Singer Identification

Submitted By

Jainesh Doshi

13MCEC31

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Singer Identification

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Jainesh Doshi

(13MCEC31)

Guided By

Prof. Sapan H. Mankad

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Certificate

This is to certify that the major project entitled **"Singer Identification"** submitted by **Jainesh Doshi (Roll No: 13MCEC31)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Sapan H. Mankad
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Priyanka Sharma
Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr P. N. Tekwani
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Jainesh Doshi**, Roll No. **13MCEC31**, give undertaking that the Major Project entitled "**Singer Identification**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place: Ahmedabad

Endorsed by

Prof. Sapan H. Mankad

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Sapan H. Mankad**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr P. N. Tekwani**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Jainesh Doshi

13MCEC31

Abstract

Survey of the Music Information retrieval (MIR), in particular paying attention to latest developments, such as singer identification. First elaborate on well-established and proven methods for feature extraction and singer identification, from sound files. But in music information retrieval domain, singer identification is very difficult topic. Because with singing voice, background instrumental music is also included which reduces the performance of the system. Subsequently, review of current work on user analysis and modeling in the context of music recommendation and retrieval, addressing the recent trend towards singer identification.

A discussion follows about the important aspect of how various Music Information Retrieval approaches to different problems are evaluated and compared. Eventually, a discussion about the major open challenges concludes the survey.

Abbreviations

MIR	Music Information Retrieval.
MFCC	Mel Frequency Cepstral coefficient.
LPC	Linear Predictive Coefficient.
GMM	Gaussian Mixture Model.
ISMIR	International Society for Music Information Retrieval conference

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction to MIR System

Music is a language in itself. It requires a understanding of practical, theoretical, and analytical approaches in its various environments. Now a days, there is a growing amount of MP3 music data available on the Internet. With the more MP3 music data the problem related to music classification and content based music retrieval are getting more attention.

Music data are huge because of growth of the Internet and so that techniques for content based retrieval of multimedia data are in demand in the area of the multimedia database. MP3 is an ISO international standard for the compression of digital audio data. [2]

Music is a complex form of information. Sometimes users ask "why do I need MIR?", the answer is users do need MIR for complete and simple access.

To retrieve the content based data from the multimedia data, Music Information Retrieval system is very useful. Music Information retrieval is the science of retrieving information from music.

There are many information to retrieve from the music like which instruments are played, which singers are sung, who is the composer of the music, given sound file is only instrumental or only vocal or instrumental and vocal, how many instruments are played, how many singers are sung, etc. [2]

Music is different from text in many ways. Mostly it is different from information retrieval because text data is simple compare to music information data. Music infor-

mation has many different representations and formats. It is continually improving and much advancement has been made in MIR.

1.2 Motivation

In our society, music is a general topic as almost everyone enjoys listening to it and many also create. Music information retrieval is the research field in which foremost concerned with the extraction of meaningful features from music, indexing of music using these features, and the development of different search and retrieval schemes. [2] It comes as a surprise that the research field of MIR is a relatively origin less than two decades ago. Some of the most important reason for its success are audio compression technique is developed in the late 1990s, mobile music players are available widely.

1.3 History and evolution

Symbolic representation of music pieces in MIR research focused on early working. Many important features of music are related to music content, contextual aspects that can be modelled from user generated information available for instance on the Internet. [3] With respect to assessment, user-centric techniques go for considering distinctive variables in the impression of music qualities, specifically of music comparability.

1.4 Music Information Retrieval System

From the music, there are many information which are user wants to retrieve. But some information which are really need and which are very useful to user.

Here as shown in figure which is whole MIR system. Here for MIR system sound file is as a input to the system. System is performing operation on the sound file and gives the output. Now what is the output? So answer is depend on user means what user wants. If user wants to identify the sound file is purely instrumental or purely vocal. If sound file is purely instrumental, then how many instruments are playing and which instruments are playing. If given sound file is purely vocal then how many singers are performing and who the singers are. Another classification is genre classification from the system. This system is take input as a sound file and apply feature extraction techniques and extract the features. With the use of classification techniques, train the data-set.

Instruments, Singers and Genre of the sound files are important. So from the sound

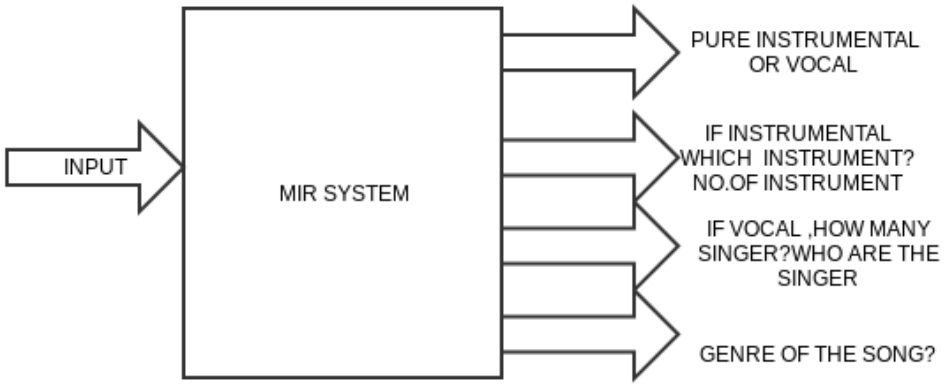

Figure 1.1: MIR System

file, we can extract that given sound file is purely instrumental or purely vocal or mixture of instrumental and vocal both.

Second thing is to extract if given sound file is instrumental then how many instruments are played and which instruments means name of the instruments.

Third thing is to extract if given sound file is vocal then how many singers are singing and who is/are the singers means name of the singers.

Last thing is to extract given sound file is in which genre means if we can listen the song then we can define the song is sad song, romantic song, etc.

1.5 Objective of Project

The main objective behind this project is to identify the singers from the sound file based on different characteristics of the singers like pitch, frequency, etc. Recognize a singer based on different types of features. With the use of feature extraction techniques, features are extracted from all of the sound file and make a data set.

1.6 Problem Definition

Research on Music Information Retrieval contains a rich and differing set of areas whose extension goes well past unimportant recovery of records. Topics are related to the extraction of useful features from music content and context. Parts of Music Information

Retrieval: First is Singer identification, second is Instrument identification (classification), third is Raaga identification, and fourth is Genre classification.

So mainly focus on purely instrument or purely vocal. For that discriminating between instrumental and vocal components in song/sound file.

Music Information Retrieval system for singer identification from the sound file.

1.7 Outline of Thesis

In the next section, literature survey can be explained. In the literature survey, Singer identification is explained. Also, explained about vocal and instrument identification. In the 3^{rd} section, Proposed architecture is explained, It is also explained about ""feature extraction" and "feature matching" with the explanation of some feature extraction techniques. The 4^{th} section shows the implementation. It is also shows the dataset which is using in this system and implementation of features dataset with neural network. The 5^{th} secton shows the conclusion. The 6^{th} section shows the future work.

Chapter 2

Literature Survey

The field of MIR has great job in introducing music terminology and categories of music features that are important for retrieval but this considerable changes during recent years. Further identifies different users on an MIR system who are discusses their individual requirements and needs. From that features and requirements, it will gives a through introduction to MIR, numerous new research headings have developed inside of the field from that point forward.

An outline of the field of MIR from a signal processing point of view. They henceforth emphatically concentrate on sound examination and music content-based comparability and recovery. A MIR is an exceptionally multidisciplinary research field the yearly International Society for Music Information Retrieval conference (ISMIR) unites scientists of fields as various as Electrical Engineering, Library Science, Psychology, Computer Science, Sociology, Mathematics, Music, Theory, and law.

Here mainly focused on singer identification and discriminating between instrumental and vocal components in song/sound file.

2.1 Singer Identification

Music IR differs in several ways from evaluation in text IR Such as:

- Availability of data.
- Multimedia information is inherently more complex than text.

Starts with identifying the first audio descriptors then using these feature vectors as input to further classification using Gaussian Mixture model or Hidden Markov model. Mel Frequency Cepstral coefficient (MFCC) as extracted features, given as input to neural network classifier. [4]

Singers voice is the most important part in the music and it is very useful. But the major problem is to segmentation of vocal and non-vocal parts in a sound. [5] One proposed method is carious acoustic features extracted from the sound file for singers identity but this method failed to define background music.[6] Other is identify the singer using music structure knowledge but this method is not reliable for short test sound files.[7]

Other proposed method was using MFCC feature vectors and artificial neural networks classifiers. But accuracy is decrease when number of sound file is increase. Also LPC as a feature vectors and Gaussian Mixture Model (GMM) as a classifier. [8]

A large portion of the current systems considered music recordings with no background music and for the recognition, features were used LPC (Linear Predictive Coefficient). In any case, human auditory system doesn't treat all frequencies on a linear scale.

2.2 Vocal and instrument identification

Instruments are not registered as a musical instrument or we can say that user wants to identify new instruments. To solve this problem, distinguishing between registered and non-registered instruments. When given sound is registered, its instrument name is identify. Even it is not registered, its category name can be identify.

But issue is achieving such identification is to adopt a musical instrument hierarchy reflecting the acoustical similarity.[9]

To identify the instrument, comparison of two classification methods are used. But it is not a proper way to identify the instrument. [10]

We can use features which are extract from the sound file, to identify the instruments. Features like MFCC is used to identify the instruments. Several set of features derived from MFCC and compare their performance against other features. But MFCC over several time scales in the context of automatic instrument identification.[11]

Chapter 3

Proposed Architecture

3.1 Feature Extraction and Feature Matching

To build any MIR system used some kind of information, which represent each music uniquely,it must be extracted from the sound file. This information represented by certain "parameters" or "features". This task is called as "feature extraction" from the sound file.

In "feature matching" task, we try to match the features of different sound file. For the feature matching task, neural network is used. Also, there are many techniques to match the features.

There are many techniques which are used for extracting the features. MFCC(Mel Frequency Cepstral Coefficients), LPC (Linear Predictive Coding), LPCC(Linear Predictive Cepstral Coefficients), PLP (Perceptual Linear Predictive Coefficients). Some of them are explained here.

3.1.1 Lineare Predictive Coding(LPC)

LPC is one of the traditional techniques in Feature Extraction. It is very useful in Encoding the Speech at a lower Bit-rate. The linear combination of the past speech signals can be used to approximate the value of the current Speech Signal. LPC is based on the Speech Production System of Humans. It uses the Source-Filter Model. The goal of LPC is to reduce the Summation of Squared difference between the original and predicted speech signal. The output is the set of Unique Coefficients. These coefficients are predicted over each frame of the speech signal, which is generally of 20ms[12]. The

complete process of LPC is shown in the figure:

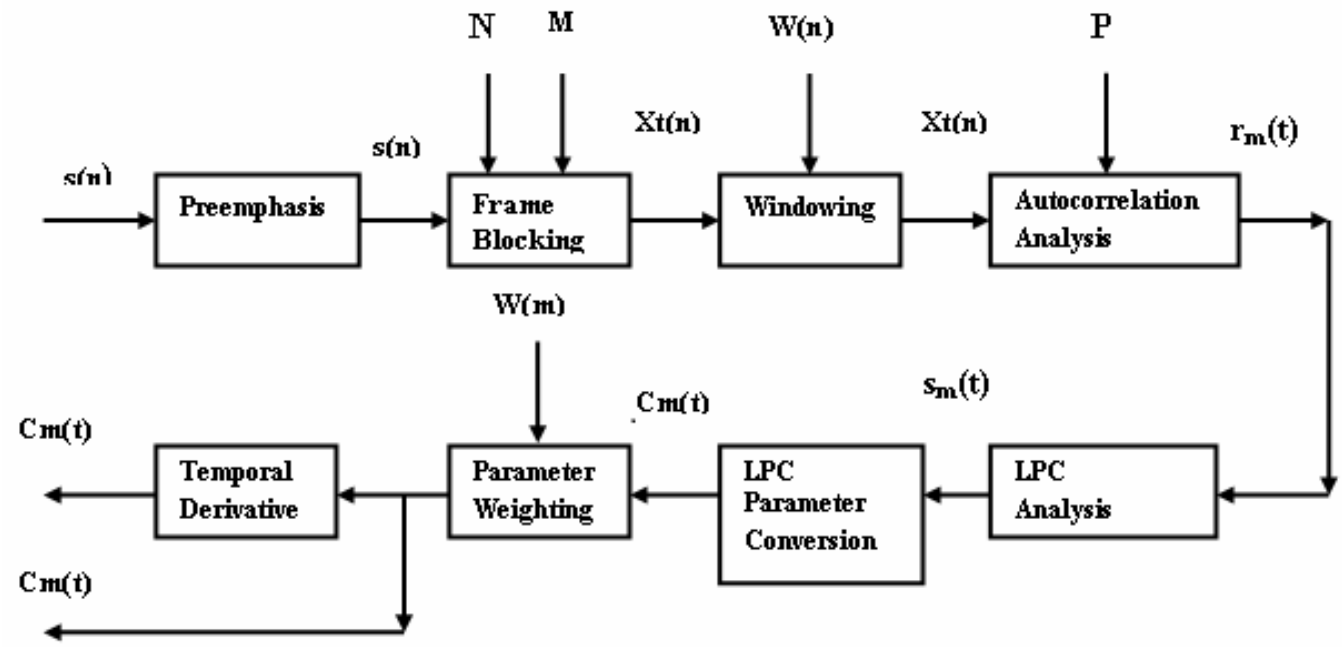

Figure 3.1: The LPC Process

3.1.2 Mel Frequency Cepstral Coefficient(MFCC)

MFCC Technique is the most widely used technique for the Feature Extraction of the Speech Signal. It is used in the most of the Applications of Speech Processing. MFCC is based on the Human Ear Perception System. Human ear is capable of detecting the frequency ranging from 20 to 20,000 Hz. Human Ear acts as a Filter, which focuses on only some components of the frequency range. These filters lies non-uniformly on the Frequency Axis. They are spaced linearly at low frequencies, below 1000 Hz and logarithmically at High frequencies, above 1000 Hz [1]. This entire spacing scale of the filters is known as Mel Frequency Scale. It can be shown as below :

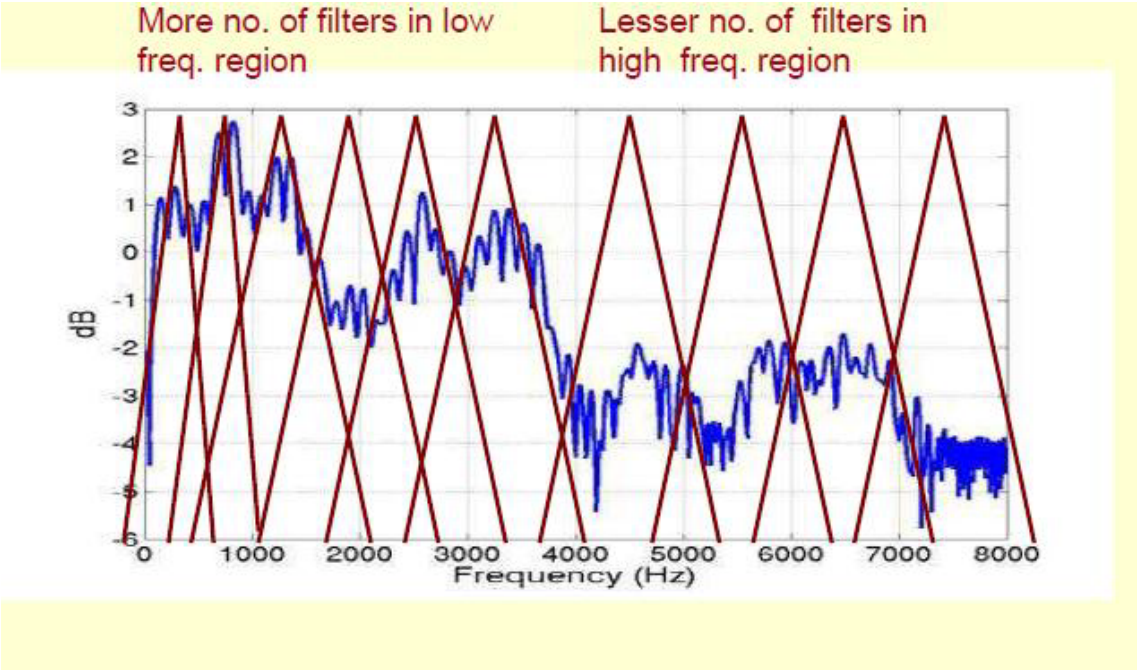

Figure 3.2: Mel Frequency Filters[1]

The Human Ear Cochlea is shown below:

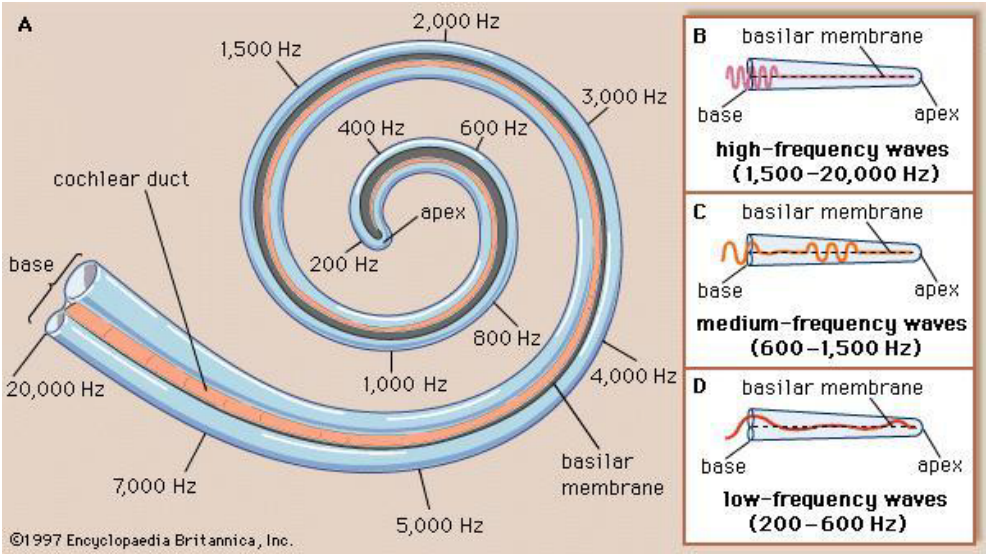

Figure 3.3: Human Ear Cochlea[1]

As shown in the figure, the frequencies below 1000 Hz are linearly spaced on the Frequency Axis. Above that, they are spaced logarithmically. The Low Frequency Waves contains the range of 200-600 Hz. The Medium Frequency Waves have the range of 600-1500 Hz. And, the High Frequency Waves have the range of 1500- 20,000 Hz. The MFCC technique works on the same concept. Generally, 20 coefficients of MFCC are used, but 10-12 are also considerably sufficient to be used in the Speech Processing Applications.

The MFCC Process can be described as below:

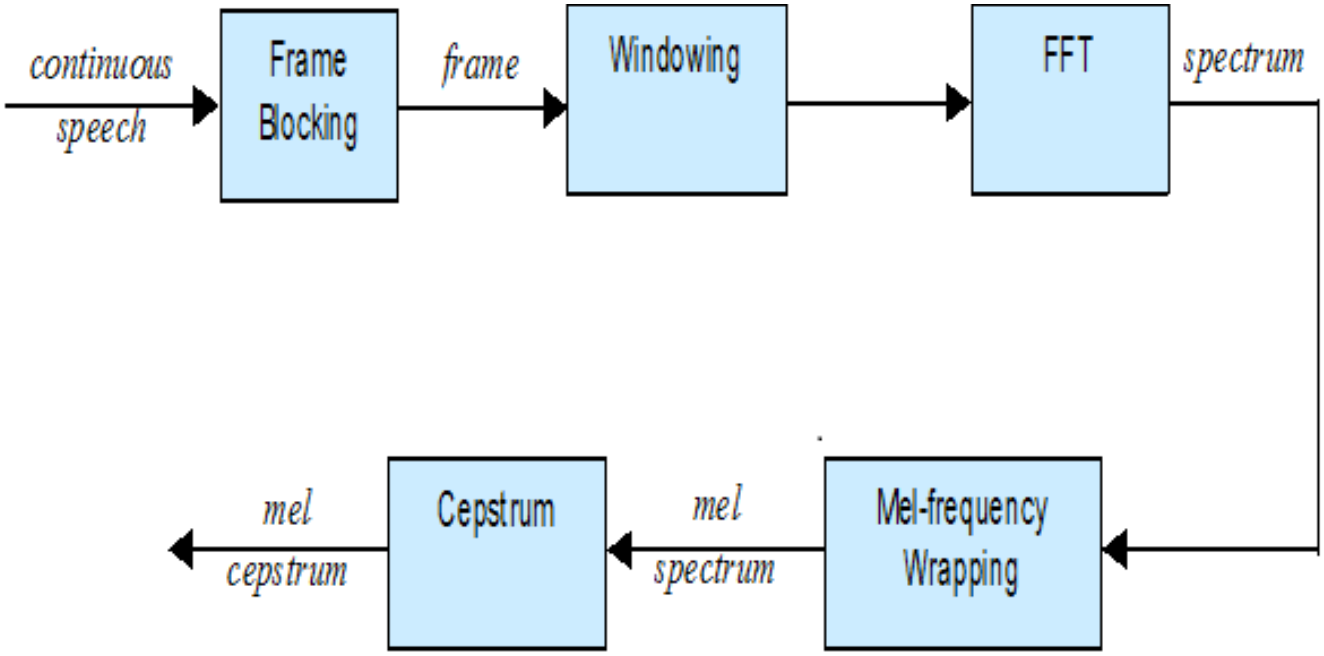

Figure 3.4: MFCC Block Diagram[1]

First of all, the sampling rate is chosen for the input. Then the process of Noise Removal is done. After that, we select the size of the frame, which is denoted by N. Then, we select the no. of samples after which the new frame will start and will overlap the current frame, which is denoted by M , where M<N. Thus, we try to block the speech into frames. Then the step of Windowing comes. We do this to Minimize the discontinuity at starting and ending of each individual frame. If the window is defined by w(n), then, the resulting signal after windowing is :

$y_l(n) = x_l(n)w(n), 0 \leq n \leq N-1$

Here, Hamming Window is used, which is :

$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), 0 \leq n \leq N-1$

Now comes the step of FFT. It converts each frame from Time Domain to Frequency domain. Here, FFT is used to implement DFT(Discrete Fourier Transform). The result gives a Spectrum.

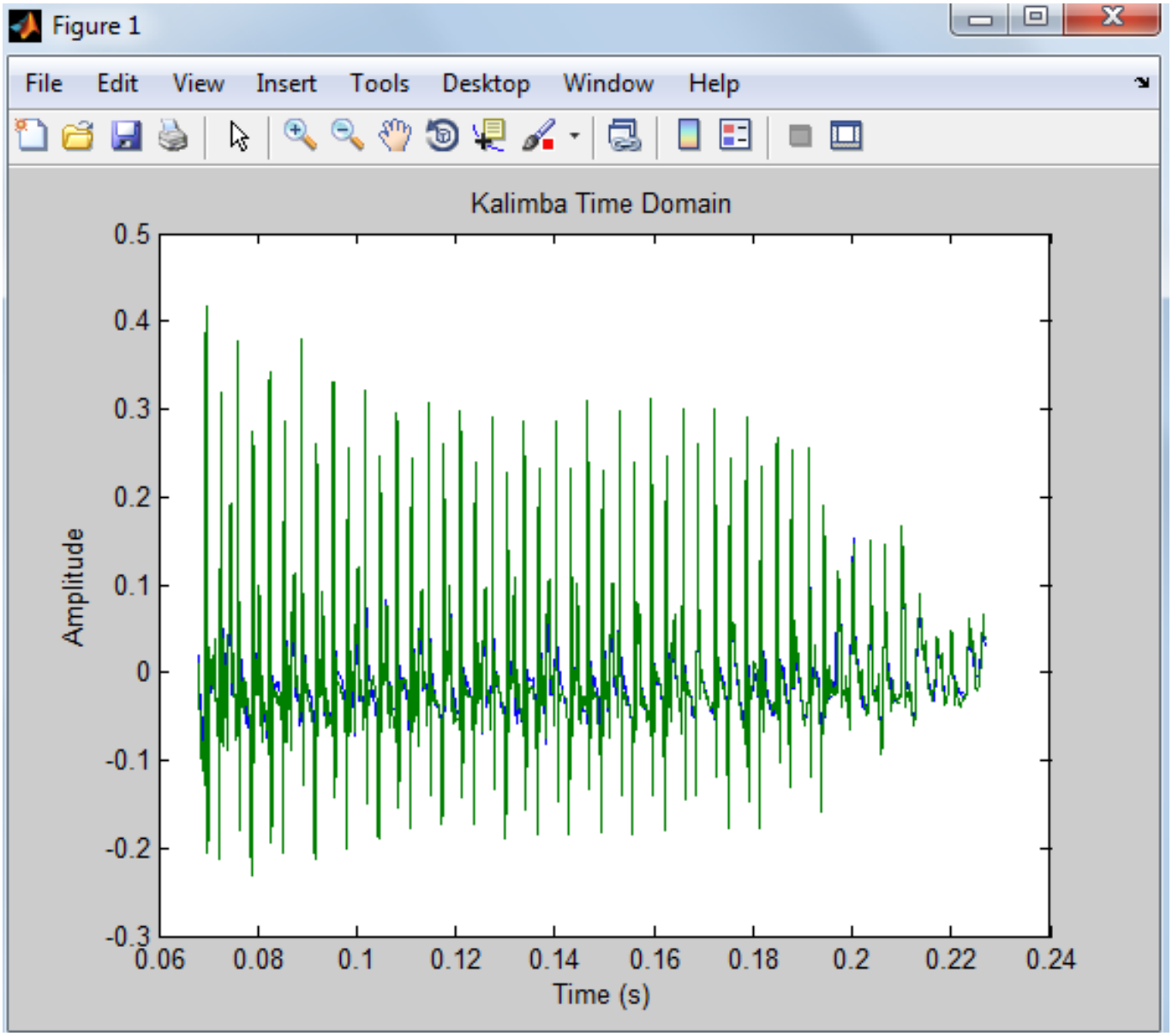

Figure 3.5: Time Domain of an Audio File

And its Frequency domain is shown below:

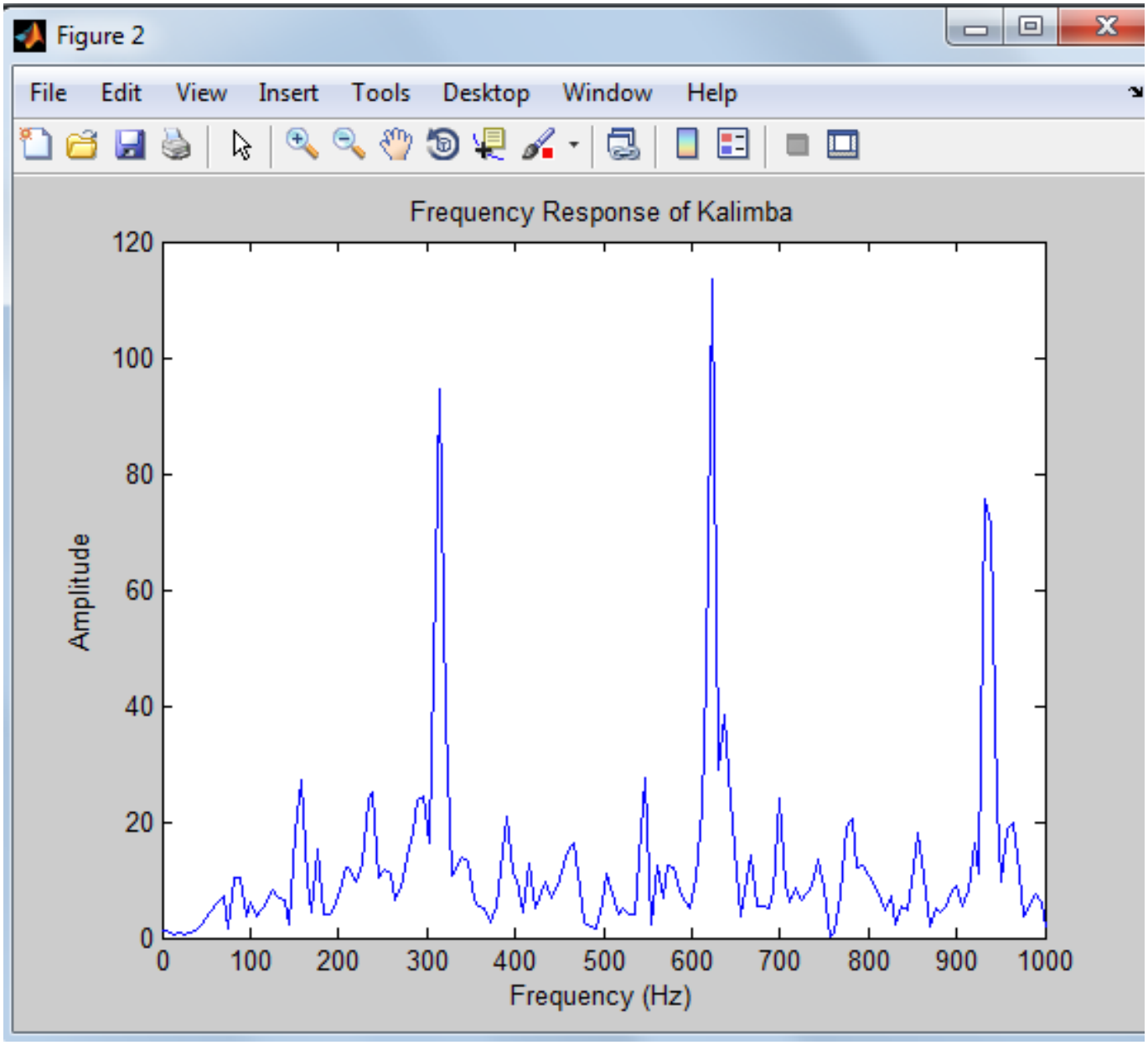

Figure 3.6: Frequency Domain of an Audio File

The next step is Mel Frequency Warping. For each sound of original frequency f, a subjective pitch is calculated on the Mel Scale. We use Filter-Bank, spaced uniformly over the Mel Scale. The Mel-Frequencys interval is defined and the no. of mel cepstrum coefficients (K) are chosen. In this implementation , it is chosen as K=13. Here, the Triangular Shaped Window is applied. This Mel warping Filter bank is applied in the Frequency Domain of the Signal. Each filter can be considered as a Histogram Bin.

The final step is Cepstrum where we convert the Log Mel Spectrum to the time domain. The result of this is called Mel Frequency Cepstral Coefficients. We convert

them to the time domain by use of Discrete Cosine Transform(DCT). Here we exclude the first coefficient of MFCC, as it represents the Mean Value of the signal, which has very less useful information of the Speaker[1].

Thus, applying the steps as illustrated above to the frame of 20-30 ms, we get the MFCCs for that frame. The set of MFCCs is called a Acoustic Vector, which is the input to any Speech Processing Application.

3.2 The Proposed System's Architecture

The architecture of the proposed "Singer Identification" system is shown below:

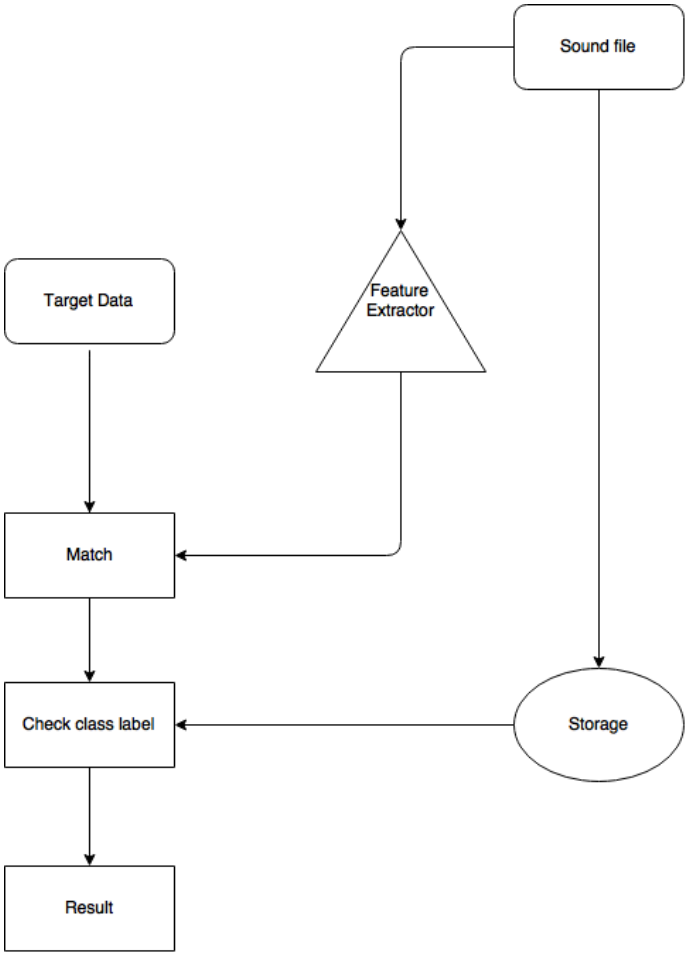

Figure 3.7: The architecture of the proposed system

In this figure, one block is sound file. From that one or more than one sound file is given to the feature extractor to extract the features. Features are like MFCC, LPC, etc. Another block is for target data in which, features are manually given. In target data features are like class label. Resulting features are match with the target data and

from that output can be generated. Storage block is for store the features so no need to generate same sound file's features every time and directly check the class label with the storage data.

Chapter 4

Implementation

4.1 Tools used for the Project

First of all downloading the songs of 10 singers and 60 songs from each singers. So total 600 songs are used in this system.

For segmentation of the songs, Mixpad Multitrack Recording software was used. With the use of this software we can segment all the songs into 20 seconds sound file with the .wav extension.

In this chapter, going to describe the tools and details about to implementation of Singer identification. Initially we have been implemented the model of Singer identification with MFCC and LPC feature extraction techniques. For feature extraction, using Matlab is the best way because MATLAB have some signal processing libraries which make our task little easy and reliable.

Also, there is neural network library in python. So with the use of python code, implement neural network with the MFCC and LPC features and generate the graph.

4.2 Datasets

- **MusiClef 2012 Multimodel music data set**

  To distinguish the music things in the information set and connect them to other information sources, we give arrangements of craftsman and tunes, the comparing collection data for the melodies, and in addition relating MusicBrainz identifiers. For specialists, we facilitate give abbreviated representations ("webartist"), which are utilized to recognize craftsmen in the web creeping subsets. [13]

- **Music Speech**

  In this dataset, music and speech sound files are included. For both the category, extension of the sound files is wav.

- **Singing Database**

  In this dataset Chinese and western sound files are added. There are two types of sound file, first is polyphonic and second is monophonic.[14]

Here, for singer identification sound files are required. For that 600 songs of 10 singers(each singers have 60 songs) download and segment into time period of 20 seconds. Name of singers are as shown below:

Name of Singers	Number of Songs
Sonu Nigam	60
Kishore Kumar	60
Atif Aslam	60
Asha Bhosle	60
Arijit Singh	60
K K	60
Mohit Chauhan	60
Mohammad Rafi	60
Rahat Fateh Ali Khan	60
Shreya Ghoshal	60

4.3 Feature extraction

4.3.1 MFCC

First of all, MFCC feature extract from the sound file using mfcc library in Matlab. From one sound file, 12 mfcc features extracted. All the MFCC features are saved in one excel file or any tabular form. So it is like 12×600 matrix.

4.3.2 LPC

Now for lpc feature extraction, python is used. Here 9 features are extracted from one sound file. Like mfcc features, here also lpc features can save into tabular form. For lpc it is like 9×600 matrix.

4.4 Implementation of dataset with Neural Network

Now, the nntool(Neural Network Tool) of matlab is used to increase the accuracy of the system. Matlabs nntool provides 4 options for selection of appropriate tool.They are :

- Fitting Tool(Input-Output and Curve Fitting)
- Pattern Recognition Tool(Pattern Recognition and Classification)
- Clustering Tool
- Time Series Tool

Out of these 4, we use the Pattern Recognition Tool.

4.4.1 MFCC features as a Training dataset

MFCC of 600 songs are calculated and a 12×600 matrix is generated. Here we have 10 singers so 10 class labels. Now, to identify the singers from the mfcc features, generate 10×600 matrix in which, 1's will be allocated to class labels where the singers belongs to. This matrix will act as the "target dataset" for the neural network.

Now, the Pattern Recognition Tool of matlab is used. Following is the screenshot after entering the input and target dataset.

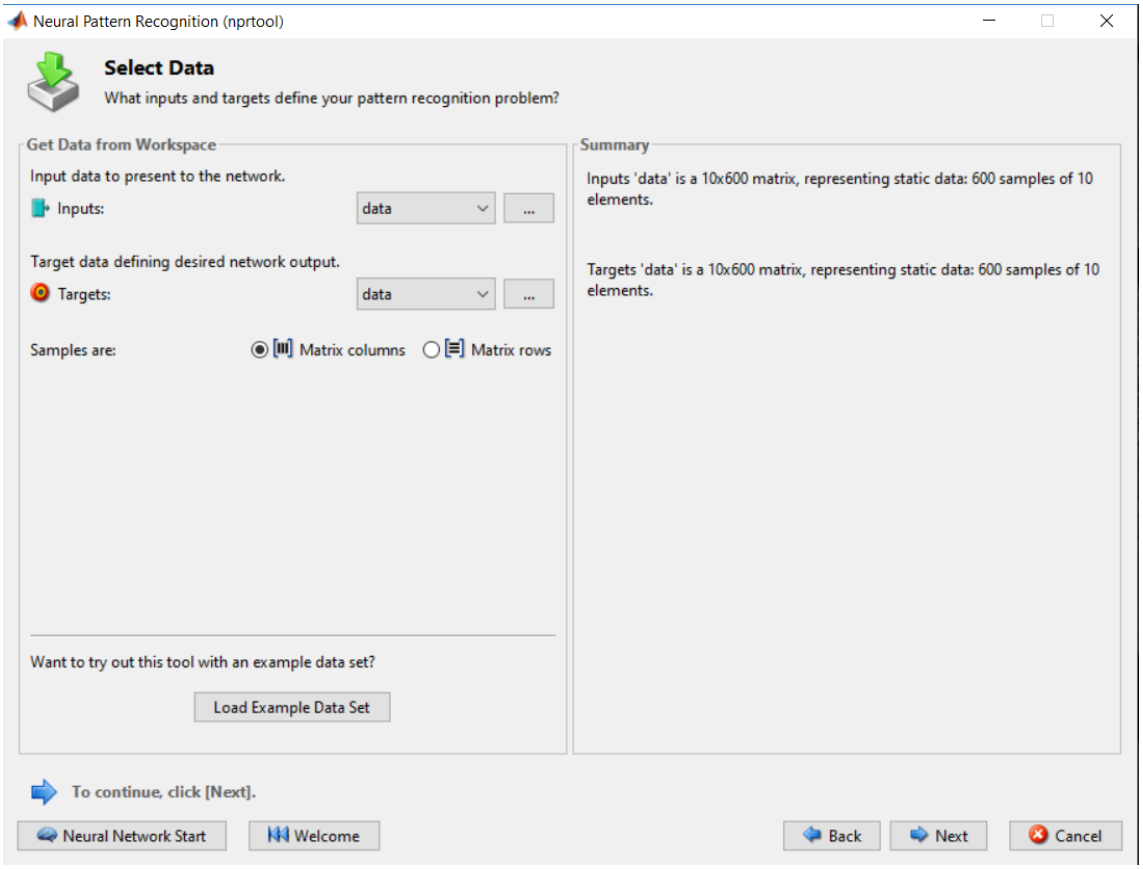

Figure 4.1: Screenshot after loading the Dataset to the nn-pr-tool

Now, the division of 600 samples into 3 categories is done in the following manner:

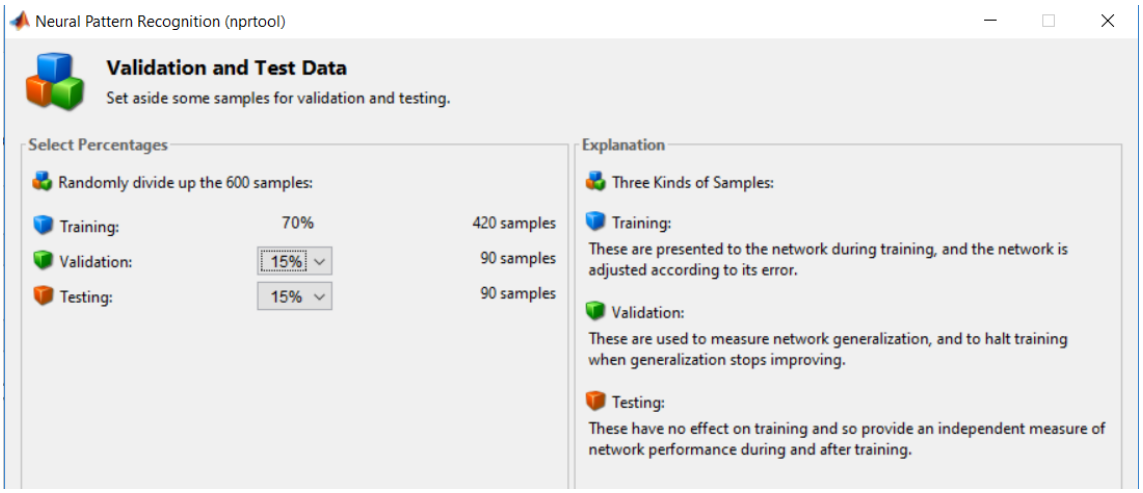

Figure 4.2: Division of the 600 samples

Here the no of hidden neurons are kept to 10, which can be changed later. The generated trained neural network is as below:

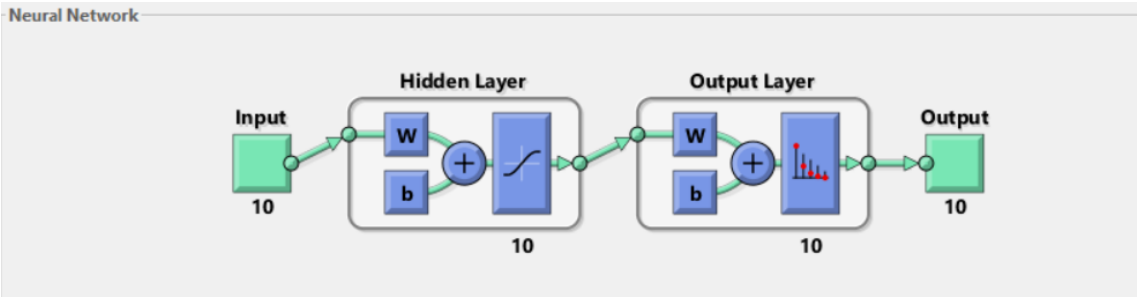

Figure 4.3: Generated neural network for 600 samples

The mean squared error and percentage error are as below:

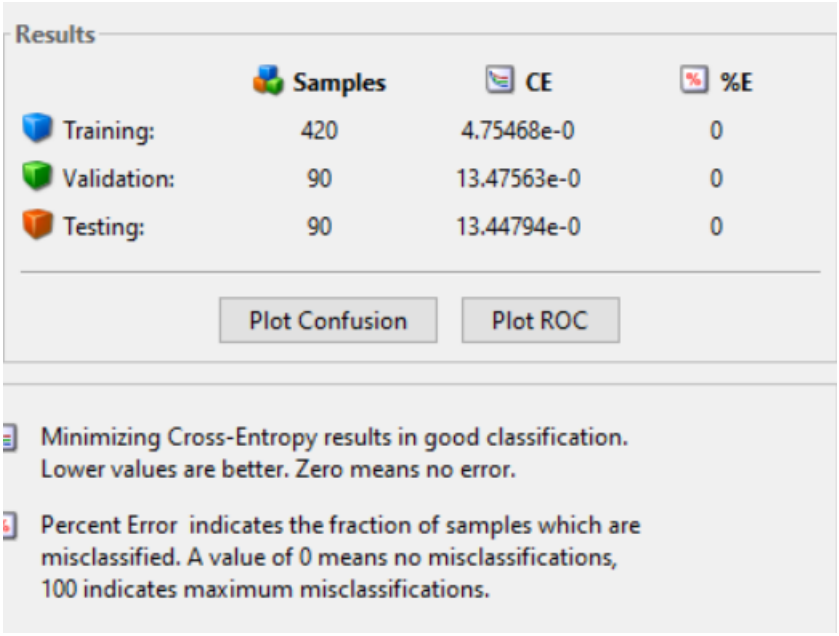

Figure 4.4: "Mean Squared error" and "Percentage error" for 600 samples

The Performance Plot after 50 epochs is shown below :

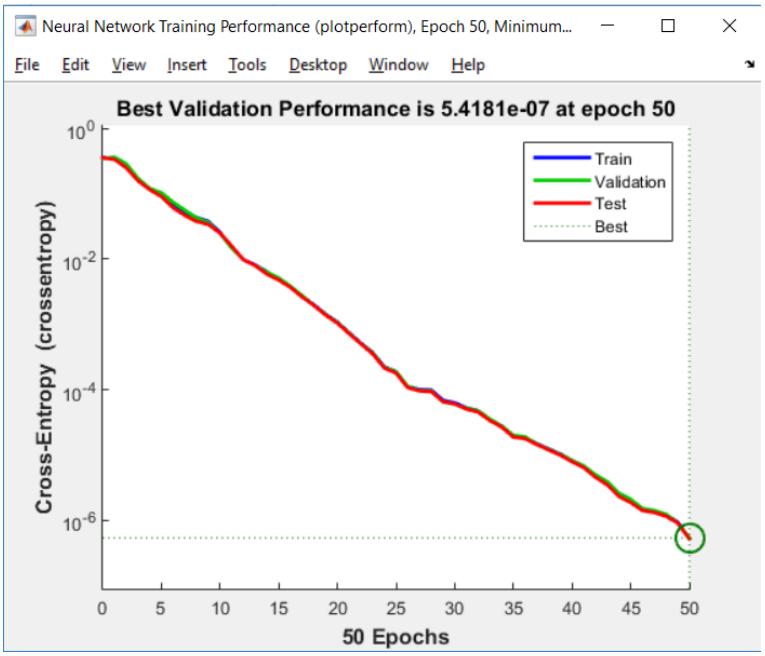

Figure 4.5: Performance Plot after 50 Epochs

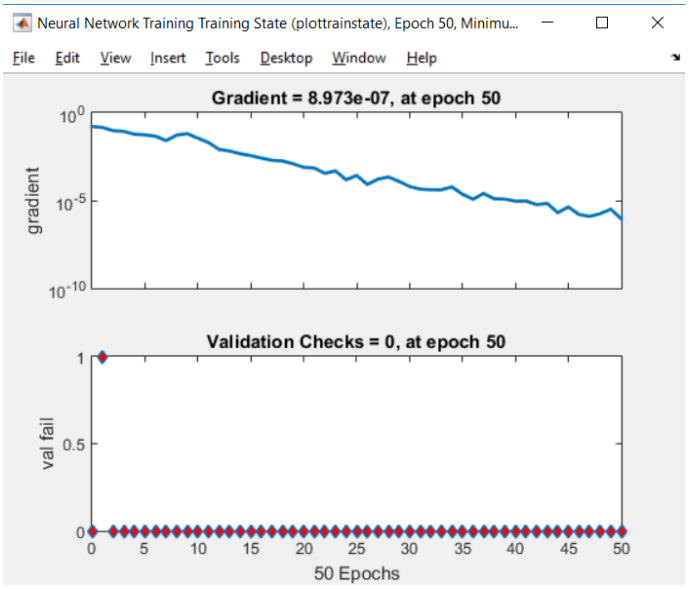

Figure 4.6: Training State Diagram after 50 Epochs

The Error Histogram with 20 Bins is as below:

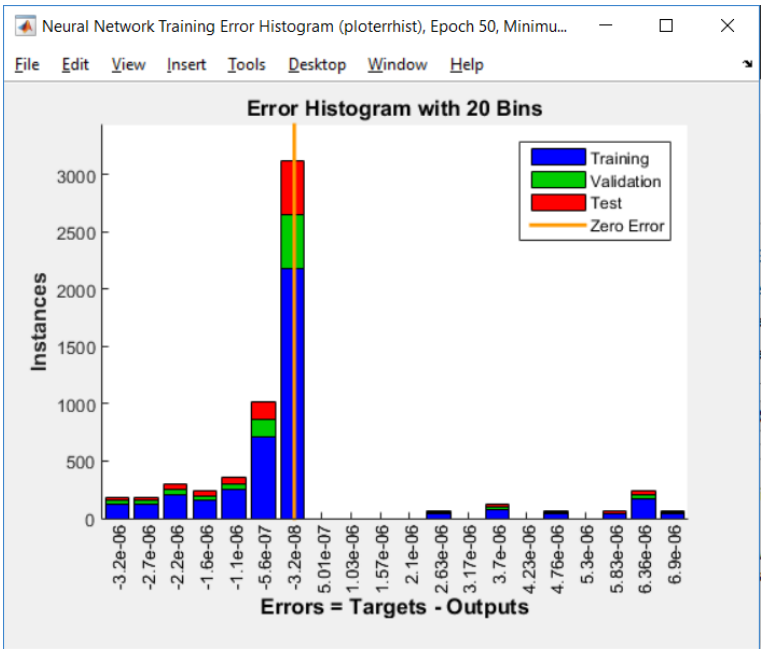

Figure 4.7: Error Histogram with 20 Bins after 50 Epochss

Now, after retraining the network, the Performance Plot after 30 Epochs is shown below:

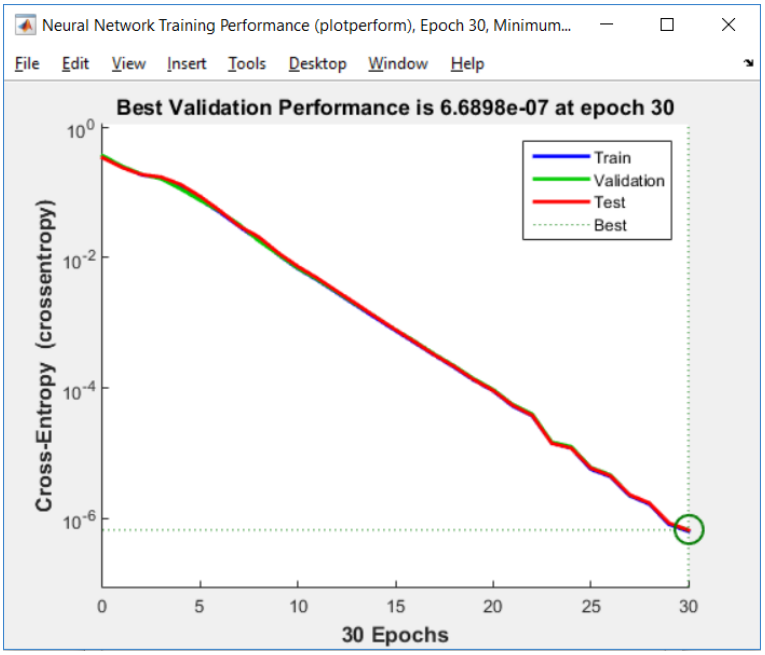

Figure 4.8: Performance Plot after 30 Epochs

It shows that, at epoch 50, the mean squared error is $5.4181e^{-07}$ whereas after 30 epochs its value is $6.6898e^{-07}$.

Now, in python neural network is used with 2 layers and 50 epochs. Input in the neural network is mfcc and target dataset. Graph is as shown below:

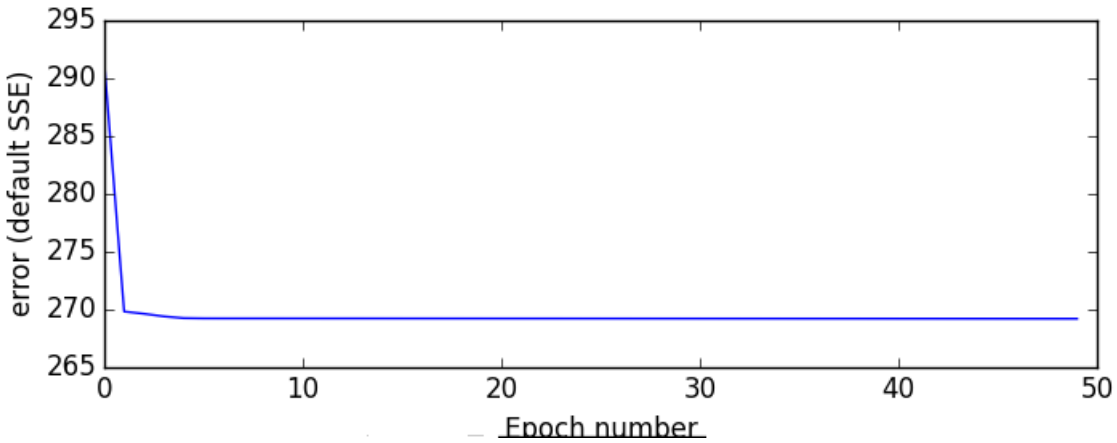

Figure 4.9: Error(SSE) vs Epoch number

4.4.2 LPC features as a Training dataset

Same as MFCC, here 9×600 matrix is generated. Also, class label is same. Now same target dataset is used for neural network.

Now, the Pattern Recognition Tool of matlab is used. Following is the screenshot after entering the input and target dataset.

All the procesures are same as MFCC. Input data is LPC features and target dataset.

The mean squared error and percentage error are as below:

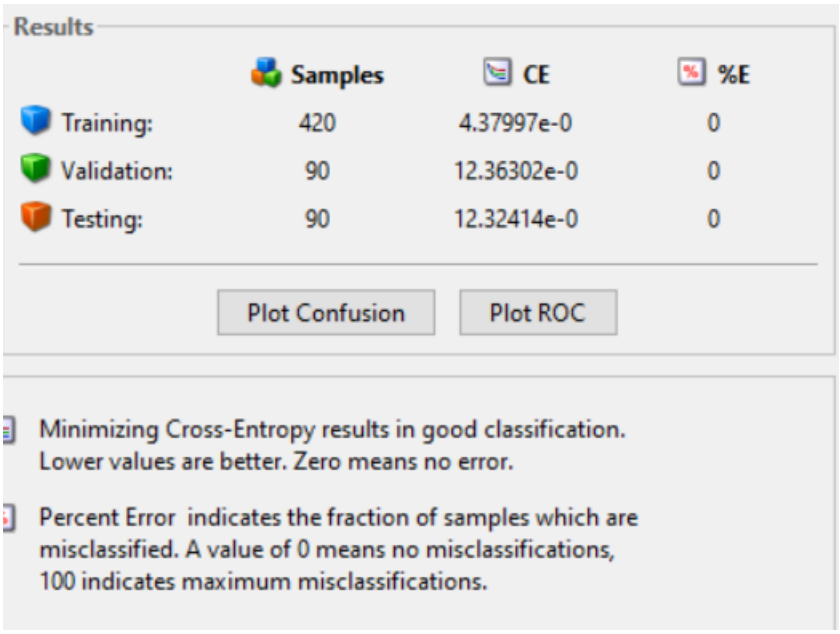

Figure 4.10: "Mean Squared error" and "Percentage error" for 600 samples

The Performance Plot after 28 epochs is shown below :

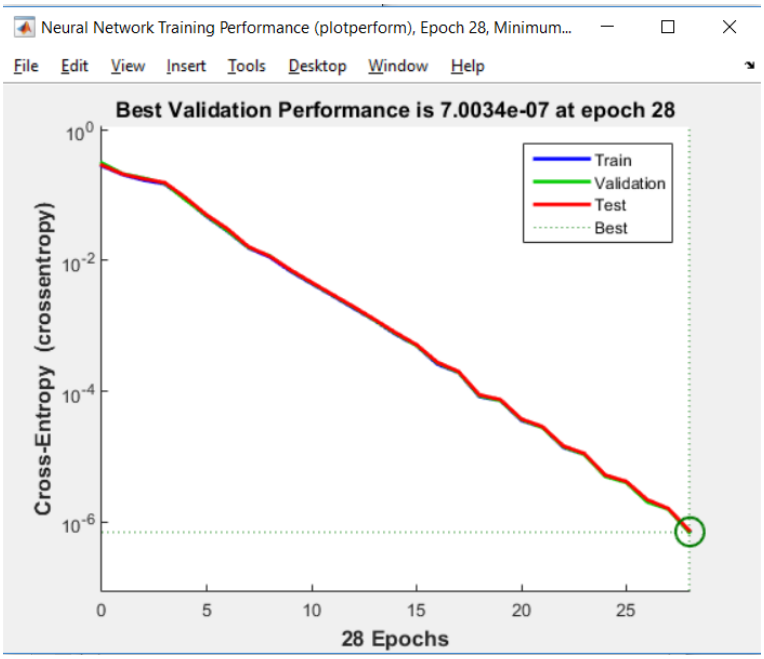

Figure 4.11: Performance Plot after 28 Epochs

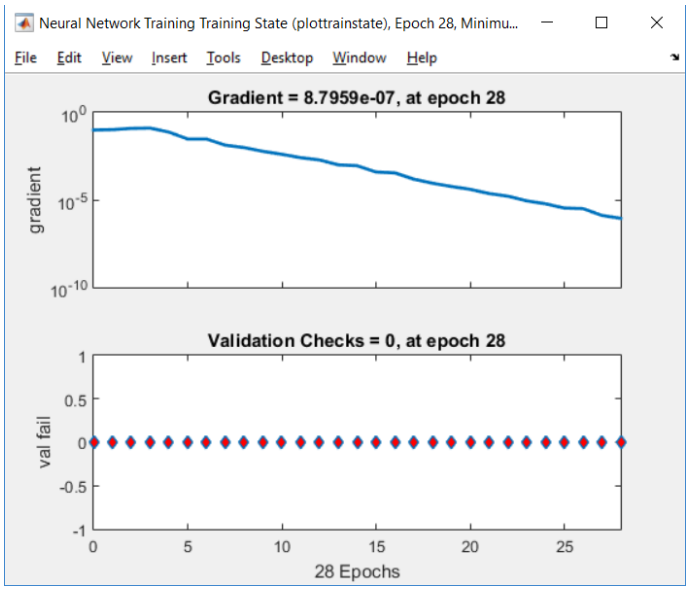

Figure 4.12: Training State Diagram after 28 Epochs

The Error Histogram with 20 Bins is as below:

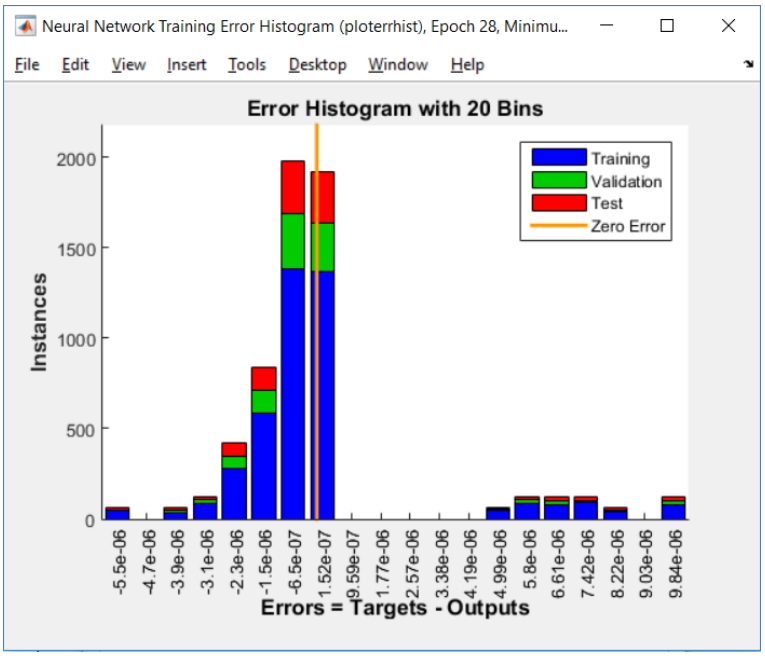

Figure 4.13: Error Histogram with 20 Bins after 28 Epochss

Now, after retraining the network, the Performance Plot after 27 Epochs is shown below:

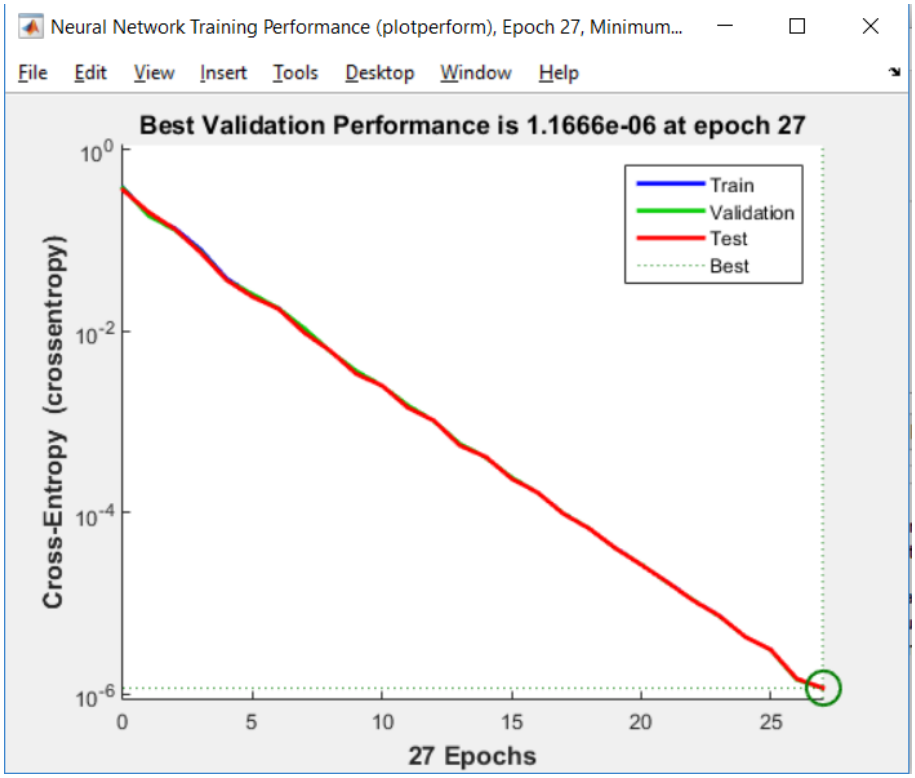

Figure 4.14: Performance Plot after 27 Epochs

It shows that, at epoch 28, the mean squared error is $7.0034e^{-07}$ whereas after 27 epochs its value is $1.1666e^{-06}$.

Now, in python neural network is used with 2 layers and 56 epochs. Input in the neural network is lpc and target dataset. Graph is as shown below:

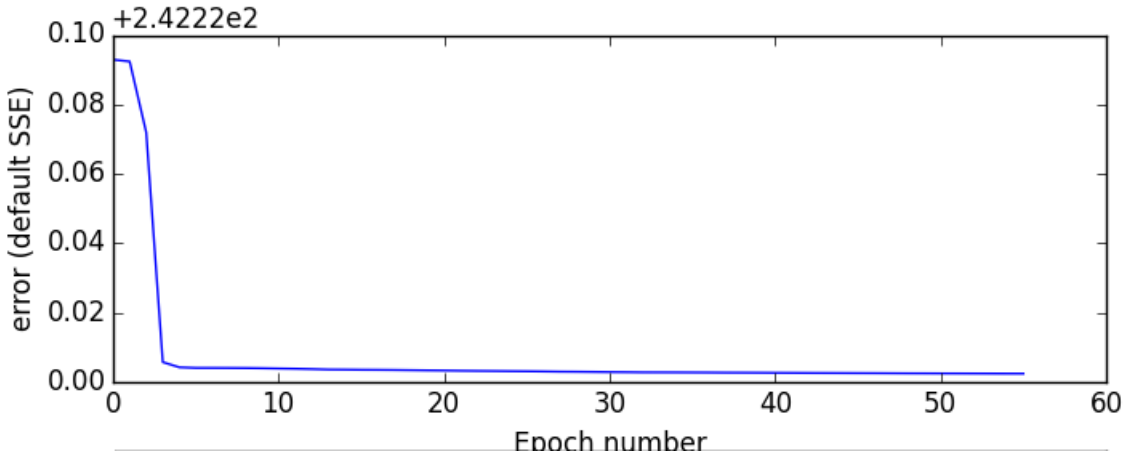

Figure 4.15: Error(SSE) vs Epoch number

Chapter 5

Conclusion

We conclude that MIR system is content-based system and from that extract the features. From the features fetch the content of the given sound file and identify the singers. Features are pitch, LPC, MFCC etc. Training the Neural network again and again will make the it able to distinguish between the various class-labels. The use of other Feature Extraction techniques should also give better results. Here implementation of mfcc and lpc with neural network in matlab and python. For mfcc features, best validation performance is $5.4181e^{-07}$ at 50 epochs and for lpc features, best validation performance is $7.0034e^{-07}$ at 28 epochs. If epochs are change then peformance is degraded.

Chapter 6

Future Work

In this system mfcc and lpc features are extracted and implement it using neural network in matlab and python. So this features are also implement some different classification algorithm. Large daatset can be used. Some other parameters which are affected also. Parameters are change of length of the sound file, some different features of the sound file can be extracted.

References

[1] "Digital signal processing mini-project: An automatic speaker recognition system." http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/.

[2] M. Schedl, E. Gómez, and J. Urbano, *Music Information Retrieval: Recent Developments and Applications*. now Publishers, 2014.

[3] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, *et al.*, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[4] S. Deshmukh and S. G. Bhirud, "A hybrid selection method of audio descriptors for singer identification in north indian classical music," in *Emerging Trends in Engineering and Technology (ICETET), 2012 Fifth International Conference on*, pp. 224–227, IEEE, 2012.

[5] C. Nithin and J. Cheriyan, "A novel approach to automatic singer identification in duet recordings with background accompaniments," in *Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD), 2014 Annual International Conference on*, pp. 1–6, IEEE, 2014.

[6] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proceedings of the 3rd International Conference on Music Information Retrieval*, vol. 13, p. 17, 2002.

[7] N. C. Maddage, C. Xu, and Y. Wang, "Singer identification based on vocal and instrumental models," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, pp. 375–378, IEEE, 2004.

[8] B. Whitman, G. Flake, and S. Lawrence, "Artist detection in music with minnowmatch," in *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pp. 559–568, IEEE, 2001.

[9] T. Kitahara, M. Goto, and H. G. Okuno, "Category-level identification of non-registered musical instrument sounds," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 4, pp. iv–253, IEEE, 2004.

[10] Y. Takahashi and K. Kondo, "Comparison of two classification methods for musical instrument identification," in *Consumer Electronics (GCCE), 2014 IEEE 3rd Global Conference on*, pp. 67–68, IEEE, 2014.

[11] B. L. Sturm, M. Morvidone, and L. Daudet, "Musical instrument identification using multiscale mel-frequency cepstral coefficients," in *Proc. EUSIPCO*, no. 1, pp. 477–481, 2010.

[12] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *arXiv preprint arXiv:1305.1145*, 2013.

[13] M. Schedl, N. Orio, C. Liem, and G. Peeters, "A professionally annotated and enriched multimodal data set on popular music," in *Proceedings of the 4th ACM Multimedia Systems Conference*, pp. 78–83, ACM, 2013.

[14] D. Black, "Singing voice dataset," 2014.