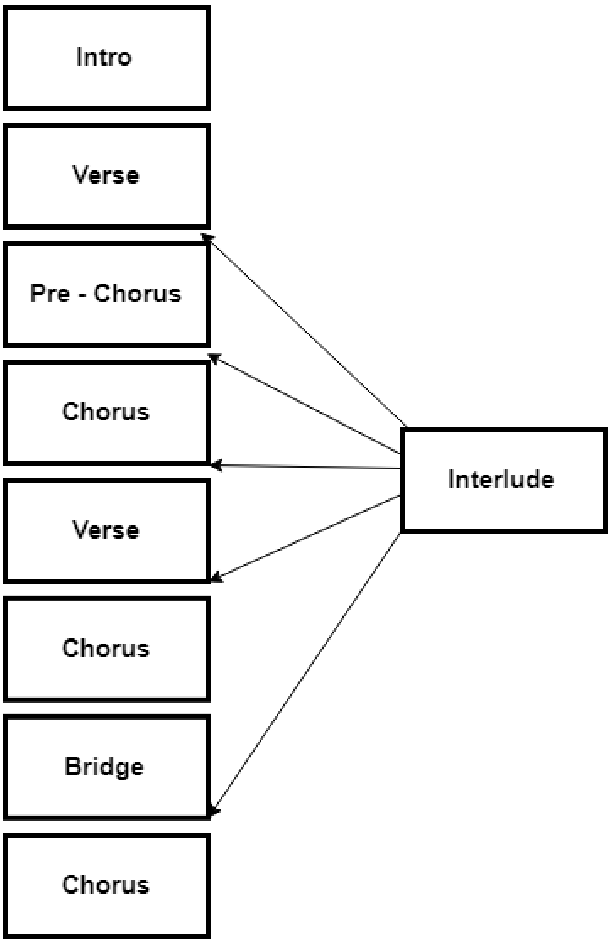
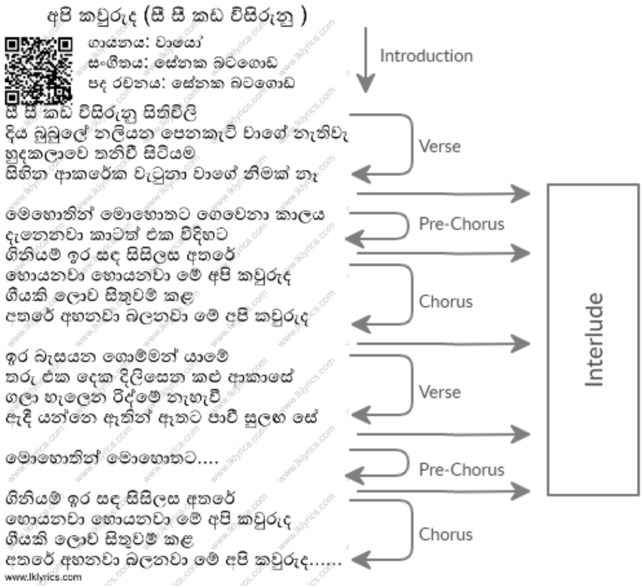


Figure 3.3: The structure of a typical Sri Lankan song

How a Sri Lankan song can be divided into so-called partitions can be depicted in Figure 3.4. The song used here is ”Api kauruda” which was sung originally by Mr. Senaka Batagoda.



අපි කවුරුද (සී සී කඩ විසිරුනු)

ගායනය: වායෝ
සංගීතය: සේනක බටගොඩ
පද රචනය: සේනක බටගොඩ

සී සී කඩ විසිරුනු සිතිවිලි
දිය බුබුලේ නලියන පෙනකැටි වාගේ නැතිවැ
හුදකලාවෙ තනිවී සිටියම
සිහින ආකරේක වැටුනා වාගේ නිමක් නෑ

මෙහොතින් මොහොතට ගෙවෙනා කාලය
දැනෙනවා කාටත් එක විදිහට
ගිනියම් ඉර සඳ සිසිලස අතරේ
හොයනවා හොයනවා මේ අපි කවුරුද
ගීයකි ලොව සිතුවම් කළ
අතරේ අහනවා බලනවා මේ අපි කවුරුද

ඉර බැසයන ගොම්මන් යාමේ
තරු එක දෙක දිලිසෙන කළු ආකාසේ
ගලා හැලෙන රිද්මේ නැහැවී
ඇදී යන්නෙ ඈතින් ඈතට පාවී සුලඟ සේ

මොහොතින් මොහොතට....

ගිනියම් ඉර සඳ සිසිලස අතරේ
හොයනවා හොයනවා මේ අපි කවුරුද
ගීයකි ලොව සිතුවම් කළ
අතරේ අහනවා බලනවා මේ අපි කවුරුද......
www.lklyrics.com

Figure 3.4: The song “Api kawruda” partitioned

The Sinhala song consists of background music which consists of sounds of various musical instruments. In this research, the Sri Lankan song has been examined thoroughly and how the features of the vocals and music differ from each other is observed. The observations made are,

1. The song can be divided into harmonic and percussive partitions. Vocals are inside the harmonic partition.

2. The background music comprises of the same pattern. (Guitars,piano chords, drums etc.)

3. The vocals are inside the frequency range 85 to 855 Hz.

4. The introductory and ending parts of the song mostly consist of instrumental music or silence.

When listening to a song, there exists a wide variety of different sounds. However, on a very coarse level, many sounds can be categorized to belong in either one of two classes, harmonic or percussive sounds. Harmonic sounds are the ones which we perceive to have a certain pitch such that we could for example sing along to them. The sound of a violin is a harmonic sound. Percussive sounds often stem from two colliding objects like for example, hitting the drums. Percussive sounds do not have a pitch but a clear localization in time. Mostly singing vocals have harmonic features(*Harmonic Percussive Source Separation*, 2014). Therefore, in order to remove the percussive sounds(drum beats, symbol beats), harmonic-percussive source separation has been used in this project.

The repeating pattern of a song in the instrumentals comprises of the same melody in almost all the songs. Most Sri Lankan songs do not exhibit a sudden change in rhythm or melodiousness. In a musical composition, a chord progression is a succession of chords. In tonal music, chord progressions have the capacity of setting up or negating a tonality, the specialized name for what is ordinarily comprehended as the "key" of a tune or piece. The chord progression gives the song a particular color. In order to maintain that specific key or the color of the song, the chord progression, and the rhythm is kept the same throughout the song. Therefore, instrumental music can be unmasked by detecting a similar pattern of the song as depicted in the spectrogram. Afterward, the amplitude of the frames in which has repetition is lowered to enhance vocals.

The voice is produced from the sound when air from the lungs vibrates the vocal chords in the throat. The air in the empty spaces of the chest, throat, and mouth vibrates and intensifies the sound of the voice. The vibrations of the vocal chords reverberate in the cavities of both the chest (in the lower register) and the head (in the upper register). Low notes are produced by loose vocal strings whereas high notes are produced by tight vocal strings. The artist naturally changes the shapes and sizes of these cavities to create the required notes. Women typically sing in four groups of voice ranges: soprano, mezzo-soprano, and contralto. Men are typically separated into four groups: countertenor, tenor, baritone, and bass. Men's voices are deeper than ladies' as their vocal chords are longer. When people sing together, men actually sing an octave lower: the ordinary scope of ladies' voices is in the treble clef, and the men's is in the bass clef. Each voice has its own regular scope of the pitch. The highest range of lady's voice is the soprano, and the lowest the contralto, or alto. The deepest male voice is the bass and the highest is normally the tenor. Some male artists have a characteristic augmentation at the highest point of their range which permits them to sing in the alto, or countertenor territory. (*Voice Classification: An Examination of Methodology*, 2013) These are the ranges of vocals,

1. Soprano: the highest female voice, being able to sing C4 (middle C) to C6 (high C), and possibly higher.
2. Mezzo-soprano: a female voice between A3 (A below middle C) and A5 (2nd A above middle C).
3. Contralto: the lowest female voice, F3 (F below middle C) to E5. Rare contraltos have a range similar to the tenor.
4. Tenor: the highest male voice, B2 (2nd B below middle C) to A4 (A above Middle C).
5. Baritone: a male voice, G2 (two Gs below middle C) to F4 (F above middle C).
6. Bass: the lowest male voice, E2 (two Es below middle C) to E4 (the E above middle C)

This classification depicts that the highest frequency range of humans as Soprano while the lowest as the Bass. The soprano's vocal range (utilizing logical pitch documentation) is considered from around middle C (C4) = 261 Hz to "high A" (A5) = 880 Hz in choral music. The frequency range of a typical adult bass singer is said to be from 85 to 180 Hz. This concludes that the singing vocal frequency range can be defined as 85 Hz to 880 Hz.

The introduction of a song is longer than the interludes of the song, it is found at the beginning and sets up the song, establishing many of the song's key, tempo, rhythmic feel, energy and attitude. The goal of the introduction of a song is to make the listener interested of the song. In contrast, interludes try to link verse with the chorus and do not need the attention required when composing the introduction of the song. (*Basic Song Structure Explained*, 2011) This feature of songs had made the introduction of the song more stronger instrumentally than the interlude. If that stronger introduction partition is removed, it can be assumed that the features of the singer's voice are extracted in a better manner. Even the instrumental music is removed completely from a song in this research, a long period of silence is existent in the track. Lots of research had been conducted to show how the removal of silenced and unvoiced segments had improved the efficiency of the performance of the system.

Silence and unvoiced signal removal can be considered as a pre-processing technique used to remove silence (background noise) and unvoiced segments from the input signal. Silence removal had been very helpful portion of proposed technique to reduce processing time and increase the performance of system by eliminating unvoiced segments from the input signal.(Sahoo and Patra, 2014) Researching had been done in order to capture unvoiced and silenced signals, therefore, it enhances the performance of the speech/vocal signal processing.This adaptation is conformed in this research to see if that enhances the voice isolation process. When the silence of the introductory and ending partitions is reduced, the extraction of the features

By using those observations, some efforts had been made to reduce the effect of the background music and enhance the vocal part of the singer. The efficiency of these observations were made in the evaluation, which will be discussed in the Results and Evaluation section. The following are the steps which were performed to strengthen the previous observations.

1. Harmonic Percussive source separation using median filtering.

2. Voice extraction using Similarity Matrix.

3. Introducing a Butterworth Band-Pass Filter.

4. Eliminate the introduction and ending of song. (Complementary Step)

3.2.1.1 Harmonic Percussive source separation

The following Figure 3.5 depicts what this filter is capable of.

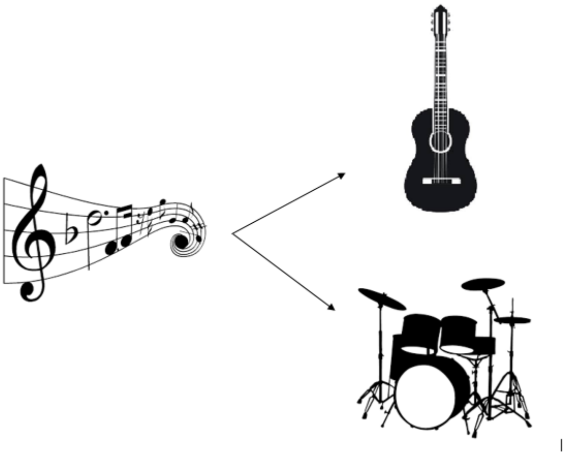


Figure 3.5: Harmonic and Percussive source Separation

The system of this technique includes the usage of median filtering on a spectrogram of the sound signal, with median filtering performed across progressive frames to stifle percussive occasions and improve harmonic partitions, while median filtering is also performed across frequency bins to strengthen percussive occasions and supress consonant segments. The two emerging median filtered spectrograms would generate masks which are then applied to the main spectrogram to isolate the harmonic and percussive pieces of the sign.

The approximation in this strategy is that considering harmonic occasions as vertical lines and percussive occasions as horizontal lines in a spectrogram. It very well may be considered as a valuable estimation when trying to isolate harmonic and percussive sources. Median filters work by replacing a given sample in a signal by the median of the sign values in a window around the example. Given an input vector x(n) and then y(n) is the yield of a median filter of length l where l characterizes the quantity of samples over which median filtering happens. Where l is odd, the middle channel can be characterized as:

$y(n) = median * (x(n{-}k : n + k), k = (l{-}1)/2)$

In the past strategy, there may be issues emerging. One issue is that the computed components are frequently not of purely harmonic or percussive in nature yet in addition contain commotion like sounds that are neither clearly harmonic nor percussive. Besides, depending on the parameter settings, one often can watch a spillage of harmonic sounds into the percussive segment and a spillage of percussive sounds into the harmonic segment. Consequently another methodology is extended utilizing two expansions to a state-of-the-art harmonic-percussive separation procedure to focus on the issues. Initially, a partition factor parameter is brought into the disintegration procedure that permits for fixing separation results and for upholding the segments to be unmistakably harmonic or percussive. As the second commitment, inspired by the classical sines+transients+noise (STN) sound model, this novel idea is exploited to highlight a third residual segment to the decomposition which catches the sounds that stay between the distinctly harmonic and percussive sounds of the audio signal.

3.2.1.2 Voice Extraction using Similarity Matrix

A similarity matrix is defined as two-dimensional representation where each point (a, b) measures the dissimilarity between any two elements a and b of a given sequence. Since, repetition is mostly used in the instrumental parts of Sinhala songs, it can be considered as what makes the structure in music. A similarity matrix calculated from an audio signal aids to reveal the musical structure that underlies it. Given a single-channel mixture signal x, first, its Short-Time Fourier Transform would be calculated using half overlapping Hamming windows of a particular length. Then, the magnitude spectrogram V is derived by taking the absolute value of the elements of X, after discarding the symmetric part, while keeping the direct current component. The similarity matrix S is then defined as the matrix multiplication between transposed V and V, after normalization of the columns of V by their Euclidean norm. The Figure 3.6 shows how the similarity matrix is generated.

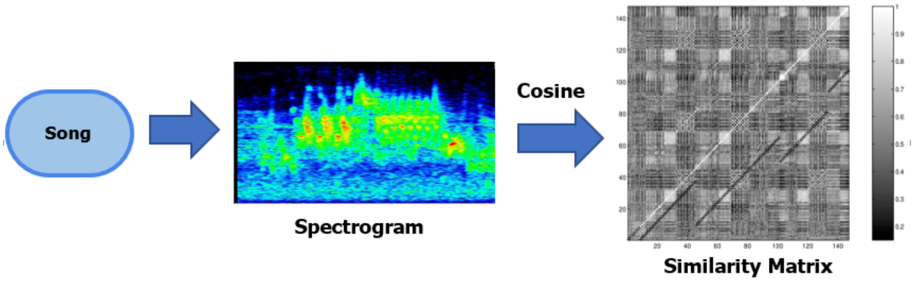


Figure 3.6: Generation of the similarity matrix

The calculation of the similarity matrix S is shown in Figure 3.7.

$$S(j^a, j^b) = \frac{\sum_{i=1}^{n} V(i, j^a) V(i, j^b)}{\sqrt{\sum_{i=1}^{n} V(i, j^{a^2})}\ \sqrt{\sum_{i=1}^{n} V(i, j^{b^2})}}$$

Figure 3.7: Calculation of similarity matrix

Where n is the number of frequency channels.

Once the similarity matrix S is calculated, the repeating elements can be identified in the mixture spectrogram V. For all the frames j in V, the frames that are the most similar to the given frame j are identified and saved in a vector of indices.

Assuming that the non-repeating foreground (vocal part) is sparse and varied compared to the repeating background (music part), a reasonable assumption could be taken as the repeating elements revealed by the similarity matrix should be those that form the bottom-line repeating structure. This approach proved to be better as it allowed not only the identification of patterns which are periodic, but also patterns which did not necessarily happen in a periodic fashion.

In order to limit the number of repeating frames considered similar to the given frame j, k was defined as the maximum allowed number of repeating frames. Correspondingly, t was defined as the minimum allowed threshold for the similarity between a repeating frame and the given frame. Another parameter d was defined as the minimum allowed distance between two consecutive repeating frames deemed to be similar enough to indicate a repeating element.

By following the rationale, “the non-repeating foreground (voice) has a sparse time-frequency representation compare to the time-frequency representation of the repeating background (music)”, the researches had come to a conclusion that time-frequency bins with little deviations between repeating frames would constitute a repeating pattern and would be captured by the median. Once the repeating elements have been identified for all the frames j in the mixture spectrogram V through their corresponding vectors of indices, they had been used to derive a repeating spectrogram model W for the background by taking the median of the k number of frames.

The process of generating the repeating spectogram using the similarity matrix is illustrated in the Figure 3.8.

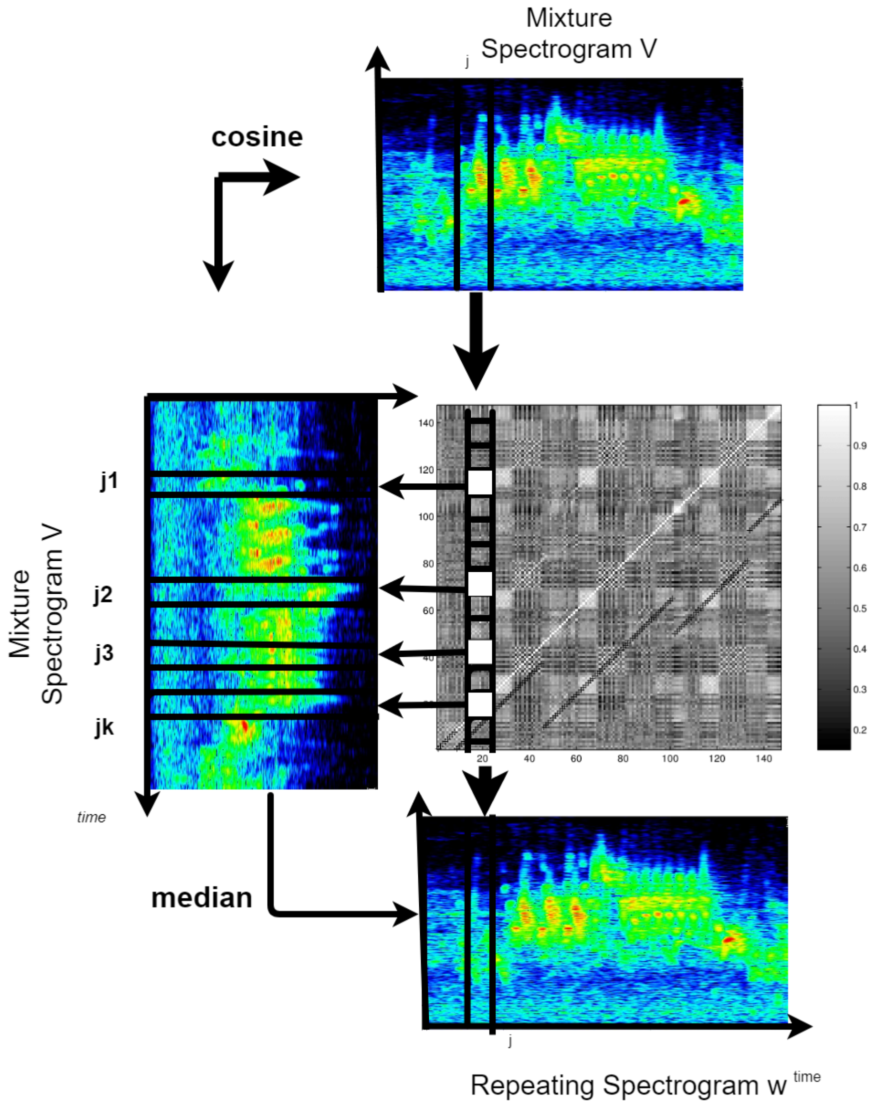


Figure 3.8: Generation of the repeating spectogram using similarity matrix

After generating the repeating spectrogram model, a time-frequency mask is derived by normalizing the repeating spectrogram model. The time-frequency mask is then symmetrized and applied to the Short time Fourier transform of the mixture signal x. The estimated music signal is finally obtained by inverting the resulting STFT into the time domain. The estimated voice signal is obtained by simply subtracting the music signal from the mixture signal.

3.2.1.3 Frequency Filtering using a Band-pass filter

The Butterworth filter is expressed as a form of signal processing filter designed to have a frequency response as flat as acheivable in the passband. It is additionally referred to as a maximally flat magnitude filter.The frequency response of the Butterworth filter is maximally flat (i.e. has no ripples) within the passband and rolls off approaching zero in the stopband.(Abubakar Sadiq et al., 2018) When viewed on a logarithmic plot, the response is a slope which declines off linearly towards negative infinity. A first order filter's response falls off at −6 dB per octave (−20 dB per decade) (all first-order lowpass filters have identical normalized frequency response). A second-order filter decreases at −12 dB per octave, a third-order at −18 dB and likewise. Butterworth filters have a monotonically wavering magnitude function with ω, unlike other filter forms that have non-monotonic ripple in the passband and/or the stopband.

When compared with a Chebyshev Type I/Type II filter or an elliptic filter, the Butterworth filter has a lethargic roll-off, and so woould require a higher order to implement a specific stopband specification, however Butterworth filters have an extra linear phase response in the pass-band than Chebyshev Type I/Type II and elliptic filters can accompish. The following Figure 3.9 represents a butterworth band pass filter.

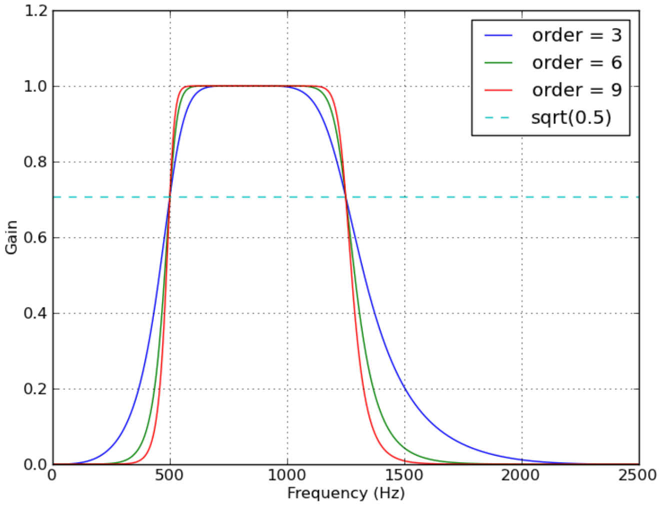


Figure 3.9: Butterworth Band-pass filter

3.2.1.4 Eliminating introductory and ending parts

A survey analysis has been conducted to observe the significant amount of time used in the introduction and the end of the song consisting with only instrumental music or silence in the case where after vocals were being isolated. As sound is a vibration, there is a capability to access frame by frame and to check whether it is silent or not. This has been achieved in this research using the PyDub library in python. The silence threshold used in this project has been −50 decibels. The threshold had been found through a trial and error methodology, as per the audios have been of the same quality. The reasoning behind the usage of trail and error method for this functionality is the silence threshold depends hugely on the quality of the audio and the duration of the silence in the audio. When listening to the audio after trimming the introduction, the audio had been in a satisfactory level in this project.

3.2.2 Feature Extraction

The audio signal is a three-dimensional signal in which represent time, amplitude and frequency. The features suitable for speech signals were selected for this project. The features chosen are, MFCC, zero Crossings Rate, Spectral Centroid and Spectral Rolloff.

3.2.2.1 MFCC

MFCC features are included in the recognized discrepancy of the human ear's bandwidths with frs spaced nearly at low frequencies and logarithmically at high frequencies have been used to retain the phonetically vital properties of the speech signal.(Alim and Alang Md Rashid, 2018) As they had discussed in their paper, MFCC computation is considered as a replication of the human hearing system intending to artificially implement the ear's working principle with the assumption that the human ear is a reliable speaker recognizer which means that this would act as an artificial human ear. Therefore, it had been essential to use MFCC features in this project which gives information of the whole spectrum of the vocals.

Mel-Frequency Cepstral Coefficients (MFCCs) can be defined as a form of dimensionality reduction. One might pass a collection of audio samples, and receive

10 to 20 cepstral coefficients that describes that sound in a typical MFCC computation. While MFCCs were initially developed to represent the sounds made by the human vocal tract, they have turned out to be a pretty solid timbral, pitch invariant feature, that has all sorts of uses other than automatic speech recognition tasks.When obtaining MFCCs, the first step is to compute the Fourier transform of the audio data, which converts time domain signal into a frequency domain signal.

Then the power spectrum from the frequencies computed are taken and Mel-Filter bank is applied to them. This process can be simplified as summing the energies in each filter. The Mel-Frequency scale relates to perceived frequency of a pre tone compared to its actual measurement of the pitch which means that us humans are much better at noticing small perturbations in lower frequencies that we are at high frequencies. Applying this scale to the power spectrum is how it is related to the features to what humans actually perceive. Then the logarithm on each of the filtered energies are computed, which is motivated by human hearing that doesn't perceive loudness in a linear scale.

Finally, the cepstrum is computed. A cepstrum can be simplified as a spectrum of a spectrum. In order to retrieve the cepstrum, Discrete Cosine Transform (DCT) of the log filter bank energies should be computed, which gives the periodicity of the spectrum. The periodicity of the spectrum shows how quickly the frequencies themselves are changing. The DCT is a similar transform the Fourier transforms, but the DCT only returns values that are real numbers where the DFT returns a complex signal of imaginary and real numbers.The MFCC geneation from speech signals can be depicted as Figure 3.10.

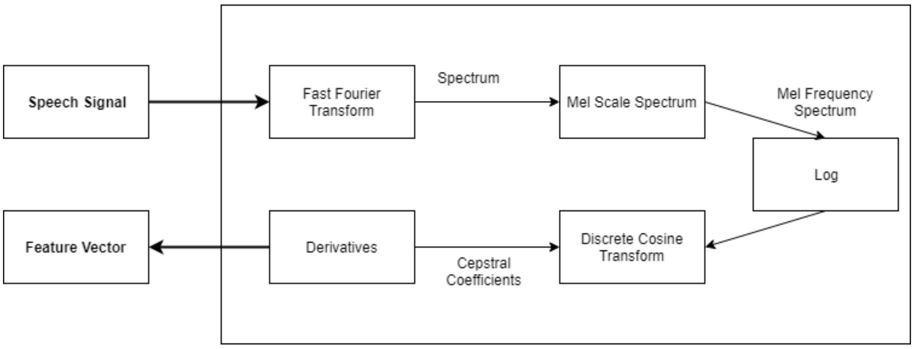


Figure 3.10: Generation of MFCC from the speech signals

3.2.2.2 Zero Crossings Rate

The zero-crossing rate is the rate of the sign being changed along a signal, the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval whereas it usually has higher values for highly percussive sounds like those in metal and rock. Therefore the artists are classified according to the significant features of their voices for and example, a rock singer usually has higher zero crossings rate when compared with a classical singer. It had seem that artist classification can be achieved through this feature and hence this feature was extracted from each and every isolated vocal.

3.2.2.3 Spectral Centroid

This indicates where the " center of mass" for a sound is located and is calculated as the weighted mean of the frequencies present in the sound. If the frequencies in music are same throughout then spectral centroid would be around a center and if there are high frequencies at the end of sound then the centroid would be towards its end. Spectral centroid is considered a good indicator of brightness. In music, timbre, also known as tone color or tone quality, is the perceived sound quality of a musical note, sound or note. Timbre is the term referred to as brightness here. Timbre distinguishes different types of vocals. As the spectral centroid can distinguish between the tonal colour or the timbre of vocals this feature had to be extracted. (*Introduction to Audio Analysis*, 2014)

3.2.2.4 Spectral Rolloff

Spectral rolloff is the frequency below which a specified percentage of the total spectral energy, e.g. 85 percent, lies.It also gives results for each frame. Kos et.al. (Kos et al., 2013) had discussed the usage of spectral roll-off in acoustic classification and emphasized music/voice classification in their paper The spectral roll-off is a timbre feature. As it produces features of the timbre of voice, that feature had been extracted and used in this research project.

3.2.3 Signature Generation

Saini et. al.(Saini and Jain, 2013) have discussed in the research paper "Comprehensive Analysis of Signal Processing Techniques Used For Speaker Identification" how using MFCC features with the GMM model had given an exceptional performance in most of the speaker identification tasks. They had also mentioned as Speaker recognition is more of a biometric task, speaker-related recognition activities like artist recognition can presumably have a better ending when a Gaussian Mixure Model is used. But they had also stated the importance knowing the domain knowledge of accoustics in necessary. Therefore, a Gaussian Mixture model had been used in the final process: signature generation.

A Gaussian mixture model is considered as a probabilistic clustering model to represent the presence of sub populations within an overall population. The reason behind raining a GMM is to approximate the probability distribution of a class by a linear combination of 'k' Gaussian distributions. The likelihood of feature vectors for a model is given by following equation:

$$P(X/\lambda) = \Sigma_{k=1}^{K} w_k P_k(X/\mu_k \Sigma k)$$

, where $P_k(X/\mu_k \Sigma k)$ is the Gaussian distribution.

$$P_k(X/\mu_k \Sigma k) = 1/\sqrt{\frac{2\pi}{\Sigma k}} \exp^{1/2(x-\mu k)\Sigma(x-\mu k)}$$

The training data X_i of the class λ are used to estimate the parameters mean μ, co-variance matrices Σ and weights w of these k components. Initially, it identifies k clusters in the data by the K-means algorithm and assigns equal weight w = 1/k to each cluster. 'k' Gaussian distributions are then fitted to these k clusters. The parameters μ, σ and w of all the clusters are updated in iterations until the converge. The most popularly used method for this estimation is the Expectation-Maximization (EM) algorithm. Therefore, it can be concluded that when feature vectors unique to each singer are provided, a GMM model unique to each of the singer can be retrieved. As the feature which is getting extracted would not be similar to two singers, it would be safe enough to come to a conclusion that the model generated for a particular artist can be considered as the digital signature generated for that singer. In the determination of the artist phase, a signature dump is used where comparing would take place. The signatures for singers would be saved in a signature dump. Whenever a signature is generated from a new song, that signature will be compared with the signatures in the signature dump and would prompt the most matching singer's signature. The Figure 3.11 depicts the approach of signature generation.

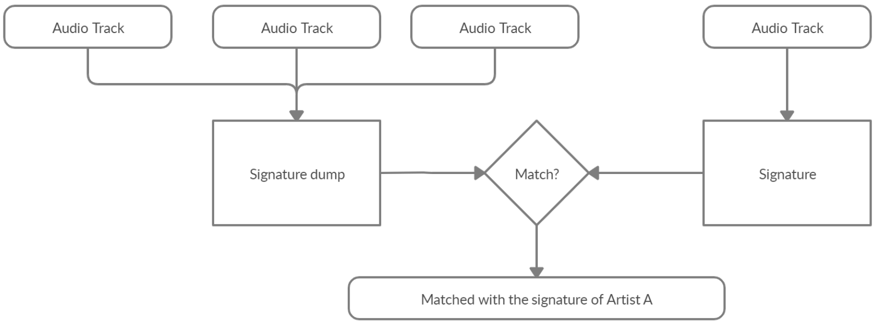


Figure 3.11: Design of signature generation

Let X be a time series of feature vectors selected and λ be the GMM for the singer s. Then, the signature of the singer is determined through the following equation,

$$S = \arg\max_i 1/T \Sigma_{t=1}^{T} logp(x_t/\lambda_i)$$

The signatures for each and every singer which resembles the equation mentioned above are generated using audio tracks of that artist. This signature can be considered as a unique model of that particular artist.

3.3 Evaluation Design

The evaluation design of this research mainly focuses on the effectiveness of the pre-processing stages used in the voice isolation section, and the ability of the introduced unique signature to differentiate between presumed artists and artistes. An effective and efficient procedure is followed in order to preserve the quality and value of the research.

3.3.1 Evaluation of Voice Isolation

Many pre-processing steps had been used in the stage of voice isolation in this research namely, voice extraction using a similarity matrix, harmonic percussive source separation, frequency filtering using a Band-pass filter and the elimination of introduction and the end. REPET (RepeatingPattern Extraction Technique) introduced by Rafii and Pardo [(Rafii and Pardo, 2011)]is kept as the base and the other filters are tested out combined, and separately to see their effectiveness in the voice isolation process. The following Figure 3.12 depicts how the evaluation takes place under voice isolation. Evaluation tasks that are to be carried on are also depicted in the following figure.

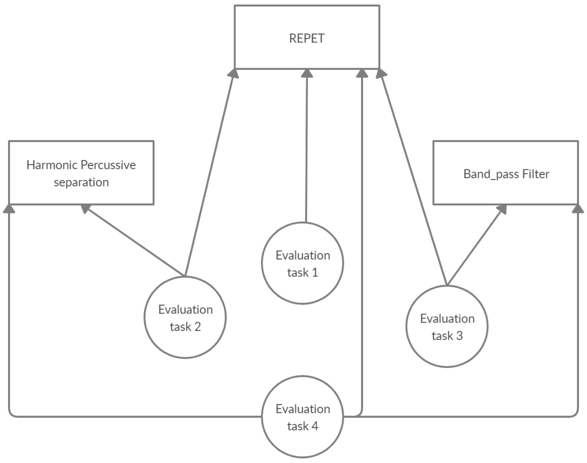


Figure 3.12: Evaluating tasks of voice isolation

The following 4 cases are the evaluation tasks as depicted in the above diagram. These evaluation tasks are carried out and the accuracy of artist identification is calculated for each case.

1. Case 1: When using REPET alone for voice isolation.

2. Case 2: When using REPET + Harmonic-Percussive source separation for voice isolation.

3. Case 3: When using REPET + Band-pass filter for voice isolation.

4. Case 4: When using REPET + Harmonic-Percussive source separation + Band-pass filter for voice isolation.

The first case is conducted in order to see the progress when using a state-of-art method in this research. The second and third cases are conducted in order to clarify if each filter separately improves the evaluation or not. The whole process from voice isolation to signature generation is carried out and finally, the accuracy of the artists is identified as a percentage and those accuracy percentages are produced for each and every evaluation task. The evaluation task which yields the highest accuracy or in other words, the winner of these subtasks is therefore recognized as the most suited voice isolation process to be used in this research.

The winner from those four cases will be evaluated once again against the tracks where the silenced partitions of introduction and ending are eliminated. This is done in order to see if there is an improvement in eliminating silence and unvoiced partitions of the song. The following Figure 3.13 depicts the final process of the evaluation design of the vocal isolation section.

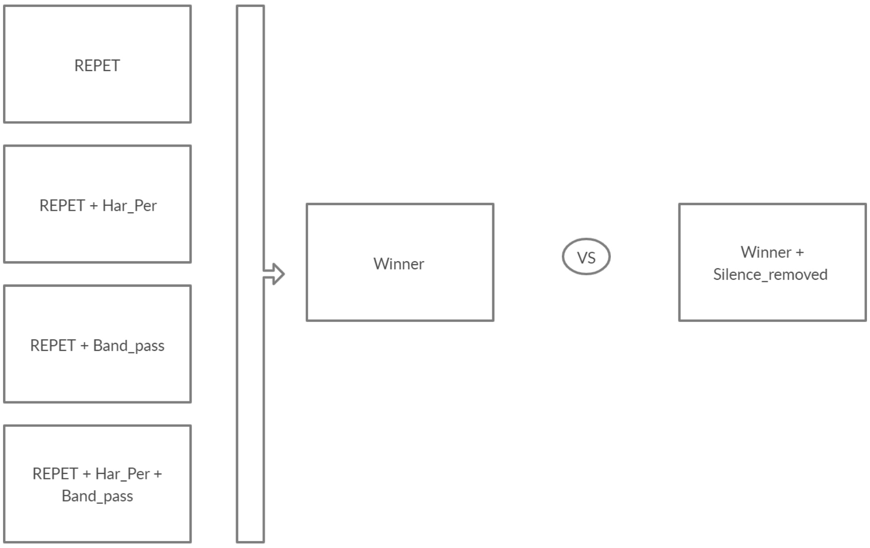


Figure 3.13: Evaluating task to compare with silence removed vocals

The best accuracy holder out of these two processes will be considered and treated as the best voice isolation methodology suited for the signature generation process.

3.3.2 Evaluation of Signature Generation

After identifying the most suited approach in this researching problem, to isolate vocals, the final evaluation is performed. The final evaluation mainly focuses on two scenarios. The following Figure 3.14 depicts how the final evaluation takes place.

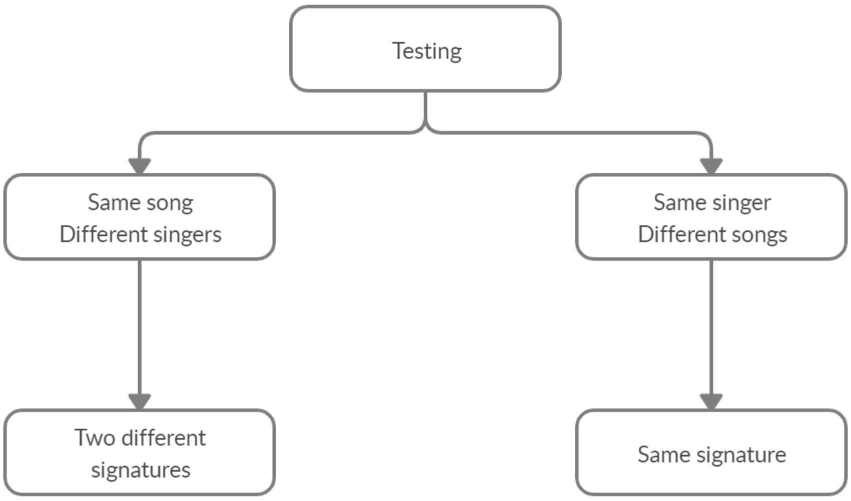


Figure 3.14: Evaluating task of signature generation

The reasoning behind this evaluation model is when trying to learn what is counterfeit and what is real, comparing will take place between real and fake singing. And mostly the same song will be sung by both the parties. In the dataset used, most similar voices are the voices of father-son or mother-daughter combinations. The same songs had also been sung by the sons of famous artists. The first evaluation is performed on the voices of father-son and mother-daughter combinations in the dataset. Assuming that the research is successful, they should generate different digital signatures and the singers should be identified specifically from each other.

The second evaluation is performed on the same singer using different songs. This should generate the same digital signature and hence should be identified as the same singer in the evaluation.

Chapter 4

Implementation

This project is implemented using open source software. It is written in Python language using the Spyder IDE. Appropriate libraries including python Librosa, MatplotLib, Pandas have been used throughout this project.

4.1 Discussion on technology used

This section discusses the technology that is used in this project and code implementations can be found in Appendix A. The overall architecture of the libraries and the connections among them is given in Figure 4.1. Librosa package has been used mainly for audio signal processing activities, while Matplotlib had aided in graph and plot generation. Numpy and Pandas had been mostly used for data analysis using data frames and multi dimensional arrays. Pickle module had been used for model storage and evaluation purposes.

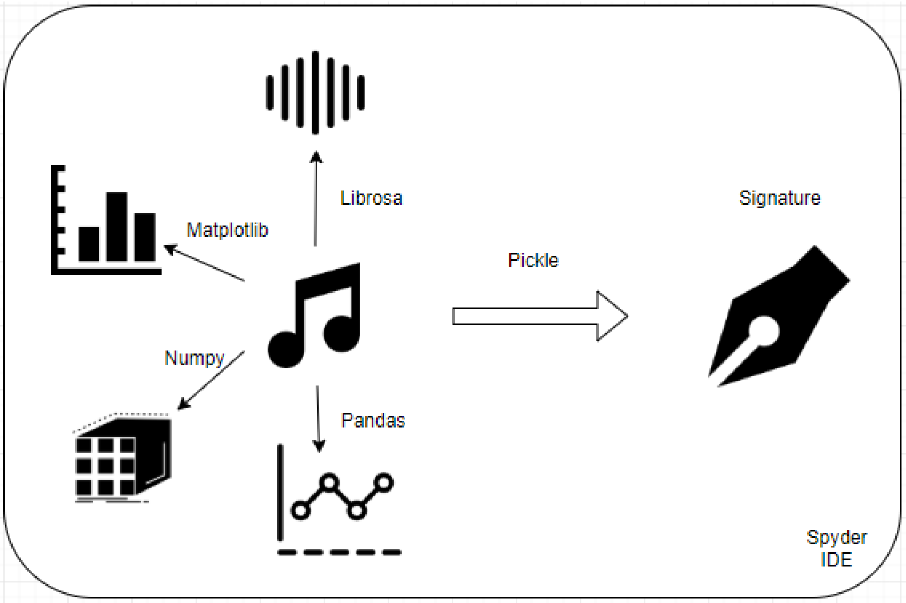


Figure 4.1: Overall Architecture of libraries

4.1.1 Personal Computer(PC)

The PC used in this project has the following specifications, as it has to have the processing power to accommodate the requirements of the training and model generation.

- Processor: Intel(R) Core i5-3210M
- RAM: 8GB
- Operating System: Windows 10

4.1.2 Spyder IDE

Spyder is considered as an open source cross-platform integrated development environment for scientific programming in the Python language. It integrates with a variety of prominent packages within the scientific Python stack, including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, including some other open source software. It is released under the MIT license. It is a strong scientific environment written in the python language which offers a combination of advanced editing, analysis, debugging functionality of a comprehensive development tool with the incorporation of the data exploration, interactive execution, deep inspection, and enticing visualization capabilities.

It has many built in features, but those are extended further by using its API and the plugin system. Furthermore, Spyder is also used as a PyQt5 extension library, which allows developers to build their own functionality and embed its components, like the interactive console, in their own PyQt software.

4.1.3 Python Librosa

LibROSA is considered as python package for developed specifically for music and audio analysis. It provides the infra-structure necessary to create music information retrieval systems. It includes core functionalities such as loading audio from disk, computing various spectrogram representations, and having a variety of commonly used tools for music analysis.

There exist functionalities for harmonic-percussive source separation (HPSS) and generic spectrogram decomposition using matrix decomposition methods implemented in scikit-learn.Time-domain audio processing, such as pitch shifting and time stretching can also be acheived using this library. This also provides time-domain wrappers for the decompose submodule.

Feature extraction and manipulation which includes low-level feature extraction, such as chromagrams, pseudo-constant-Q (log-frequency) transforms, Mel spectrogram, MFCC, and tuning estimation. Also provided are feature manipulation methods, such as delta features, memory embedding, and event-synchronous feature alignment can also be acheived using this module.

4.1.4 Matplotlib

Matplotlib is considered as a plotting library for the Python language and its numerical mathematics extension NumPy. It provides an object-oriented Application Proggramming Interface for embedding plots into applications using toolkits. There is also a procedural "pylab" interface supported on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

Pyplot is defined as a Matplotlib module which provides a MATLAB-like interface. It is meant to be designed as usable as MATLAB, with the power to use

Python, and therefore having the advantage of being free and open-source. One of the best benefits of visualization in Matplotlib is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots including line, bar, scatter and, histogram.

4.1.5 Pandas

Pandas is considered a software library written for the Python language. The major use of this library is for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and statistic time series. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the similar individuals.

There are huge amount of features in pandas library such as having data frame object for data manipulation with integrated indexing,having tools for reading and writing data between in-memory data structures and different file formats, being capable of data alignment and integrated handling of missing data. Reshaping and pivoting of data sets, label-based slicing, fancy indexing, and subsetting of large data sets can also be observed in pandas library. Pandas provides data filteration and dataset meging and joining additionally.

4.1.6 Scikit-learn

Scikit-learn is an open-souce software machine learning library for the Python language. It supports many classification, regression and clustering algorithms. These include support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is meant to operate parallel with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-learn is essentially written in Python, and uses numpy for algebra with high-performance and array operations likely. Furthermore, some core algorithms are written in Cython to enhance performance. Support vector machines are implemented by a Cython wrapper, logistic regression and linear support vector machines are written by an identical wrapper. In such cases, extending these methods with Python might not be possible.

Scikit-learn integrates well with many other Python libraries. Some of the integrations are integrating with matplotlib and plotly for plotting, integrating with numpy for array vectorization, pandas for dataframes, scipy, and many other libraries.

4.1.7 Pickle Module

Python pickle module is employed for serializing and de-serializing a Python object structure. Any object in Python are often pickled in such a way that it can be saved on disk. First Pickle "serializes" the object before saving it to file. Pickling is a way to convert a python object (list, dictionary) into a character stream. This gives the idea that this character stream consists of all the information necessary to reconstruct the object in another script.

4.1.8 Pydub

Pydub is another module introduced for audio analysis in python. It supports many analysis tasks including loading and saving different audio types, audio trimming, audio mixing, audio level changing and many more. It lets the user achieve various forms of manipulation within the audio. The major use of pydub in this project is audio trimming. Audio trimming had been beneficial when eliminating silence parts of the audio in voice isolation process.

4.2 Implementation of the functionalities

This section discusses how the required functionalities were implemented using predefined libraries in python. Steps which were implemented separately will be discussed here. In addition, benefits and drawbacks of each stage will further be discussed.

4.2.1 Data Gathering

This research approach uses for monaural (single channel) songs in the mp3(MPEG-1 standard) format as input and produces the digital signature of the corresponding singer as the output. The data set gathered consists of almost 600 songs of 30 artists and artistes. All the songs used were downloaded free from websites Ananmanan.lk and sarigama.lk. The dataset contains songs of 16 Male artists and 14 Female artists. This dataset includes songs of artists who have similar voices (eg: H.R. Jothipala and Greshan Ananda) including voices of father-son combinations. (eg: Milton and Ranil Mallawaarachchi, Mervin and Amal Perera)

Table 4.1 shows the list of the artists used in this project. M and F represents Male and Female.

Table 4.1: Artists of the tracks used in this project

#	Singer	Number of Songs	Gender
1	Amal Perera	20	M
2	W.D. Amaradewa	20	M
3	Amarasiri Peiris	20	M
4	Anjaline Gunathilake	20	F
5	Chandralekha Perera	20	F
6	Clarance Wijewardana	20	M
7	Dayan Witharana	20	M
8	Deepika Priyadarshani	20	F
9	Karunarathna Diwulgane	20	M
10	Greshan Ananda	20	M
11	Indrani Perera	20	F
12	Jagath Wickramasingha	20	M
13	H.R. Jothipala	20	M
14	Gunadasa Kapuge	20	M
15	Kasun Kalhara	20	M
16	Latha Walpola	20	F
17	Malani Bulathsinhala	20	F
18	Mervin Perera	20	M
19	Milton Mallawaarachchi	20	M
20	Nanda Malani	20	F
21	Neela Wickramasingha	20	F
22	Nelu Adhikari	20	F
23	Nirosha Virajini	20	F
24	Ranil Mallawaarachchi	20	M
25	Rookantha Goonathilaka	20	M
26	Samitha Mudunkotuwa	20	F
27	Shashika Nisansala	20	F
28	Sujatha Aththanayaka	20	F
29	T.M. Jayarathna	20	M
30	Umaria Sinhawansa	20	F

4.2.2 Voice Extraction

Preliminary stage of this project is to isolate vocals as extraction of features raw from the audio itself may lead to a research similar to song or genre identification rather than artist identification. This task had been attempted by previous researchers as discussed in the literature survey many having positive in conjunction with negative outputs. This section discusses the steps taken for the backing of the mentioned observations in chapter 1.

Degenerate Unmixing Estimation Technique (DUET) [39] uses cross channel timing and amplitude differences in a song to differentiate between accompaniment and voice. It is complex to apply to Sinhala songs due to reverberant effects added, also due to the violation of the sparsity assumption for music mixtures. There are many approaches which separate music from vocals by generally training an accompaniment model from non-vocal segments [40]. These methods require a training phase on audio with labeled vocal/non vocal segments.The inability to find labeled data for vocal and non vocal segments separately had directed to discard that methodology.

Some methodologies had been implemented and tested for purpose of voice isolation in this research project. Some methods had been discarded and some had been enhanced and tested to see the effect.

4.2.2.1 Using the tool Audacity

Audacity is a freely available and and open-source digital audio editor and recording application software, available for Windows, macOS, Linux, and other Unix-like operating systems. In addition to recording audio from multiple sources, Audacity can be used for post-processing of all types of audio, including podcasts by adding effects such as normalization, trimming, and fading in and out. Audacity has also been used to record and mix entire albums, such as by Tune-Yards. It is also currently used in the UK OCR National Level 2 ICT course for the sound creation unit.

This tool can be used to isolate vocals in an audio, by using the Audacity's noise reduction feature. It comes in as a Nyquist plug-in. If the process is explained, a copy of the original stereo track is made. The noise profile of the copied track is captured. Using the noise profile, the instrumental sounds are reduced in the original track. The copied track can be deleted afterwards.

The following Figure 4.2 depict the spectrograms of the song by M.S. Fernando, "Aadarawanthi" before isolating vocals and after isolating vocals using Audacity.

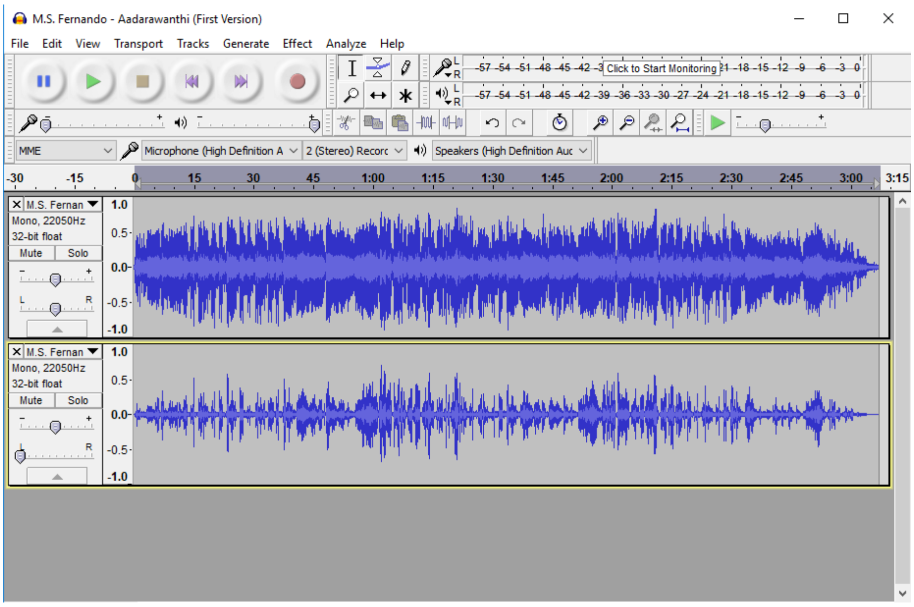


Figure 4.2: Isolation of vocals using Audacity

By examining the graphical representation in Figure 4.2 it is indisputable that this approach has affected the amplitude of the track drastically. The loudness of the resulting track had reduced to a level such that the features extracted from this track did not seem different from features extracted from other songs. Therefore this approach had to be discarded from use in this project.

4.2.2.2 Harmonic Percussive Source Separation

Separation of the audio signal into harmonic and percussive sources had been done using the python librosa library. In built function to separate sources can be found in librosa.decompose sub module. Librosa.decomposition is used to decompose a spectrogram to analyse its features and transformation. Librosa.decompose.hpss function is built in librosa with regard to the original median-filtering based approach of Fitzgerald, 2010 (Fitzgerald, 2010) and its margin-based extension due to Dreidger, Mueller and Disch, 2014 (Driedger et al., 2014).

The following code snippet is used to separate sources.

```
librosa.decompose.hpss(S, kernel_size=31, power=2.0, mask=False, mar
 gin=1.0)
```

- If margin = 1.0, decomposes an input spectrogram S = H + P where H contains the harmonic components, and P contains the percussive components.
- If margin > 1.0, decomposes an input spectrogram S = H + P + R where R contains residual components not included in H or P.

The parameter kernal size of this code snippet stands for the kernal size of the median filters. If the kernal size is a scalar, the same size is used for both harmonic and percussive.If else it is a tuple, the first value specifies the width of the harmonic filter, and the second value specifies the width of the percussive filter. Here in this implementation the value already in the method specified is used which is the kernal size of 31. As it is a scalar, the same size is used for both harmonic and percussive.

The source separation can be obtained by using the following procedure.

Load the audio by giving the pathname using the librosa package.

```
y, sr = librosa.load('audio/Ammawarune.mp3')
```

Compute the short time fourier transform (STFT) of y.

```
D = librosa.stft(y)
```

Decompose D into harmonic and percussive components.

$$D = D_{percussive} + D_{harmonic}$$

```
D_harmonic, D_percussive = librosa.decompose.hpss(D)
```

The specified code snippet would decompose the STFT into harmonic and percussive components. The following code snippet will generate the spectrograms of those two components along with the original component. The resulting graphs of this can be seen in Figure **??** in Results chapter.

```
rp = np.max(np.abs(D))
plt.figure(figsize=(12, 8))
```

```
 plt.subplot(3, 1, 1)
librosa.display.specshow(librosa.amplitude_to_db(D, ref=rp), y_axis='log')
 plt.colorbar()
 plt.title('Full spectrogram')

 plt.subplot(3, 1, 2)
librosa.display.specshow(librosa.amplitude_to_db(D_harmonic, ref=rp),y_ax
 is='log')
 plt.colorbar()
 plt.title('Harmonic spectrogram')

 plt.subplot(3, 1, 3)
librosa.display.specshow(librosa.amplitude_to_db(D_percussive, ref=rp),y_a
 xis='log', x_axis='time')
 plt.colorbar()
 plt.title('Percussive spectrogram')
 plt.tight_layout()
```

The harmonic partition (the assumed isolated vocals) will be transformed back to an audio snippet.

```
harmonic = librosa.istft(H)
librosa.output.write_wav('nadeeGangaharmonic.mp3',harmonic, sr)
```

Complete Codes will be annexed to the Appendix A.

4.2.2.3 Using Similarity Matrix

Vocal separation in this section is implemented by using librosa, numpy and matplotlib libraries. The song had been divided into frames of a 3 seconds. The frames which have the similar cosine similarity are aggregated and the amplitude of that precise frames is reduced. This is based on the “REPET-SIM” method, (Rafii and Pardo, 2011) but includes a couple of modifications and extensions including FFT windows overlap by a quater, instead of a half and non-local filtering is converted into a soft mask by Wiener filtering. This is similar in spirit to the soft-masking method (Fitzgerald and Jaiswal, 2012), but is a bit more numerically stable in practice.

The implementation of this filter will be discussed subsequently.

Load the audio.

```
y, sr = librosa.load('audio/Ammawarune.mp3')
```

Compute the spectrogram magnitude and the phase.

```
S_full, phase = librosa.magphase(librosa.stft(y))
```

A five second slice of the spectrum is plotted below, in the Figure 4.3 for a song.

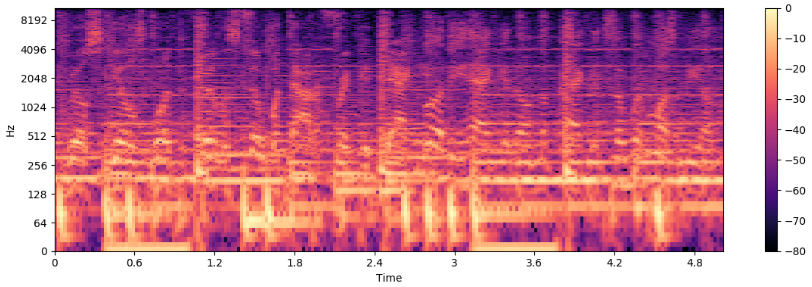


Figure 4.3: Spectrogram of original song

The dizzy lines above depict the vocal component. The goal had been to separate them from the accompanying instrumentation. The frames are compared using co-sine similarity, and similar frames are aggregated by taking their (per-frequency) median value. To avoid being biased by local continuity, a constraint is applied for similar frames to be separated by at least 2 seconds. This suppresses sparse/non-repetitive deviations from the average spectrum, and works well to discard vocal elements. The following code section implements the discussed filter.

```
S_filter = librosa.decompose.nn_filter(S_full,
                                   aggregate=np.median,
                                   metric='cosine',
                                  width=int(librosa.time_to_frames
                                   (2, sr=sr)))
```

The output of the filter shouldn't be greater than the input when the assumption is made that signals are additive. Taking the point wise minimium with the input spectrum forces this.

```
S_filter = np.minimum(S_full, S_filter)
```

The raw filter output can now be used as a mask, but to make the sound better soft-masking needs to be used. A margin is also used to reduce bleed between the vocals and instrumentation masks.

```
margin_i, margin_v = 2, 10
power = 2

mask_i = librosa.util.softmask(S_filter,
                               margin_i * (S_full - S_filter),
                               power=power)

mask_v = librosa.util.softmask(S_full - S_filter,
                               margin_v * S_filter,
                               power=power)
```

After obtaining the masks, they are to be multiplied with the input spectrum to separate the components.

```
S_foreground = mask_v * S_full
S_background = mask_i * S_full
```

The foreground(voice) and the background(instrumental music) can be separated when the defined procedure is followed. The complete code snippet will be annexed to Appendix A.

Finally the foreground is converted into an mp3 format audio.

```
D_foreground = S_foreground * phase
y_foreground = librosa.istft(D_foreground)
librosa.output.write_wav('final.wav', y_foreground, sr)
```

4.2.2.4 Using Band-pass Filter

The Butterworth band-pass filter is implemented using the Scipython library. Scipy signal sub module is considered the signal processing toolbox currently contains some filtering functions, a limited set of filter design tools, and a few B-spline interpolation algorithms for uni dimensional and 2 dimensional data. B-spline algorithms are usually placed in the interpolation category, they are included as a filtering function as they only work with data which are equally spaced. Furthermore they have made use of filter theory and transfer-function formalism to make the transform faster.A signal in SciPy is an array of real or complex numbers.While scipy.signal.freqz is used to compute the frequency response, the scipy.signal.lfilter is used to apply the filter to a signal.

Here, a function was written in order to apply the butterworth bandpass filter to a specific audio.

```
def butter_bandpass(lowcut, highcut, fs, order=5):
    nyq = 0.5 * fs
    low = lowcut / nyq
    high = highcut / nyq
    b, a = butter(order, [low, high], btype='band')
    return b, a

def butter_bandpass_filter(data, lowcut, highcut, fs, order=5):
    b, a = butter_bandpass(lowcut, highcut, fs, order=order)
    y = lfilter(b, a, data)
    return y
```

Here, the filter order of order=5 is given to the audio which attenuated the signal sequentially starting from -6db per octave and second order at -12db per octave and likewise. The reason behind the order being higher is the necessity to lower the frequencies which do not fit in to the range as much as possible.

4.2.2.5 Elimination of Silence

As cited and discussed in the design chapter, introduction of a song is the partition where the strongest music is present. Therefore, the first path which was followed had been eliminating a small time period from the start of the song. To show the considerable amount of time allocated for the introduction of a song, a survey analysis had been taken place. In order to remove the introductory part of the song, two approaches had been considered.

- Take an average number of seconds allocated for introductory parts of all songs of the dataset and remove that number of seconds from every song.
- Take the highest number of seconds which had been allocated to the introduction of a song and reduce that number of seconds from every song.

Both these approaches had issues. The average number of seconds was 14 seconds, and the longest time duration was 47 seconds. They were not compatible with each other. After vocal isolation it seemed that the resulting audios did not actually contain music in the introduction but silence mostly. After the voice isolation that time frames had been nothing but silence. Therefore, the approach had been changed to remove the silence from the audios. Removal of silence in audio processing tasks had been considered as a preprocessing task. Scientists had always sided up with removal of silence and noise in many audio signal processing researches as they had believed that would increase the accuracy of the final result.[100]

Implementation of this functionality had been done using Audiosegment in pydub module. Starting signal of an energy lesser than -50db had been considered as silence and the audio is trimmed until anything larger than -50db is heard. That is done reversed as well, which ends up removing silence of the ending part of the song. Frames are iterated until a frame with a sound is found.

```
def detect_leading_silence(sound, silence_threshold=-50.0,
chunk_size=10):
    trim_ms = 0
    while sound[trim_ms:trim_ms+chunk_size].dBFS < silence_thresho
    ld and trim_ms < len(sound):
        trim_ms += chunk_size

    return trim_ms
```

In the resulting audios, the vocals were extracted starting from the first frame and nevertheless in the last frame.

4.2.3 Feature Extraction

Python librosa had been used for analyzing and extracting features of an audio signal. The features extracted in this project were, Zero Crossings Rate, Spectral Centroid, Spectral Rolloff and MFCC.

A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. It's a representation of frequencies changing with respect to time for given music signals. The following code snippet will display the spectrogram with its transformation.

```
X = librosa.stft(x)
Xdb = librosa.amplitude_to_db(abs(X))
plt.figure(figsize=(14, 5))
librosa.display.specshow(Xdb, sr=sr, x_axis='time', y_axis='hz')
```

.stft converts data into short term Fourier transform. STFT converts signal such that we can know the amplitude of given frequency at a given time. Using STFT we can determine the amplitude of various frequencies playing at a given time of an audio signal. .specshow is used to display spectogram.

Zero crossings can also be represented or counted using code. There can be found an inbuilt function in librosa to get the zero crossings of a signal.

```
zero_crossings = librosa.zero_crossings(x[n0:n1], pad=False)
print(sum(zero_crossings))
```

spectral-centroid is used to calculate the spectral centroid for each frame. So it would return an array with columns equal to a number of frames present in the sample.The following code snippet has represented the spectral centroid of a given audio. The resulting graphical representations are represented in the results and the evaluation section.

```
spectral_centroids = librosa.feature.spectral_centroid(x, sr=sr)[0]
spectral_centroids.shape
```

To compute the time variable for visualization,

```
frames = range(len(spectral_centroids))
t = librosa.frames_to_time(frames)
```

To normalize the spectral centroid for visualization,

```
def normalize(x, axis=0):
return sklearn.preprocessing.minmax_scale(x, axis=axis)
```

Spectral Rolloff would give results to each frame.

```
spectral_rolloff = librosa.feature.spectral_rolloff(x, sr=sr)[0]
```

MFCCs can be extracted using code as well. They can be graphically represented as depicted in in Results section.

```
mfccs = librosa.feature.mfcc(x, sr=sr)
```

All these features are extracted and stored in a CSV(Comma Separated Values) file for model generation.

4.2.4 Signature Generation

Python's sklearn.mixture package is used in this research project to learn a GMM from the features matrix containing the defined features in audio.The following Python code is used to train the GMM speaker models (signatures). The code is run once for each artist and there exists a text filename containing path to all the audios for the respective singer. Also, a new directory is created where all the models will be dumped after training.

The generated signatures were enclosed in a signature dump. Python cPickle library was used in the implementation of the dump.

```
for artist in artists:
gmm = GaussianMixture(n_components = 4, max_iter = 200, covariance
_type='diag',n_init = 3)
gmm.fit(artist)
picklefile = str(t)+”.gmm”
pickle.dump(gmm,open(dest + picklefile,'wb'))
print (' modeling completed for speaker:'+str(t))
t = t+1
```

For evaluation, the winning signature is selected by using the argmax function from the signature dump after checking all models.

```
for i in range(len(models)):
gmm    = models[i]
scores = np.array(gmm.score(new))
log_likelihood[i] = scores.sum()

winner = np.argmax(log_likelihood)
print ("\tdetected as - ", singers[winner])
```

This section has discussed the implementation details of this research project including the discussions on technology used, implementations of functionalities and issues and limitations which had to be dealt with, when proceeding. Some main code snippets for the implementation had been shown in the chapter where all the codes will be attached to the Appendix A. Each step to perform all filters separately had been discussed with and through argument and analysis.

Chapter 5

Results and Evaluation

As mentioned under Section 3.3, evaluation will consist of a quantitative evaluation and a qualitative artist identification evaluation. The quantitative evaluation will further be broken down into evaluation categories where pre-processing steps are used separately and combined which consists of four cases. In the quantitative analysis, there is another evaluation design to calculate the effect of complementary preprossessing step. The qualitative evaluation will be based on different paired combinations of singers. They start from male and female voice distinction to father,son voice distinction. This chapter details the entire evaluation process including the rationale and assumptions behind the methods of evaluation. The evaluation results of both the quantitative and qualitative evaluations are summarized and analyzed under this chapter.

Furthermore, the results obtained from every task will be depicted and analysed in this chapter.

5.1 Results

In this section the results obtained from voice isolation, feature extraction and signature generation are outlined. The observations which were made, and the decisions made by observations are detailed further.

5.1.1 Results of Voice Isolation

There had been mainly three approaches used for voice isolation namely,

- REPET
- Haronic Percussive Source Separation
- Band-pass filter

The results of voice isolation from these three filters will be analysed separately hereinafter.

5.1.1.1 Results of REPET

From looking at the spectrogram of the original song, one may be incapable of identifying what is voice and what is instrumental sounds. When using REPET, the spectrogram of the song is divided into the foreground and background where foreground is the isolated voice of the song and background is the remaining music of the song. The Figure 5.1 depicts how the voice is separated from the instrumental music.

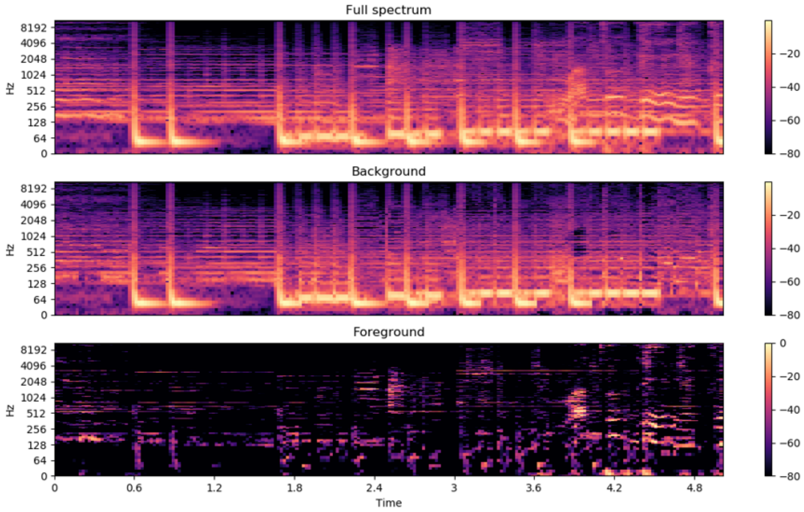


Figure 5.1: Isolation of vocals using REPET

The remaining dizzy segments of the spectrogram are the vocals as assumed by the REPET method. This had been made the base method to isolate vocals in this research due to, the considerable amount of instrumentation music components which were reduced by using this approach.

5.1.1.2 Results of Harmonic Percussive source separation

This filter had used the argument that vertical components of the spectrogram imply the percussive sounds while horizontal components of the spectrogram imply harmonic sounds. Here in the Figure 5.2 below represent how the source separation had taken place.

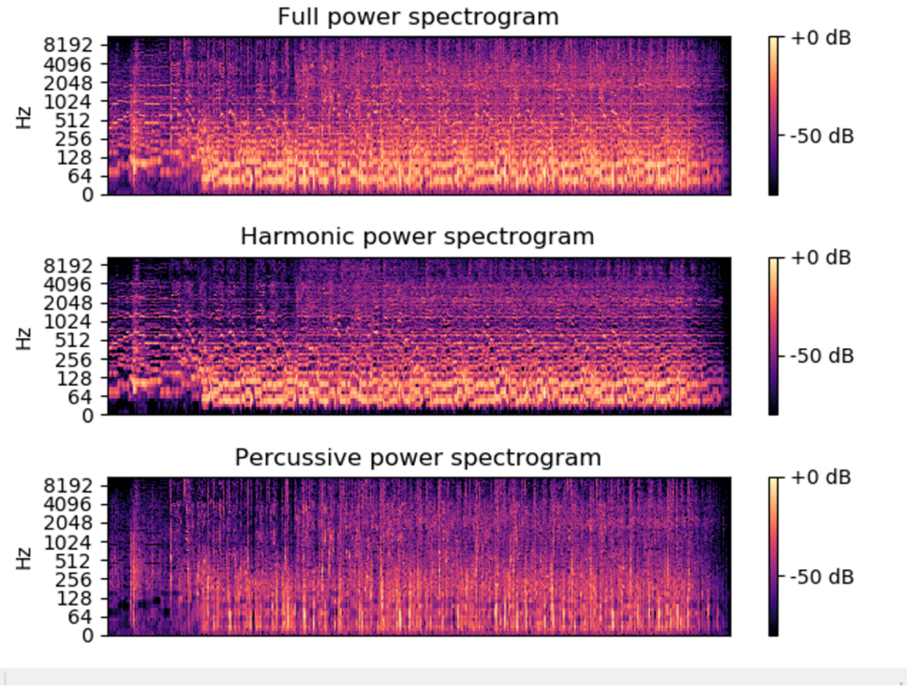


Figure 5.2: Harmonic Percussive source separation

This filter resulted in melodious sounds like voice, guitars and pianos in the harmonic partition whereas the percussive sounds like drums, symbols in the percussive partition.

5.1.1.3 Results of applying band-pass filter

The butterworth band-pass filter had removed non vocal frequency range from the audio preserving the vocal frequency range from the song. The following Figure 5.3 depicts how the top and bottom frequencies had been removed from the audio.

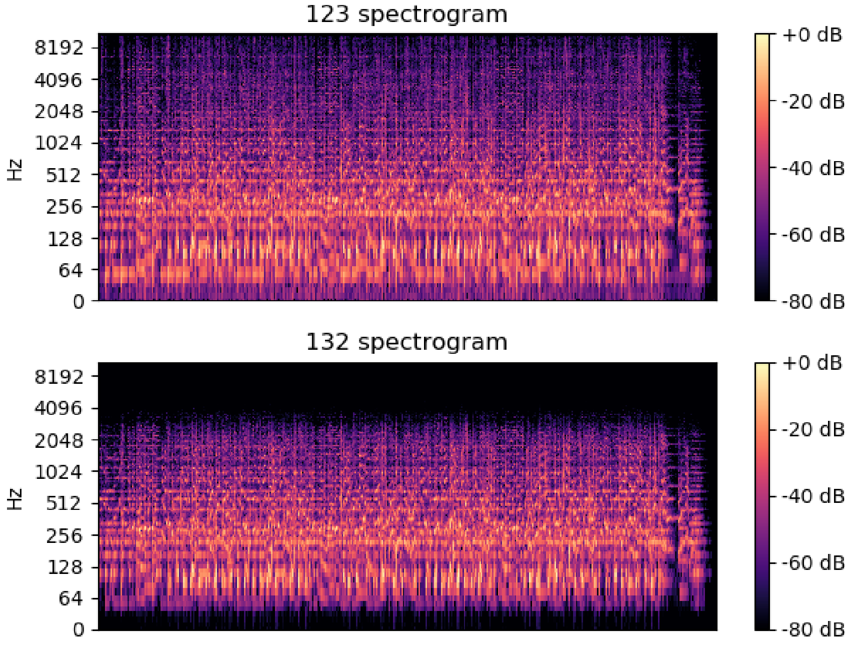
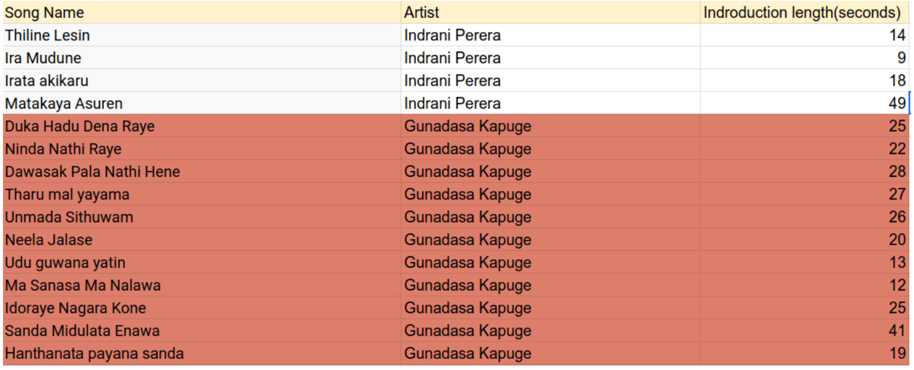


Figure 5.3: Band-pass filter

Voice isolation had been done using these three filters mainly and different combinations had been experimented in order to find the best combination of filters to isolate vocals. Those findings will be discussed in the Evaluations section.

5.1.1.4 Elimination of introductory and ending parts

There had been a complementary step to experiment the effect of removal of introductory partitions for the voice separation process. In order to observe the considerable time duration a song takes for its introduction, a survey analysis had been performed. The following Figure 5.4 shows how the evaluation had taken place.



Song Name	Artist	Indroduction length(seconds)
Thiline Lesin	Indrani Perera	14
Ira Mudune	Indrani Perera	9
Irata akikaru	Indrani Perera	18
Matakaya Asuren	Indrani Perera	49
Duka Hadu Dena Raye	Gunadasa Kapuge	25
Ninda Nathi Raye	Gunadasa Kapuge	22
Dawasak Pala Nathi Hene	Gunadasa Kapuge	28
Tharu mal yayama	Gunadasa Kapuge	27
Unmada Sithuwam	Gunadasa Kapuge	26
Neela Jalase	Gunadasa Kapuge	20
Udu guwana yatin	Gunadasa Kapuge	13
Ma Sanasa Ma Nalawa	Gunadasa Kapuge	12
Idoraye Nagara Kone	Gunadasa Kapuge	25
Sanda Midulata Enawa	Gunadasa Kapuge	41
Hanthanata payana sanda	Gunadasa Kapuge	19

Figure 5.4: Survey analysis for intro removal

5.1.2 Results of Feature Extraction

The features extracted for this project as described in the Section 3 are, zero crossings rate, spectral centroid, spectral rolloff and mfcc features. The resulting spectrograms of those features are discussed in this sub section.

The following Figure 5.5 depicts the zero crossings rate of the song "Sihina ahase ananthe " by the artist Greshan Ananda. The zero crossings rate is the number of times the signal changes from positive to negative. This plot depicts the sound signal of the song "Sihina Ahase" which has 11 zero crossings.

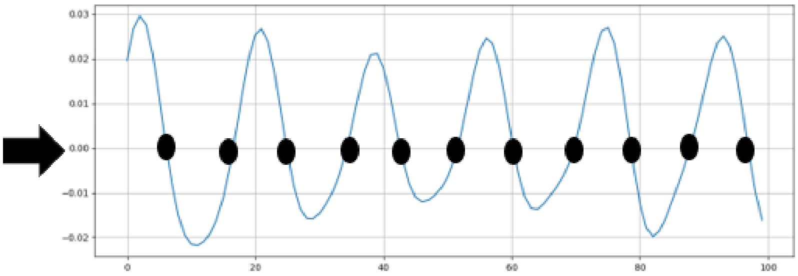


Figure 5.5: Zero crossings rate of the song "Sihina Ahase"

The spectral centroid of a song is usually the fundamental frequency of a song. The fundamental frequency of a song lies on the vocal track of the song. It can be considered as the center of mass of a sound. The Figure 5.6 below represents how the centre of mass had fallen on the song "Sihna Ahase".

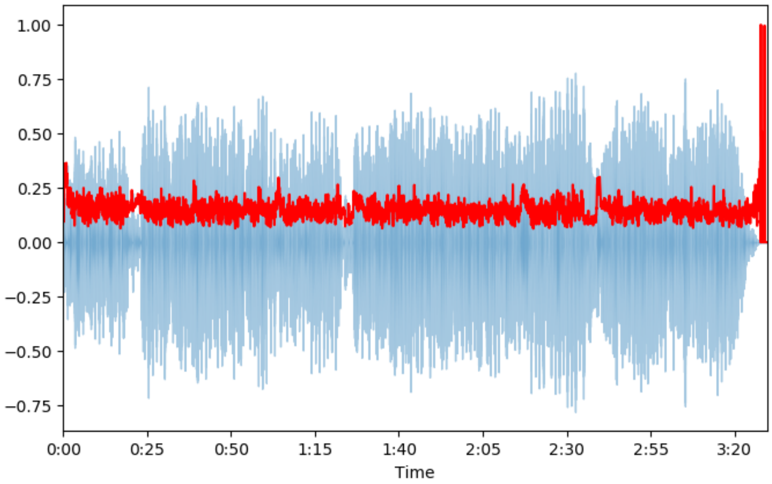


Figure 5.6: Spectral centroid for the song "Sihina Ahase"

Here the spectral centroid has fallen on the track as depicted by the red line. As explained earlier this centroid represents features of the voice of the singer. The blue partition represents the original song. The centroid had risen up towards the end of the song as the voice is slowly faded and now the centroid is occupied by the instrumental music.

Spectral rolloff of an audio shows the frequency below which a specified percentage of the total spectral energy lies. This is important because the energy of a song is amplified by the singer. Vocals possess a larger percentage of the overall energy of a song. Therefore it is considered as another important feature of voice. Figure 5.7 represents how spectral rolloff of the song ”Sihina Ahase” is spread.

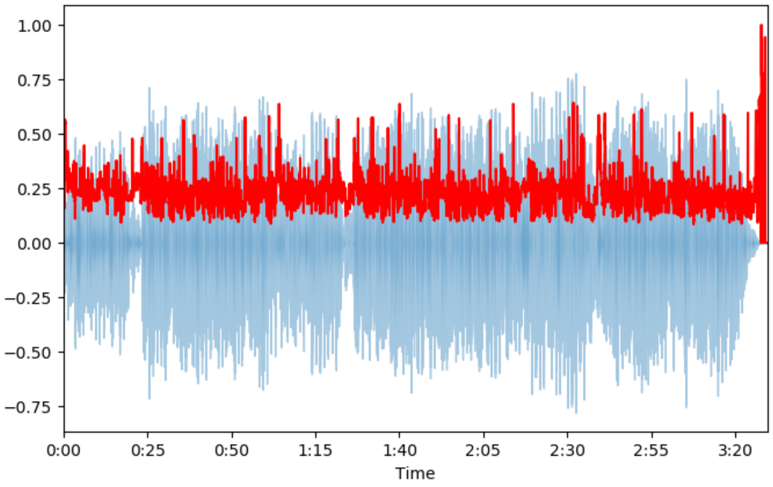


Figure 5.7: Spectral rolloff for the song ”Sihina Ahase”

The spectral rolloff of the audio is depicted above by the red lines and the original audio signal by the blue lines. It can be seen that the spectral centroid of the song mostly gathers up in the higher frequency sections, which means the singer has sung this song amplifying his voice in the higher frequency notes. Or either the singer’s voice can be in a higher frequency range.

Mel frequency spectral coefficients are considered as a feature vector which describes the human voice the best as discussed in the chapter 3.2.2 . The following Figure 5.8 is the graphical representation of the MFCC features. The 20 features here describe the overall shape of the spectral envelope.

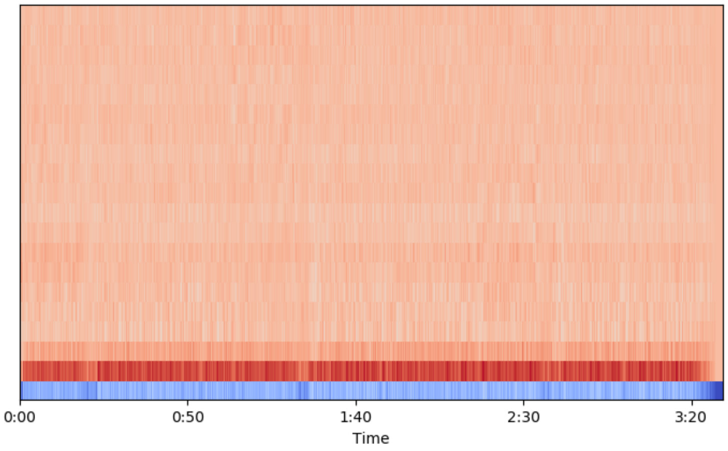


Figure 5.8: MFCC for the song ”Sihina Ahase”

5.2 Evaluation

Generating digital signatures for singers using their songs is considered as a novel research up to date. The evaluation of this research had been done using both quantitative and qualitative methodologies.

5.2.1 Quantitative approach

There have been mainly three approaches used in the voice isolation process namely,

- REPET
- Harmonic Percussive source separation
- Band-pass filter

The quantitative evaluation has analysed the effect of these pre processing steps separately and combined, for the signature and has chosen the best combination to produce signatures for singers finally.

REPET alone has the capability to isolate vocals by itself while the other two filters do not. Harmonic, percussive source separation separates harmonic and percussive sources, not vocal and instrumental sources. Therefore the harmonic source includes voice with other music from harmonic instruments. Therefore when features are extracted from the harmonic part separately, features of the sounds of those music instruments may be extracted as well. Therefore harmonic percussive source separation method could not be used as a direct approach to separate vocals and non vocal segments of a song. Therefore, evaluating that approach alone is erroneous. Usage of the band-pass filter alone too would preserve the frequency range 85 to 855 Hz and that might not only be vocals, it can be the instrumental music as well. Therefore, using band-pass filter alone and evaluating has also been discarded.

Therefore, the combinations which were evaluated were,

- REPET alone
- REPET + Harmonic Percussive source separation
- REPET + Band-pass filter
- REPET + Harmonic percussive separation + Band-pass filter

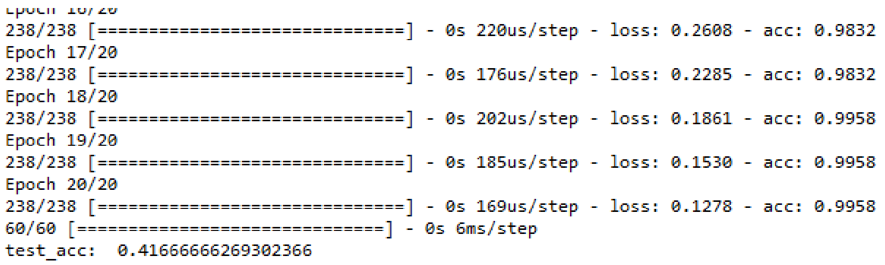
The reasons for evaluating the harmonic percussive filter and band pass filter separately with REPET is to examine whether there is any effect when they are combined.

5.2.1.1 Training and testing data

The dataset consisted of approximately 600 songs of both female and male artists. The training and testing data were split from this dataset, training set having 70% of the data while test set having 30% of the data. The data had been randomly chosen to be included in test or training sets.

5.2.1.2 REPET alone

The dataset for this evaluation criteria has been trained only using the REPET filter. GMM models are generated for every singer and the accuracy is obtained as the percentage of test data which were correctly identified the label of the model as the appropriate singer. The accuracy of the signature generation when using REPET alone had been 0.4166667 as shown in the Figure 5.9.

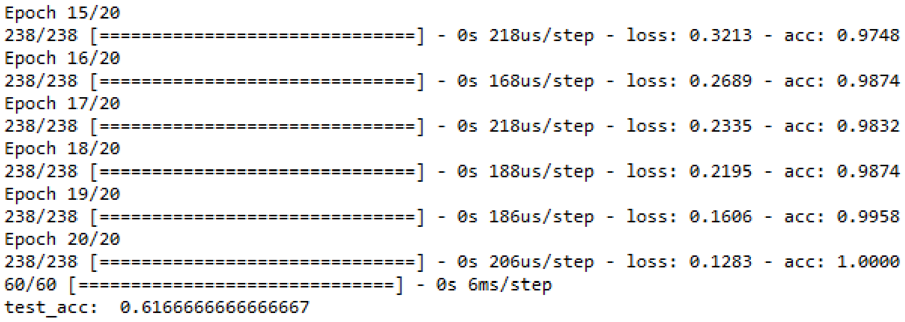


```
Epoch 16/20
238/238 [==============================] - 0s 220us/step - loss: 0.2608 - acc: 0.9832
Epoch 17/20
238/238 [==============================] - 0s 176us/step - loss: 0.2285 - acc: 0.9832
Epoch 18/20
238/238 [==============================] - 0s 202us/step - loss: 0.1861 - acc: 0.9958
Epoch 19/20
238/238 [==============================] - 0s 185us/step - loss: 0.1530 - acc: 0.9958
Epoch 20/20
238/238 [==============================] - 0s 169us/step - loss: 0.1278 - acc: 0.9958
60/60 [==============================] - 0s 6ms/step
test_acc:  0.41666666269302366
```

Figure 5.9: Accuracy of using REPET alone

5.2.1.3 REPET + Harmonic Percussive source separation

Here, the data set had been preprocessed using both harmonic percussive source separation and REPET together. The accuracy of signature generation had been 0.61111 as shown in the below Figure 5.10.



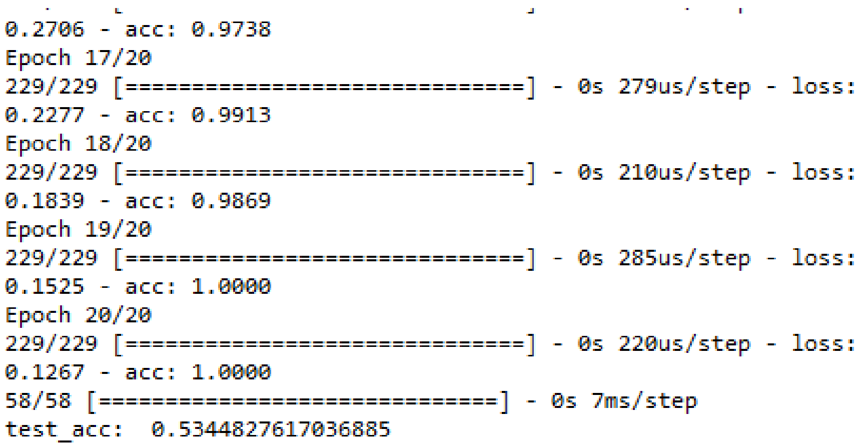
```
Epoch 15/20
238/238 [==============================] - 0s 218us/step - loss: 0.3213 - acc: 0.9748
Epoch 16/20
238/238 [==============================] - 0s 168us/step - loss: 0.2689 - acc: 0.9874
Epoch 17/20
238/238 [==============================] - 0s 218us/step - loss: 0.2335 - acc: 0.9832
Epoch 18/20
238/238 [==============================] - 0s 188us/step - loss: 0.2195 - acc: 0.9874
Epoch 19/20
238/238 [==============================] - 0s 186us/step - loss: 0.1606 - acc: 0.9958
Epoch 20/20
238/238 [==============================] - 0s 206us/step - loss: 0.1283 - acc: 1.0000
60/60 [==============================] - 0s 6ms/step
test_acc:  0.6166666666666667
```

Figure 5.10: Accuracy of using REPET + Harmonic percussive separation

This had shown that the performance of voice isolation for the task of generating digital signatures for singers had been improved by using the harmonic percussive source separation significantly.

5.2.1.4 REPET + Band-pass filter

Both REPET and the butterworth band-pass filter had been used in the preprocessing stage of this evaluation. The accuracy obtained as shown in the Figure 5.11 below is 0.53448.



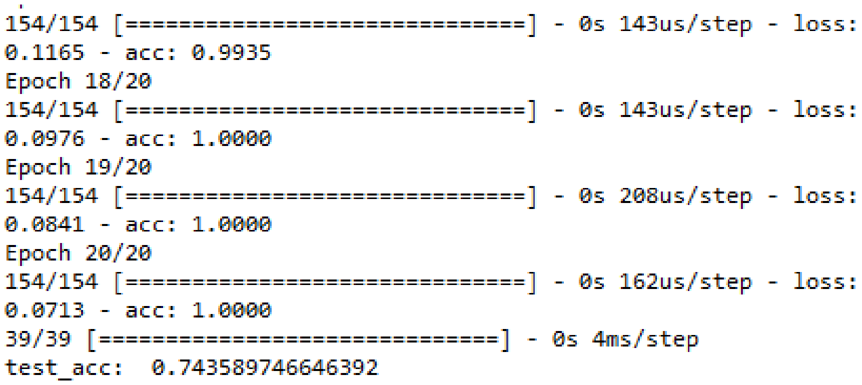
```
0.2706 - acc: 0.9738
Epoch 17/20
229/229 [==============================] - 0s 279us/step - loss:
0.2277 - acc: 0.9913
Epoch 18/20
229/229 [==============================] - 0s 210us/step - loss:
0.1839 - acc: 0.9869
Epoch 19/20
229/229 [==============================] - 0s 285us/step - loss:
0.1525 - acc: 1.0000
Epoch 20/20
229/229 [==============================] - 0s 220us/step - loss:
0.1267 - acc: 1.0000
58/58 [==============================] - 0s 7ms/step
test_acc:  0.5344827617036885
```

Figure 5.11: Accuracy of using REPET + Harmonic percussive separation

Even the performance of this filter is better than using REPET alone, it is less than using REPET with harmonic percussive source separation.

5.2.1.5 REPET + Harmonic percussive separation + Band-pass filter

All three filters had been combined in this experiment. All the filters had been performed on the songs in training and test datasets. The result had been the best as of then, resulting in 0.743589 accuracy. The result is shown below in the Figure 5.12.



```
154/154 [==============================] - 0s 143us/step - loss:
0.1165 - acc: 0.9935
Epoch 18/20
154/154 [==============================] - 0s 143us/step - loss:
0.0976 - acc: 1.0000
Epoch 19/20
154/154 [==============================] - 0s 208us/step - loss:
0.0841 - acc: 1.0000
Epoch 20/20
154/154 [==============================] - 0s 162us/step - loss:
0.0713 - acc: 1.0000
39/39 [==============================] - 0s 4ms/step
test_acc:  0.743589746646392
```

Figure 5.12: Accuracy of using REPET + Harmonic percussive separation + Band-pass filter

5.2.1.6 Discussion of combined approaches

These results have been represented as rounded off percentages in the following Table 5.1.

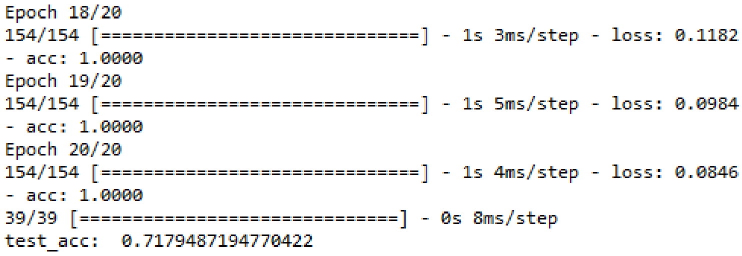
Table 5.1: Accuracies of the filter combinations

#	Method	Accuracy
1	REPET	42%
2	REPET + HP	62%
3	REPET + BP	54%
4	REPET + HP + BP	74%

These results show that the best approach for voice isolation is the combination of all three REPET, Harmonic percussive source separation and using the Band-pass filter. These percentages show some interesting theories like, harmonic percussive source separation being more convenient than the band-pass filter for this particular research question. These findings will further be addressed in the Chapter 6.

5.2.1.7 Effect of silence removal

As discussed the chapter 3, the winning approach is again evaluated against the winning approach combined with silence removal. Here the winning approach is the combination of all three filters.Therefore, after applying all REPET, HP and BP filters, silence of introduction and ending are removed and evaluated. The resulted accuracy of this combination was 0.717948 as depicted in the Figure 5.13 below.



```
Epoch 18/20
154/154 [==============================] - 1s 3ms/step - loss: 0.1182
- acc: 1.0000
Epoch 19/20
154/154 [==============================] - 1s 5ms/step - loss: 0.0984
- acc: 1.0000
Epoch 20/20
154/154 [==============================] - 1s 4ms/step - loss: 0.0846
- acc: 1.0000
39/39 [==============================] - 0s 8ms/step
test_acc:  0.7179487194770422
```

Figure 5.13: Accuracy of using silence removal with winner

The silence removal had not been fruitful in voice isolation as it has decreased the performance of the winning approach. Why this has happened can be because even though the silence had been removed from the start and the end, still there are remaining silence chunks in the middle of the songs due to the interludes.

The quantitative evaluation of this research project had been concluded with declaring the best approach to generate the digital signature by using REPET, Harmonic and percussive source separation and the band-pass filter.

5.2.2 Qualitative approach

The evaluation, is performed using a pairwise evaluation. Several pairs of singers have been classified according to some defined relationships. Signatures for both the singers are generated and when given a song of one artist, and examined if the singer is identified correctly when given a song. The reason for pairwise evaluation is because this research focuses on reducing counterfeiting. In counterfeiting a song of a singer is sung by another person. Therefore, the signature generated by the song will be compared with only the real singer's signature. Hence, a binary evaluation can be considered as the mostly suited evaluation approach for this project.

The pairs which had been made into consideration of this research are,

1. Gender classification - Can the signatures of male and female singers be identified correctly?

2. Classification of male singers - Can the signatures of different male singers identified properly?

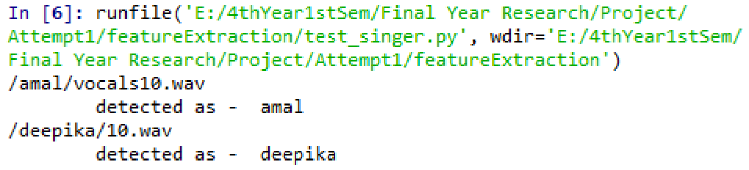
3. Classification of female singers - Can the signatures of different female singers be identified correctly?

4. Classification of father son combinations - Can the signatures of a father and a son be identified properly?

5. Classification of vocals of siblings - Can the signatures of sisters or brothers be identified properly?

5.2.2.1 Gender classification

An example for a pair of singers whose signatures were evaluated is Amal Perera with Deepika Priyadarshani. They had generated two different signatures and whenever a test song was provided, they were accurately identified as Amal Perera's or Deepika Priyadarshani's. The following Figure 5.14 shows how the results were generated.



```
In [6]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/amal/vocals10.wav
        detected as -  amal
/deepika/10.wav
        detected as -  deepika
```

Figure 5.14: Artist signature identification

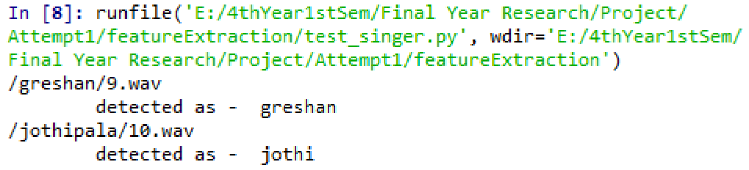
This evaluation had been done using other artists like, H. R. Jothipala, Nelu Adhikari, Jagath Wickramasinghe, Anjaline Gunathilaka. All the voices were correctly identified as female and male when given a pairwise combination.

5.2.2.2 Classification of male singers

For this evaluation, singers whose voices are considered similar in Sri Lanka were chosen.

- Greshan Ananda and H.R. Jothipala
- Dayan Witharana and Jagath Wickramasinghe

Both the evaluations were successful. In both situations, test song was correctly classified. This observation has depicted the potential of these signatures to be robust and rigid even under similar features. The following Figure 5.15 depicts an instance where Greshan Ananda was identified properly from H.R. Jothipala and himself.



```
In [8]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/greshan/9.wav
        detected as -  greshan
/jothipala/10.wav
        detected as -  jothi
```

Figure 5.15: Artist signature identification of male similar voiced singers

5.2.2.3 Classification of female singers

This evaluation had also been done using female artist pairs who are said to have similar voices. The pairs which were considered in this project were,

- Neela Wickramasinghe and Nelu Adikari
- Anjaline Gunathilake and Latha Walpola

Both these evaluations had given positive results. All singers were correctly identified as themselves when given test songs even though they sound similar.

5.2.2.4 Classification of father-son combinations

It is a fact that in Sri Lanka there are some father-son pairs who sound very similar. Therefore this signature generation approach had to be tested on them as well. The father-son pairs which were considered in this evaluation were,

- Milton and Ranil Mallawaarachchi
- Mervin and Amal Perera

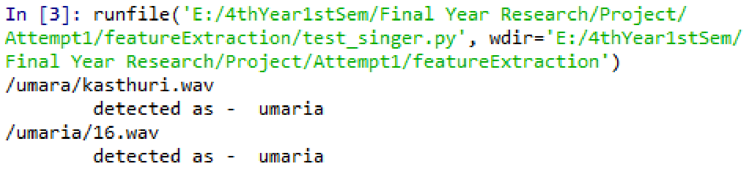
The results of this evaluation were again accurately identified as the father or the son exhibiting the robustness of the signature. The signature generated shows the ability of being vulnerable to sensitive features of the vocalist.

5.2.2.5 Classification of siblings

The vocals which were compared for this evaluation was,

- Umara and Umaria Sinhawansa

Both these singers have similar voices, and the signatures were unable to classify Umara from Umaria. Songs of Umara Sinhawansa were classified as Umaria Sinhawansa's while Umaria Sinhawansa's voice was correctly identified as Umaria Sinhawansa's. The inability of this approach to identify these two singers can not be answered directly but through some more experiments. It can be assumed that as these two are female singers and the voices are higher in frequency range, there might be issues when extracting features from them. The following Figure 5.16 shows how the two singers were incorrectly classified.



```
In [3]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/umara/kasthuri.wav
        detected as -  umaria
/umaria/16.wav
        detected as -  umaria
```

Figure 5.16: Artist signature identification of siblings

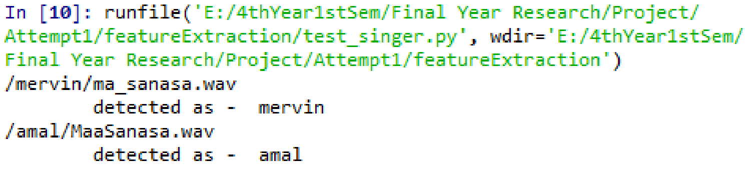
5.2.2.6 Further Evaluation

The evaluation of this research had been done in two other routines as well. The evaluation must satisfy the following two conditions according to the evaluation plan described in the Chapter 3.

1. Same song sung by two different singers must generate two different signatures.

2. Different songs sung by the same singer must generate the same digital signature.

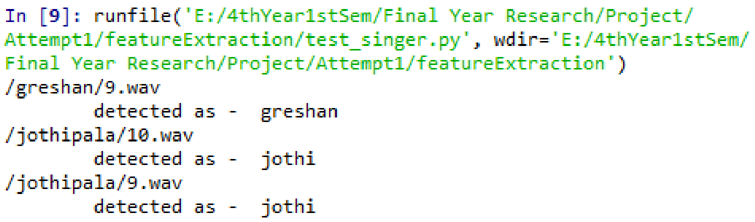
The first condition had been tested by using the song " Ma sanasa " sung by both Mervin Perera and Amal Perera as the test song. They had specifically resulted in identifying the voices accurately as Mervin's and Amal's regardless of the test song being the same song. The following Figure 5.17 depicts how they were identified.



```
In [10]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/mervin/ma_sanasa.wav
        detected as -  mervin
/amal/MaaSanasa.wav
        detected as -  amal
```

Figure 5.17: Artist signature identification of same song - different singers

The second condition had been tested by using two different songs of H.R. Jothipala. Both the songs were classified as H.R. Jothipala's songs discarding other signatures. The following Figure 5.18 depicts how they were identified.



```
In [9]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/greshan/9.wav
        detected as -  greshan
/jothipala/10.wav
        detected as -  jothi
/jothipala/9.wav
        detected as -  jothi
```

Figure 5.18: Artist signature identification of same singer- different singers

5.2.2.7 Discussion of qualitative approach

The pairwise evaluation had been successful in almost all the cases. It had shown just one negative result. A brief analysis is shown in the Table 5.2 below, depicting how the evaluation had taken place.

Table 5.2: Qualitative Evaluation

#	Combination	Result
1	Gender classification	✓
2	Male singer classification	✓
3	Female singer classification	✓
4	Father son classification	✓
5	Sibling classification	✗
6	Same song - different singers	✓
7	Same singer - different songs	✓

Evaluation of the research had been successful. It is safe to say that the generating signature is done using very specific features of voices. Identification of father and son can be considered as a valuable result generated from the research as they have very similar vocals.

This section has addressed the results generated in all phases of the research, and it also has included the how the evaluation of the research had taken place.

Chapter 6

Conclusion

6.1 Introduction

This chapter focuses on the conclusions drawn upon the completion of the research. The research aim stated in Section 1.2.2 has been accomplished by using the technologies and tools that are mentioned in Section 4.1. It is possible to generate digital signatures for singers using their songs.The subsequent sections in this chapter will further discuss the conclusions of this research.

6.2 Conclusions about research questions

- RQ1 : How can Audio signal processing and music information retrieval be used to distinguish between voice extracted from songs?

Audio signal processing and music information retrieval has been used in this research to isolate vocals in a song, extract features from an audio signal and generate unique models(signatures) for each artist. The process of isolating vocals had been made the pre-processing step of this research, which had been tested using three main techniques separated and combined. The best technique for voice separation had been selected out of those combinations after evaluation which is using REPET method, Harmonic Percussive source separation and Band-pass filter. This technique had been used for the final evaluation of generating a precise digital signature for singers. These signatures had been compared with other signatures to examine the sensitivity to capture specific features of vocalists. In all except one case, the singer had been identified correctly. It is safe to declare this research successful as it had resulted in an accuracy closer to 75%. Singers who tend to have similar voices like Milton and Ranil Mallawaarachchi , Neela Wickramasinghe with Nelu Adikari had shown the sensitivity of this generated signature.

In this research,the best voice isolation technique had been declared as the combination of REPET with Band-pass Filter and Harmonic Percussive source separation. It had also pointed out some interesting theories like harmonic percussive source separation is more favourable than using a band pass filter because it gives a better percentage accuracy when combined with REPET separately. It had also depicted that voices of singers who have similar voices can be separated using these signatures. The signatures also had shown the consistency by identifying different songs of the same singer and when the same song is sung by different singers. The silence removal had not been effective as it had shown a less percentage accuracy than the combination of REPET, Harmonic Percussive source separation and band-pass filter. That can be due to the remaining silenced partitions in the song after music removal in interludes.

- RQ2 : What gains can be made using proposing method over usual voice recognition methods used?

This research is a novel research. The literature review has specified voice isolation and singer identification as two different research questions. The proposing voice isolation method is a combination of a usual voice isolation method REPET and audio analysis methods. The gains of the proposing voice isolation methodology had been compared in the Chapter 5.2.1. It had shown a better accuracy than the usual REPET. Signaure generation had been done using GMM models rather than using HMM models. An evaluation of GMM model over HMM model is not discussed in this research but can be performed as future work.

- RQ3 : Would this proposing signature meet the other requirements? (can be easily stored, indexed and compared)

This signature can be easily stored as it is a GMM model of 22KB. For every singer this signature can be created using this approach. The comparing process is easily done using the cpickle library of python.

6.3 Conclusion about the research problem

This research has given a solution for the imitation of voices of popular artists and creating counterfeit audio tracks using imitated voice. It has shown how singers can be recognized as the original and the counterfeiting singer, using the way the qualitative evaluation had been taken place. The qualitative evaluation has represented how the similar voices are distinguished accurately. Therefore this can be adapted in order to find the real singer in any case. In conclusion, this study has yielded a productive solution for the problem of counterfeiting of songs in Sri Lanka. This solution in this study has shown promise of being accurate for the target artists.

6.4 Implications for further research

This research project accuracy can further be increased by experiment. The preprocessing steps used in this research had been a combination of three signal processing techniques. This approach can be modified, and be experimented against the proposed method. If voice isolation becomes more successful, it might grant a better final result.

It is a fact that the Sri Lankan Sinhala song has strings with Indian Hindi songs, where in fact Sinhala songs with the exact melody of Hindi songs can still be heard in Sri Lanka. Therefore, it can be said with confidence that this reseach can be extended using Indian songs.

The model used as the signature in this research is a GMM model. This model can be changed and the results be compared with the results of the proposing method which might give out a better percentage accuracy than 74%. Likewise, this research can be performed further until a better result is achieved.

References

Abubakar Sadiq, A., Othman, N. and Abdul Jamil, M. M. (2018), 'Fourth-order butterworth active bandpass filter design for single-sided magnetic particle imaging scanner', *Computers Electrical Engineering* **10**, 17–21.

Alim, S. and Alang Md Rashid, N. K. (2018), *Some Commonly Used Speech Feature Extraction Algorithms*.

Basic Song Structure Explained (2011), https://www.fender.com/articles/play/parts-of-a-song-keep-it-straight//.

Bonjyotsna, A. and Bhuyan, M. (2013), Signal processing for segmentation of vocal and non-vocal regions in songs: A review, *in* '2013 International Conference on Signal Processing , Image Processing Pattern Recognition', pp. 87–91.

Candes, E. J., Li, X., Ma, Y. and Wright, J. (2009), 'Robust principal component analysis?'.

Cano, P., Batlle, E., Kalker, T. and Haitsma, J. (2003), 'A review of algorithms for audio fingerprinting'.

Demetriou, A., Jansson, A., Kumar, A. and Bittner, R. M. (2018), Vocals in music matter: the relevance of vocals in the minds of listeners, *in* 'ISMIR'.

Deshmukh, S. and Bhirud, G. (2014), 'Analysis and application of audio features extraction and classification method to be used for north indian classical music's singer identification problem', *International Journal of Advanced Research in Computer and Communication Engineering* **3**.

Dharini, D. and Revathy, A. (2014), Singer identification using clustering algorithm, *in* '2014 International Conference on Communication and Signal Processing', pp. 1927–1931.

Dissanayake, M. (2016), 'Intellectual property law of sri lanka', https://www.hg.org/legal-articles/intellectual-property-law-in-sri-lanka-21205/.

Driedger, J., Müller, M. and Disch, S. (2014), Extending harmonic-percussive separation of audio signals.

Dunn, R. (2013), 'The phenomenon of the voice: A comparison', *The Phenomenon of Singing* **1**(0).
URL: *https://journals.library.mun.ca/ojs/index.php/singing/article/view/932*

Fitzgerald, D. (2010), 'Harmonic/percussive separation using median filtering', *13th International Conference on Digital Audio Effects (DAFx-10)* .

Fitzgerald, D. and Jaiswal, R. (2012), On the use of masking filters in sound source separation.

Froitzheim, S. (2017), A short introduction to audio fingerprinting with a focus on shazam.

Fränti, P. and Sieranoja, S. (2019), 'How much can k-means be improved by using better initialization and repeats?', *Pattern Recognition* **93**, 95 – 112.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320319301608*

Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. (2006), F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search, Vol. 5, pp. V – V.

Haitsma, J. and Kalker, T. (2002), A highly robust audio fingerprinting system, Vol. 32.

Harmonic Percussive Source Separation (2014), https://www.audiolabs-erlangen.de/content/05-fau/professor/00-mueller/02-teaching/2016w_mpa/LabCourse_HPSS.pdf//.

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

Hsu, C., Wang, D., Jang, J. R. and Hu, K. (2012), 'A tandem algorithm for singing pitch extraction and voice separation from music accompaniment', *IEEE Transactions on Audio, Speech, and Language Processing* **20**(5), 1482–1491.

Huang, P.-S., Chen, S., Smaragdis, P. and Hasegawa-Johnson, M. (2012), 'Singing-voice separation from monaural recordings using robust principal component analysis', *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on* .

Introduction to Audio Analysis (2014), *in* T. Giannakopoulos and A. Pikrakis, eds, 'Introduction to Audio Analysis', Academic Press, Oxford, pp. 251 – 258.
URL: *http://www.sciencedirect.com/science/article/pii/B9780080993881000200*

Jensen, K. K. (1999), Timbre models of musical sounds - from the model of one sound to the model of one instrument, *in* 'Technical report / University of Copenhagen / Datalogisk institut'.

Kenneth, B. (2004), 'Voiced vs. voiceless consonants', https://www.thoughtco.com/voiced-and-voiceless-consonants-1212092/.

Kim, Y. and Whitman, B. (2002), 'Singer identification in popular music recordings using voice coding features'.

Kos, M., Kacic, Z. and Vlaj, D. (2013), 'Acoustic classification and segmentation using modified spectral roll-off and variance-based features', *Digital Signal Processing* **23**, 659–674.

Li, Y. and Wang, D. (2007), 'Separation of singing voice from music accompaniment for monaural recordings', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(4), 1475–1487.

Lin, Z., Chen, M. and Ma, Y. (2010), 'The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices', *Mathematical Programming* **9**.

Murray, C. (2014), 'Legal pop, murray law group, when is imitation infringement', https://offthemarkipsolutions.com/intellectual-property/when-is-imitation-infringement/.

Ozerov, A., Philippe, P., Bimbot, F. and Gribonval, R. (2007), 'Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs', *IEEE Transactions on Audio, Speech, and Language Processing* **15**.

O'Neil, J. (2004), 'Advertising agency's use of sound alike infringes singer's right of publicity', https://www.theexpertinstitute.com/case-studies/advertising-agencys-use-sound-alike-infringes-singers-right-publicity/.

Pampalk, E., Flexer, A. and Widmer, G. (2005), Improvements of audio-based music similarity and genre classificaton., pp. 628–633.

Rafii, Z. and Pardo, B. (2011), A simple music/voice separation method based on the extraction of the repeating musical structure, pp. 221 – 224.

Sahoo, T. and Patra, S. (2014), 'Silence removal and endpoint detection of speech signal for text independent speaker identification', *International Journal of Image, Graphics and Signal Processing* **6**, 27–35.

Saini, M. and Jain, S. (2013), Comprehensive analysis of signal processing techniques used for speaker identification.

Salamon, J. and Gómez, E. (2010), Melody extraction from polyphonic music signals, Vol. 31.

Salamon, J., Gómez, E., W., E. and Richard, G. (2014), 'Melody extraction from polyphonic music signals: Approaches, applications and challenges', *IEEE Signal Processing Magazine* **31**, 118–134.

Singing scales (2000), https://www.musikalessons.com/blog/2016/11/singing-scales//.

Subramanian, H., Rao, P. and Roy, D. (2004), Audio signal classification.

Tachibana, H., Ono, N. and Sagayama, S. (2014), 'Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(1), 228–237.

Umesh, S. and Sinha, R. (2007), 'A study of filter bank smoothing in mfcc features for recognition of children's speech', *IEEE Transactions on Audio, Speech Language Processing* **15**, 2418–2430.

Velankar, M. (2013), 'Study paper for timbre identification in sound', *http://www.ijert.org/view.php?id=5942title=study-paper-for-timbre-identification-in-sound* **2**.

Velankar, M. (2014), 'Automatic classification of instrumental music human voice', *http://www.ijarcst.com/doc/vol2-issue2/ver.2/diksha.pdf* **2**, 242.

Vembu, S. and Baumann, S. (2005), Separation of vocals from polyphonic audio recordings ., pp. 337–344.

Voice Classification: An Examination of Methodology (2013), https://en.wikipedia.org/wiki/Special:BookSources/978-1-56593-940-0//.

What are the different styles of music played on the banjo? (2010), https://www.quora.com/What-are-the-different-styles-of-music-played-on-the-banjo//.

Whitman, B., Flake, G. and Lawrence, S. (2001), Artist detection in music with minnowmatch, pp. 559 – 568.

Wright, J., Ganesh, A., Rao, S., Peng, Y. and Ma, Y. (2009), Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization., Vol. 1, pp. 2080–2088.

Zhu, B., Li, W., Li, R. and Xue, X. (2013), 'Multi-stage non-negative matrix factorization for monaural singing voice separation', *Audio, Speech, and Language Processing, IEEE Transactions on* **21**, 2096–2107.

Appendices

Appendix A

Codings

```
import librosa
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
import csv
import pydub
import cpickle

def repet(audio_path):
    y , sr = librosa.load(audio_path)
    S_full, phase = librosa.magphase(librosa.stft(y))
    idx = slice(*librosa.time_to_frames([30, 35], sr=sr))
    S_filter = librosa.decompose.nn_filter(S_full,
    aggregate=np.median,metric='cosine',width=int
    (librosa.time_to_frames(2, sr=sr)))
    S_filter = np.minimum(S_full, S_filter)
    margin_i, margin_v = 2, 10
    power = 2

    mask_i = librosa.util.softmask(S_filter, margin_i *
                          (S_full - S_filter),power=power)
    mask_v = librosa.util.softmask(S_full - S_filter,
                                   margin_v * S_filter,
                                   power=power)

    S_foreground = mask_v * S_full
    S_background = mask_i * S_full
    D_foreground = S_foreground * phase
    y_foreground = librosa.istft(D_foreground)
    librosa.output.write_wav('final321.wav',
    y_foreground, sr)
def band_pass(audio_path):
    lo,hi=85,880
    y,sr = librosa.load(audio_path)
    b,a=butter(N=6, Wn=[2*lo/sr, 2*hi/sr], btype='band')
    x = lfilter(b,a,y)
    librosa.output.write_wav(outname, x, sr)


def harmonicPercussive(audio_path):
    y, sr = librosa.load(audio_path)
    D = librosa.stft(y)
    H, P = librosa.decompose.hpss(D)
    harmonic = librosa.istft(H)
    percussive = librosa.istft(P)
    librosa.output.write_wav('percuss.mp3',percussive, sr)
    librosa.output.write_wav('harmonic.wav',harmonic,sr)


def extractFeatures(audio_path):
    y, sr = librosa.load(songname)
    chroma_stft = librosa.feature.chroma_stft(y=z, sr=sr)
    rmse = librosa.feature.rmse(y=z)
```

```
spec_cent = librosa.feature.spectral_centroid
(y=z, sr=sr)
spec_bw = librosa.feature.spectral_bandwidth
(y=z, sr=sr)
rolloff = librosa.feature.spectral_rolloff(y=z, sr=sr)
zcr = librosa.feature.zero_crossing_rate(z)
mfcc = librosa.feature.mfcc(y=z, sr=sr)
file = open('data_test_harrep.csv', 'a', newline='')
    with file:
        writer = csv.writer(file)
        writer.writerow(to_append.split())
data = pd.read_csv('data.csv')
```