SPEECH EMOTION RECOGNITION USING TAMIL CORPUS

A PROJECT REPORT

Submitted by

ARUN GOPAL G.
CHRISTY XAVIER RAJ K.

Under the guidance of

Mrs. X. ARPUTHA RATHINA

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE ENGINEERING

B.S.ABDUR RAHMAN
CRESCENT UNIVERSITY
B.S.Abdur Rahman Crescent Institute of Science & Technology
(Estd. u/s 3 of the UGC Act, 1956)

May 2017

B.S.ABDUR RAHMAN
CRESCENT UNIVERSITY
B.S.Abdur Rahman Crescent Institute of Science & Technology
(Estd. u/s 3 of the UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report **"SPEECH EMOTION RECOGNITION USING TAMIL CORPUS"** is the bonafide work of **"ARUN GOPAL G. (130071601012), CHRISTY XAVIER RAJ K. (130071601021)"** who carried out the project work under my supervision. Certified further, that to the best of our knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mrs. X.ARPUTHA RATHINA
SUPERVISOR
Associate Professor
Department of CSE
B. S. Abdur Rahman
Crescent University
Vandalur, Chennai – 600 048

SIGNATURE

Dr.SHARMILA SANKAR
HEAD OF THE DEPARTMENT
Professor & Head
Department of CSE
B. S. Abdur Rahman
Crescent University
Vandalur, Chennai – 600 048

VIVA-VOCE EXAMINATION

The viva-voice examination of the project work titled **"SPEECH EMOTION RECOGNITION USING TAMIL CORPUS"**, submitted by **ARUN GOPAL G. (130071601012)** and **CHRISTY XAVIER RAJ K. (130071601021)** is held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGMENT

We sincerely thank **Sri Prof. Ir. Dr. Sahol Hamid Bin Abu Bakar, Vice Chancellor , B. S. Abdur Rahman Crescent University,** for providing us an environment to carry out our course successfully.

We extend our sincere thanks to **Professor V. Murugesan, Registrar and Director, B.S. Abdur Rahman Crescent University,** for furnishing every essential facility for doing my project.

We thank **Dr. Sharmila Sankar, Head of the Department, Department of Computer Science & Engineering** for his strong support and encouragement throughout our project.

We express deep gratitude to our guide **Mrs. X. Arputha Rathina, Associate Professor, Department of Computer Science & Engineering** for her enthusiastic motivation and continued assistance in the project.

We also extend our sincere thanks to our class advisor **Mrs. A. Radhika, Assistant Professor, Department of Computer Science & Engineering** for her constant support and motivation.

We wish to express our sincere thanks to the project review committee members of the Department of Computer Science and Engineering **Mrs. T. Nagamalar**, Associate Professor, **Mrs. J. Brindha Merin**, Assistant Professor (Sr.Gr.), **Mrs. A. Radhika**, Assistant Professor, for their constant motivation, guidance and support at every stage of this project work.

We thank all the **Faculty members** and the System Staff of Department of Computer Science and engineering for their valuable support and assistance at various stages of project development.

ARUN GOPAL G

CHRISTY XAVIER RAJ K

ABSTRACT

In human machine interaction, automatic speech emotion recognition is yet challenging but important task which paid close attention in current research area. Speech is attractive and effective medium due to its several features expressing attitude and emotions through speech is possible. It is carried out for identification of five basic emotional states of speaker's as anger, happiness, sad, surprise and neutral. Finding the user's emotion can be used for business development and psychological analysis. The motivation of the project is to build the Tamil emotional corpus and to make Tamil emotional corpus available in public domain. Tamil Movies will be used as main resource for building the emotional corpus. Basic emotions like happy, neutral, sad, fear and anger are taken for this analysis purpose. And also for the accuracy purpose from the play both Male and Female Speakers emotional speech have been considered. The observer's perception test result will be used to evaluate annotation of emotion. Tamil emotional speech corpus has been built and Emotion Recognition engine has been constructed using Support Vector Machine (SVM) classifier with the features like MFCC and Fourier Transform.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	-	Support Vector Machine
MFCC	-	Mel Frequency Cepstral Coefficient
FFT	-	Fast Fourier Transform
HMM	-	Hidden Markov Model
GMM	-	Gaussian Mixture Model
FP	-	Fourier Parameters
DSP	-	Digital Signal Processing
LPC	-	Linear Prediction Coding
DCT	-	Discrete Cosine Transform
EMODB	-	Emotional Database
CASIA	-	Chinese Language database
EESDB	-	Chinese Elderly Emotion Database
LPCC	-	Linear Predictive Cepstral Coefficients
MDT	-	Meta Decision Tree
MLP	-	Multilayer Perceptron
EAR	-	Emotion Association Rules
VQ	-	Vector Quantization
DTW	-	Dynamic Time Warping
DES	-	Danish Emotional Speech Database
BES	-	Berlin Emotional Speech Database

1. INTRODUCTION

1.1 OVERVIEW

Historically the sounds of spoken language have been studied at two different levels: (1) phonetic components of spoken words, e.g., vowel and consonant sounds, and (2) acoustic wave patterns. A language can be broken down into a very small number of basic sounds, called phonemes (English has approximately forty). An acoustic wave is a sequence of changing vibration patterns (generally in air), however we are more accustom to “seeing” acoustic waves as their electrical analog on an oscilloscope (time presentation) or spectrum analyzer (frequency presentation). Also seen in sound analysis are two-dimensional patterns called spectrograms, which display frequency (vertical axis) vs. time (horizontal axis) and represent the signal energy as the figure intensity or color. Generally, restricting the flow of air (in the vocal tract) generates that we call consonants. On the other hand modifying the shape of the passages through which the sound waves, produced by the vocal chords, travel generates vowels. The power source for consonants is airflow producing white noise, while the power for vowels is vibrations (rich in overtones) from the vocal chords. The difference in the sound of spoken vowels such as 'A' and 'E' are due to differences in the formant peaks caused by the difference in the shape of your mouth when you produce the sounds.

Henry Sweet is generally credited with starting modern phonetics in 1877 with his publishing of A Handbook of Phonetics. It is said that Sweet was the model for Professor Henry Higgins in the 1916 play, Pygmalion, by George Bernard Shaw. You may remember the story of Professor Higgins and Eliza Doolittle from the musical (and movie) My Fair Lady. The telephone companies studied speech production and recognition in an effort to improve the accuracy of word recognition by humans. Remember nine (N AY N vs. N AY AH N) shown here in one of the “standard” phoneme sets. Telephone operators were taught to

pronounce nine with two syllables as in “onion”. Also, “niner” (N AY N ER) meaning nine is common in military communications. Some work was done, during and right after the war, on speech processing (and recognition) using analog electronics. Digital was not popular let.

These analog processors generally used filter banks to segment the voice spectrum. Operational amplifiers (vacuum tube based), although an available technology, were seldom used. The expense was prohibitive because each amplifier required many tubes at several dollars each. With fairly simple electronics and passive filters, limited success was achieved for (very) small vocabulary systems. Speaker identification / verification systems were also developed.

With the advent of digital signal processing and digital computers we see the beginnings of modern automatic speech recognizers (ASR). A broad range of applications has been developed. The more common command control systems and the popular speech-to-text systems have been seen (if not used) by all of us. Voice recognition, by computer, is used in access control and security systems. An ASR coupled (through a bilingual dictionary) with a text to speech process can be used for automatic spoken language translation. And the list goes on!

1.2 OUTLINE

1.2.1 Survey of Speech Recognition

The general public’s “understanding” of speech recognition comes from such things as the HAL 9000 computer in Stanley Kubrick’s film 2001: A Space Odyssey. Notice that HAL is a perversion of IBM. At the time of the movie’s release (1968) IBM was just getting started with a large speech recognition project that led to a very successful large vocabulary isolated word dictation system and several small voculary control systems.

In the middle nineties IBM'sVoiceType, Dragon Systems' Dragon Dictate, and Kurzweil Applied Intelligence's Voice Plus were the popular personal computer speech recognition products on the market. These "early" packages typically required additional (nonstandard) digital signal processing (DSP) computer hardware. They were about 90% accurate for general dictation and required a short pause between words. They were called discrete speech recognition systems. Today the term isolated word is more common. In 1997 Kurzweil was sold to Lernout & Hauspie (L&H), a large speech and language technology company headquartered in Belgium. L&H is working on speech recognition for possible future Microsoft products. Both IBM and Dragon now have LVCSR systems on the market. The project have IBM Via Voice installed on my computer at home. Once you have used a continuous recognizer, you would not want to go back to "inserting" a pause between each word.

The scan of literature for information about speech recognition the huge scale of the subject overwhelms us. In the technology of speech recognition a number of concepts keep coming up. Generally a speech recognizer includes the following components.

1.2.2 Pre-Processing

The A/D conversion is generally accomplished by digital signal processing hardware on the computer's sound card (a standard feature on most computers today). The typical sampling rate, 8000 samples per second, is adequate. The spoken voice is considered to be 300 to 3000 Hertz.

A sampling rate 8000 gives a Nyquist frequency of 4000 Hertz, which should be adequate for a 3000 Hz voice signal. Some systems have used over sampling plus a sharp cutoff filter to reduce the effect of noise. The sample resolution is the 8 or 16 bits per second that sound cards can accomplish.

1.2.3 Pre-Emphasis Filtering

Because speech has an overall spectral tilt of 5 to 12 dB per octave, a pre emphasis filter of the form $1 - 0.99\ z^{-1}$ is normally used. This first order filter will compensate for the fact that the lower formants contain more energy than the higher. If it weren't for this filter the lower formants would be preferentially modeled with respect to the higher formants.

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

Document describes how to build a simple, yet complete and representative automatic speaker recognition system. Such a speaker recognition system has potential in many security applications. For example, users have to speak a PIN (Personal Identification Number) in order to gain access to the laboratory door, or users have to speak their credit card number over the telephone line to verify their identity. By checking the voice characteristics of the input utterance, using an automatic speaker recognition system similar to the one that we will describe, the system is able to add an extra level of security.

The people can easily identify emotions from speech by observing the speech utterance. Usually speech conveys information about the language being spoken, emotion, gender and generally the identity of the speaker. While speech recognition aims at recognizing the word spoken in speech, the goal of automatic emotion recognition systems is to extract, characterize and recognize the emotions in the speech signal. Some of the emotions are easy to identify by the machine in the way, perceive them but most of them are hard. In case of the

different language people, it's not. In this project, we built a Tamil emotional speech corpus with emotions like happy, anger, fear, neutral and sad. The corpus has been evaluated using SVM based emotion recognition engine.

According to the literature survey, there is no emotion recognition engine built based on Tamil corpus. There is a need to build Tamil corpus to observe the uniqueness of emotion representation in Tamil speech. The corpus has been built based on the acted emotional speech of audio clips. The sources of the audio clips are from various Kollywood movies and it has been categorized into five emotions based on the listeners. The listeners are of two types. They are acoustic and linguistic Listeners. The acoustic listeners classify based on what they perceive from the sound and the linguistic listeners classify the utterance based on acoustic as well as literal usage of the word. Our corpus was categorized by five linguistic listeners.

Initially the features energy and MFCC are extracted from the corpus. Then the extracted features are given as the input to train the SVM. 70% percentages of inputs are given for the training and with the remaining 30% are used for testing. After the machine get trained based upon the input type, it will classify test utterance into the nearest matching emotion classes.

2. LITERATURE SURVEY

Kunxiawang et al.,[1] have performed studies on harmony features for speech emotion recognition. It was found that the first- and second-order differences of harmony features also play an important role in speech emotion recognition. Therefore, a new Fourier parameter model using the perceptual content of voice quality and the first- and second-order differences for speaker-independent speech emotion recognition. Experimental results show that the proposed Fourier parameter (FP) features were effective in identifying various emotional states in speech signals. They improve the recognition rates over the methods using Mel Frequency Cepstral Coefficient (MFCC) features by 16.2, 6.8and 16.6 points on the German database (EMODB), Chinese language database(CASIA) and Chinese elderly emotion database (EESDB). In particular, when combining FP with MFCC, the recognition rates can be further improved on the aforementioned databases by 17.5, 10 and 10.5 points, respectively.

Md. Touseef Sumer et al., [2] studied formant frequencies for the emotion detection. Here they had taken three formant frequencies f1, f2, f3 and for different vowels the range of f1 lies between 270 to 730Hz, f2 and f3 lies between 840 to 2290HZ and 1690 to 3010Hz respectively. These frequencies are important for analysis of emotion of person. Linear predictive coding technique has been used for estimation of formant frequencies. With the formant frequencies pitch features are also used for detection of emotion. KLD and GMM is used for further process of emotion detection.

Vidhyasaharan Sethu et al., [3] used frame based features for emotion detection, in this paper temporal contours of parameters like glottal source parameter which is extracted from three component model of speech production is use as a feature for automatic emotion detection of speech. Then automatic classification system for emotion detection is used with front end and back end.

Biswajit Nayak et al., [4] extracted Mel Frequency Cepstral Coefficients (MFCC) features for emotion detection. Here eight different speakers and IITKGP-SEHSC emotional speech corpora are used for emotion detection. And classification is carried out by using GMM. It was observed the number of centers of centered GMM increases the emotion recognition performance increases.

Akshay S. Utane and Dr. S.L .Nalbalwar [5] used Mel Frequency Cepstrum Coefficient (MFCC), linear predictive cepstral coefficients (LPCC) and energy features for the emotion detection of speech. Here GMM and HMM is used as classifier for emotion detection of speech. It was observed that both the classifier methods provide relatively similar accuracy. The efficiency of emotion recognition system highly depends on database selection, so it is very necessary to select proper database.

Stavros Ntalampiras and Nikos Fakotakis [6] combined two feature sets of heterogeneous domain such as baseline set and feature based on multi resolution analysis. The first set includes Mel filter bank, pitch, and harmonic to noise ratio and second set includes wavelet packets. After extracting these features, feature integration methods like short – term statistics, spectral moments and autoregressive model are used. Then emotion of the speech is detected by doing the fusion of feature level fusion, fusion of log likelihoods which are produced by temporally integrated feature sets and fusion of temporal integration method.

Chung-Hsien Wu and Wei-Bin Liang [7] read acoustic prosodic information and semantic labels for the emotion detection of speech. For acoustic prosodic information detection, acoustic and prosodic features like spectrum, formant and pitch are extracted from input speech. For this three types of base level classifier models GMM, SVM (support vector machine), and MLP (multilayer perceptron) were used and lastly the Meta Decision Tree (MDT) is used for classifier fusion.

For SL based detection semantic labels derived from an existing Chinese knowledge base, HowNet are used to extract Emotion Association Rules (EAR) from detected word sequence of speech. Maximum entropy model is then used to explain the relationship between emotional states and EARs for emotion detection.

Yuan Yujin et al., [8] used speaker recognition in our lives is an important branch of authenticating automatically a speaker's identity based on human biological feature. Linear Prediction Cepstrum Coefficient (LPCC) and Mel Frequency Cepstrum Coefficient (MFCC) are used as the features for text-independent speaker recognition in this system. And the experiments compare the recognition rate of LPCC, MFCC or the combination of LPCC and MFCC through using Vector Quantization (VQ) and Dynamic Time Warping (DTW) to recognize a speaker's identity. It proves that the combination of LPCC and MFCC has a higher recognition rate.

Carlos Bussoet al., [9] considered pitch features or features of fundamental frequency for the emotion detection of speech. The mean, standard deviation, range, minimum, maximum, median, lower quartile, upper quartile, interquartile range, kurtosis, skewness, slope, curvature and inflexion all these statistics of pitch contour and derivative of pitch contour are taken for emotion detection. Then these statistics are grouped into sentence level and voiced level features, which are further used for emotion detection. After that these characteristics of emotional speech is compared with the characteristics of neutral speech by using KLD. Nested logistic regression models are to quantify the emotionally discriminative power of pitch feature. The results indicate that the pitch contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the pitch shape. Also analyzing the pitch statistics at the sentence level is found to be more accurate and robust than analyzing the pitch statistics for voiced regions.

Daniel Neiberg et al., [10] combined MFCC, MFCC-low and variant features for the emotion detection. MFCCs are extracted using pre-emphasized audio, using 25.6ms hamming window at every 10ms. For each frame 24, FFT based Mel warped logarithmic filter bank are placed in 300 to 3400Hz. For MFCC-low filter bank is placed in 20-300Hz. Variant features such as pitch and derivative are used for emotion detection of speech signal. GMM is used as classifier for emotion detection.

Mohammed E. Hoqueet al.,[11] considered prosodic features like pitch, energy, formants and acoustic features to extract the intonation patterns and correlates of emotion from speech samples for the emotion detection. To improve the performance features were used on word level emotional utterances. Here the classifiers from WEKA tools are used for emotion detection.

Chul Min Lee and Shrikanth S. Narayanan [12] combined three sources of information namely acoustic, lexical, and discourse for emotion detection. To capture emotion information at the language level, an information-theoretic notion of emotional salience is introduced. Optimization of the acoustic correlates of emotion with respect to classification error was accomplished by investigating different feature sets obtained from feature selection, followed by principal component analysis. The results show that, the best results are obtained when acoustic and language information are combined**.** And also combining all the information improves emotion classification by 40.7% for males and 36.4% for females.

3. PROBLEM DEFINITION AND METHODOLOGIES

3.1 PROBLEM DEFINITION

A speech signal is naturally occurring signal and hence is random in nature. The signal expresses different ideas, communication and hence has lot of information. There are number of automatic speech detection system and music synthesizer commercially available. However despite significant progress in this area there still remain many things which are not well understood. Detection of emotions from speech is such an area. The speech signal information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people. Detection of human emotions will improve communication between human and machine. The human instinct detects emotions by observing psycho-visual appearances and voices. Machines may not fully take human place but still are not behind to replicate this human ability if speech emotion detection is employed. Also it could be used to make the computer act according to actual human emotions.

This is useful in various real life applications as systems for real life emotion detection using corpus of agent client spoken dialogues from call centre like for medical emergency, security, prosody generation, etc. The alternative emotion detection is through body, face signals, and bio signals such as ECG, EEG. However in certain real life applications these methods are very complex and sometimes impossible, hence emotion detection from speech signals is the more feasible option. Good results are obtained by the signal processing tools like MATLB and various algorithms (HMM, SVM) but their performance has limitations, while combination and ensemble of classifiers could represent a new step towards better emotion detection.

3.2 SPEECH EMOTION RECOGNITION

In general, emotion detection system consist of speech normalization, feature extraction, feature selection, classification and then the emotion is detected. First noise and dc components are removed in speech normalization then the feature extraction and selection is carried out. The most important part in further processing of input speech signal to detect emotions is extraction and selection of features from speech. The speech features are usually derived from analysis of speech signal in both time as well as frequency domain. Then the data base is generated for training and testing of the extracted speech features from input speech signal. In the last stage emotions are detected by the classifiers. Various pattern recognition algorithms (HMM, GMM) are used in classifier to detect the emotion.

3.2.1 Pre-Processing

The collected emotional data usually gets degraded due to external noise (background and “hiss” of the recording machine). This will make the feature extraction and classification less accurate. Hence normalization is critical step in emotion detection. In this preprocessing stage speaker and recording variability is eliminated while keeping the emotional discrimination. Generally two types of normalization techniques are performed they are energy normalization and pitch normalization.

3.2.2 Feature Extraction and Selection from Emotional Speech

After normalization of emotional speech signal, it is divided into segments to form their meaningful units. Generally these units represent emotion in a speech signal. The next step is the extraction of relevant features. These emotional speech features can be classified into different categories. One classification is long term features and short term features. The short term

features are the short time period characteristics like formants, pitch and energy. And long term features are the statistical approach to digitized speech signal. Some of the frequently used long term features are mean and standard deviation. The larger the feature used the more improved will be the classification process. After extraction of speech features only those features which have relevant emotion information are selected. These features are then represented into n- dimensional feature vectors [10]. The prosodic features like pitch, intensity, speaking rate and variance are important to identify the different types of emotions from speech. In Table 1 acoustic characteristics of various emotions of speech is given. The observations which are expressed in below table 1 are taken by using MATLAB.

Table 3.1 Observations for different speech emotions

Characteristics / **Emotion**	**Happy**	**Anger**	**Neutral**	**Fear**	**Sad**
Pitch Mean	High	Very High	High	Very High	Very High
Pitch Range	High	High	High	High	High
Pitch Variance	High	Very High	High	Very High	Very High
Pitch Contour	Incline	Decline	Moderate	Incline	Incline
Speaking Rate	High	High	Medium	High	High

- **Prosodic and Acoustic Features**

Prosodic and acoustic features of emotion Prosody have many other functions than emotional signaling. Intonations and speech rhythm changes can indicate other common speech related functions. The stressing of words or syllables for emphasis or other conversationally relevant functions such as turn signaling in conversations are often used regardless of any emotional state. Murray et al. (1996) conclude that any emotional changes to prosody must be seen in addition to the underlying normal prosodic processes. Also, acoustic features do not specify emotional information only.

Even simpler acoustic contours such as ZCR, HNR, or STE measures are very sensitive to different speech properties that are not related to emotion. More sophisticated spectral feature contours such as MFCC or PLP features capture also much, or all, of the normal speech processes. 41 Research to find emotionally relevant prosodic components and signal features have been traditionally conducted by performing perception tests or by manually inspecting the prosodic parameters of real speech recordings to guess what kind of changes are emotionally induced (Scherer 2003).

Access to modern computational resources has also made systematic data mining approaches more effective with the capability to search through vast amounts of suspected or even randomly generated acoustic and prosody derived candidate features to find emotional correlates, e.g. by Oudeyer (2002), Batliner et al. (1999). A more fundamental approach is to model emotional speech itself and with synthesis techniques to produce candidate samples using different model parameters.

The synthesised samples are then used in perception tests to identify which parameter changes convey emotional signals (Murray & Arnott 1995, Murray et al. 1996, Schröder et al. 2001, Schröder 2001). For a review of emotionally relevant features and extraction techniques, see Cowie & Cornelius (2003) or Ververidis & Kotropoulos (2006). It is generally accepted that the most contributing factor for emotional speech is prosody through the fundamental frequency (F0) and variations in the F0 contour. Significant differences in F0 contours have been observed, especially between basic emotions (Banse & Scherer 1996, Paeschke & Sendelmeier 2000, Toivanen 2001). Other important prosodic feature categories identified are the variations in energy (intensity), variations in duration, and variations in speech quality.

Energy statistics are directly related to the perceived activation level of emotions (Banse & Scherer 1996, Ehrette et al. 2002). The timing and duration of the different parts of an utterance as well as the overall speech rate are all

emotionally sensitive features (Murray et al. 1996, Thymé-Gobbel & Hutchins 1999, Farinas & Pellegrino 2001, Bosch 2003). It has been further noted that pause lengths are different between read and spontaneous speech. The speech rate has an effect on articulation, e.g. segment deletions, that can be detected using formant and spectral analysis (Kienast & Sendlmeier 2000).

Using synthesis techniques, voice quality changes have been shown to have a significant supporting role in the signalling of emotion, particularly among milder affective states (Gobl & Ní Chasaide 2003, Grichkovtsova et al. 2012). A summary of commonly associated emotion effects in relation to normal speech is shown in Table 3, adapted from Scherer (1986), Murray & Arnott (1993), and Nwe et al. (2003).

3.2.3 Database for Training and Testing

The database is used for training, testing and development of feature vector. A good database is important for desired result. Various databases are available created by speech processing community. The databases can be divided into training data set and testing data set. The famous databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database, TIMIT. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

The databases that are used in SER are classified into 3 types. Type 1 is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion, e.g. DES, EMO-DB. Type 2 is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come from real-life applications for example call-centers.Type 3 is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where

emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated.

3.2.4 Classifiers to Detect Emotions

Various classifiers like GMM HMM are used according to their specific usage based on selected features. Emotions are predicated using classifiers and selected feature vectors to predict emotion from training data set and the development data set. For the training data sets the emotion information are known whereas for testing data set the emotion information are unknown. When performing analysis of complex data one of the major problems comes from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still the data with sufficient accuracy.

Typically, in speech recognition, we divide speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage.

4. SYSTEM DESIGN

The system design specifies the hardware and software requirements.

4.1 SYSTEM REQUIREMENTS

The specifications required for the system regarding the software and hardware aspects are described below.

4.1.1 Hardware requirements

The hardware requirements of the system are given below.

RAM	-	1 GB
Processor	-	Any Intel or AMD x86 processor supporting SSE2 instruction set
Hard Disk	-	1 GB for MATLAB only, 3–4 GB for a typical installation

4.1.2 Software requirements

The software requirements of the system are given below.

Operating System	-	Windows , Mac, Linux
Language	-	MATLAB
Tools	-	Median Filter, Audacity

4.2 SOFTWARE REQUIREMENT DESCRIPTION

It is a multi-paradigm numerical computing environment and fourth-generation programming language. A proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user

interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python.

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to abilities. An additional package, Simulink, adds graphical multi-domain simulation and model-based design for dynamic and embedded systems.

4.2.1 Overview of MATLAB

Matlab (matrix laboratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization and programming. It Allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications.

It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods. It is used in every facet of computational mathematics. Following are some commonly used mathematical calculations where it is used most commonly

- Dealing with Matrices and Arrays
- 2-D and 3-D Plotting and graphics
- Linear Algebra
- Algebraic Equations
- Non-linear Functions
- Statistics
- Data Analysis
- Calculus and Differential Equations
- Numerical Calculations
- Integration

- Transforms
- Curve Fitting
- Various other special functions

4.2.2 Features of MATLAB

- It is a high-level language for numerical computation, visualization and application development.
- It also provides an interactive environment for iterative exploration, design and problem solving.
- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality maintainability and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

4.2.3 MATLAB Environment

When MATLAB gets started, the MATLAB desktop appears, containing tools (graphical user interfaces) for managing files, variables, and applications associated with MATLAB. The first time MATLAB starts, the desktop appears as shown in the following illustration, although your Launch Pad may contain different entries. All the desktop tools provide common features such as context menus and keyboard shortcuts.

The tools are:

- Current Directory Browser
- Workspace Browser
- Array Editor
- Editor/Debugger
- Command Window
- Command History
- Launch Pad
- Help Browser

- **Command Window**

Use the Command Window to enter variables and run functions and M-files.

- **Command History**

Lines you enter in the Command Window are logged in the Command History window. In the Command History, you can view previously used functions, and copy and execute selected lines. To save the input and output from a MATLAB session to a file, use the diary function.

- **Running External Programs**

The exclamation point character! is a shell escape and indicates that the rest of the input line is a command to the operating system. This is useful for invoking utilities or running other programs without quitting MATLAB. On Linux, for example,!emacs magik.m invokes an editor called emacs for a file named magik.m. When you quit the external program, the operating system returns control to MATLAB.

- **Launch Pad**

MATLAB's Launch Pad provides easy access to tools, demos, and documentation.

- **Help Browser**

Use the Help browser to search and view documentation for all your Math Works products. The Help browser is a Web browser integrated into the MATLAB desktop that displays HTML documents.

To open the Help browser, click the help button in the toolbar, or type help browser in the Command Window. The Help browser consists of two panes, the Help Navigator, which you use to find information, and the display pane, where you view the information.

- **Help Navigator -** Use to Help Navigator to find information.
- **Product filter** - Set the filter to show documentation only for the products you specify.
- **Contents tab** - View the titles and tables of contents of documentation for your products.
- **Index tab** - Find specific index entries (selected keywords) in the MathWorks documentation for your products.
- **Search tab** - Look for a specific phrase in the documentation. To get help for a specific function, set the Search type to Function Name.
- **Favouritesta** -View a list of documents you previously designated as favorites.
- **Browse to other pages** - Use the arrows at the tops and bottoms of the pages, or use the back and forward buttons in the toolbar.
- **Bookmark pages** - Click the Add to Favorites button in the toolbar.
- **Print pages** - Click the print button in the toolbar.
- **Find a term in the page** - Type a term in the Find in page field in the toolbar and click Go.

4.2.4 Uses of MATLAB

It is widely used as a computational tool in science and engineering encompassing the fields of physics, chemistry, math and all engineering streams. It is used in a range of applications including -

- Signal Processing and Communications
- Image and Video Processing
- Control Systems
- Test and Measurement
- Computational Finance
- Computational Biology

4.2.5 Median Filter

In signal processing, it is often desirable to be able to perform some kind of noise reduction on an image or signal. The median filter is a nonlinear digital filtering technique, often used to remove noise. Such noise reduction is a typical pre-processing step to improve the results of later processing (for example, edge detection on an image). Median filtering is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise.

The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighboring entries. The pattern of neighbors is called the "window", which slides, entry by entry, over the entire signal. For 1D signals, the most obvious window is just the first few preceding and following entries, whereas for 2D (or higher-dimensional) signals such as images, more complex window patterns are possible (such as "box" or "cross" patterns). Note that if the window has an odd number of entries, then the median is simple to define: it is just the middle value after all the entries in the window are sorted numerically. For an even number of entries, there is more than one possible median, see median for more details.

4.3 ARCHITECTURE DIAGRAM

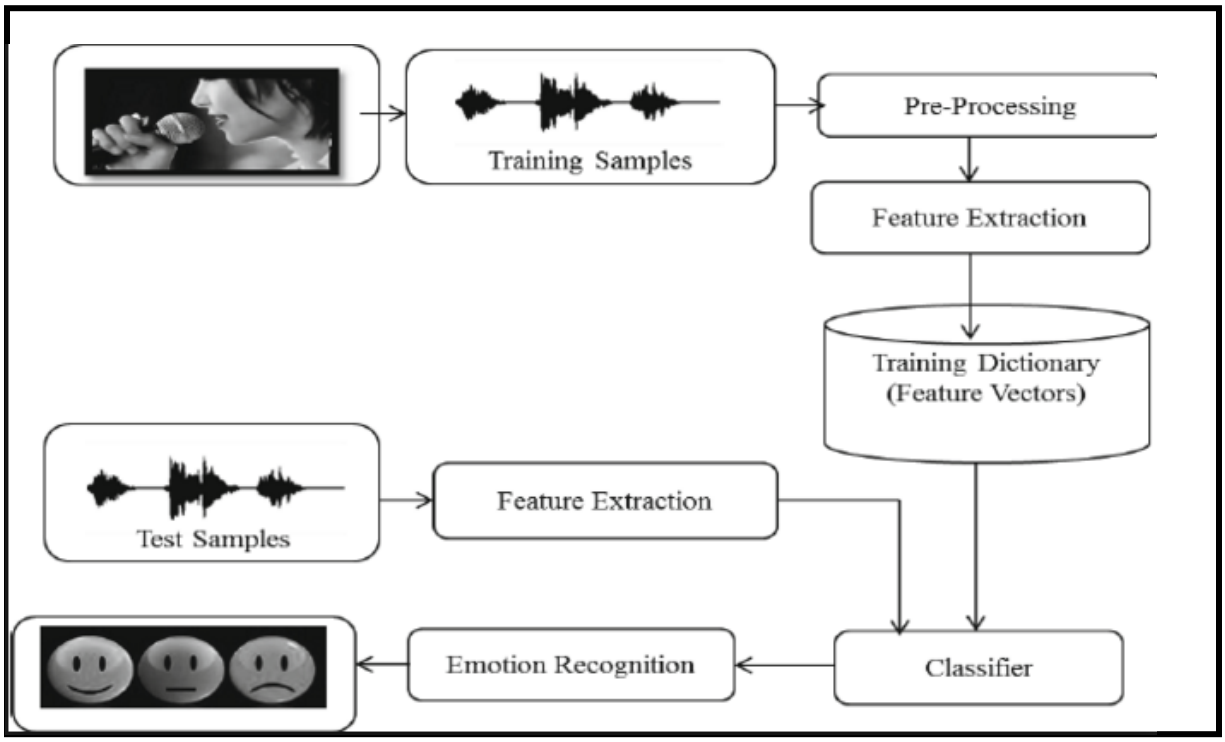


Figure 4.1 Architecture Diagram

5. IMPLEMENTATION

Emotion detection system consists of collection of audio clips, speech normalization, feature extraction, feature selection, classification and identification.

First noise and dc components are removed in speech normalization then the feature extraction and selection is carried out. The most important part in further processing of input speech signal to detect emotions is extraction and selection of features from speech. The speech features are usually derived from analysis of speech signal in both times as well as frequency domain. Then the data base is generated for training and testing of the extracted speech features from input speech signal. In the last stage emotions are detected by the classifiers. Support Vector Machine (SVM) is used as a classifier to detect the emotion.

5.1 FUNCTIONAL DESCRIPTION OF MODULES

Various modules in emotion detection are as follows.

- Creation of Emotional Database
- Speech Normalization
- Feature Extraction and Selection from Emotional Speech
- Database for Training, Testing
- Classifiers to Detect Emotions

5.1.1 CREATION OF EMOTIONAL DATABASE

The acted speech corpus is collected from various kollywood movies. These audio clippings have been categorized into five emotions based on the listeners. The listeners are of two types. They are acoustic and linguistic listeners. The acoustic listeners classify based on what they perceive and the linguistic listeners classify based on the literal usage of the word. Our corpus was categorized by five listeners. Still the number of listeners can be increased to improve the efficiency. The emotions categorized in this paper are happy, sad,

anger, fear, and neural. Each emotion is further classified into testing and training data's. The database consists of 100 audio clippings for each emotion. The database consists of acted speech of both male and female. The database with various emotions are collected and is analyzed for emotions and labeled accordingly .The emotions are categorized into various emotional groups namely happy, sad, anger, fear, neural etc. Any segment that lack in clarity in either the language content or the signal strength is ruled out. The classifications have to be evaluated by many listeners. The evaluation is done by two types of listeners, namely acoustic and linguistic listeners. The acoustic listeners observe the sound by which they perceive and linguistic listeners who observe them based on literal usage of words in the play as well as acoustical behavior.

5.1.2 SPEECH NORMALIZATION

The collected emotional data usually gets degraded due to external noise (background music and other sounds). This will make the feature extraction and classification less accurate. Hence normalization is critical step in emotion detection. In this preprocessing stage speaker and recording variability is eliminated while keeping the emotional discrimination. Generally two types of normalization techniques are performed they are energy normalization and pitch normalization.

Sampling can be done for functions varying in space, time, or any other dimension, and similar results are obtained in two or more dimensions.

For functions that vary with time, let s(t) be a continuous function (or "signal") to be sampled, and let sampling be performed by measuring the value of the continuous function every T seconds, which is called the sampling interval. Thus, the sampled function is given by the sequence**:**

s(nT), for integer values of n.

The sampling frequency or sampling rate f_s is defined as the number of samples obtained in one second (samples per second), thus $f_s = 1/T$.

Reconstructing a continuous function from samples is done by interpolation algorithms. The Whittaker–Shannon interpolation formula is mathematically equivalent to an ideal low pass filter whose input is a sequence of Dirac delta functions that are modulated (multiplied) by the sample values. When the time interval between adjacent samples is a constant (T), the sequence of delta functions is called a Dirac comb. Mathematically, the modulated Dirac comb is equivalent to the product of the comb function with s(t). That purely mathematical function is often loosely referred to as the sampled signal.

Most sampled signals are not simply stored and reconstructed. But the fidelity of a theoretical reconstruction is a customary measure of the effectiveness of sampling. That fidelity is reduced when s(t) contains frequency components higher than $\mathbf{f_s}/2$ Hz, which is known as the Nyquist frequency of the sampler. Therefore s(t) is usually the output of a low pass filter, functionally known as an "anti-aliasing" filter. Without an anti-aliasing filter, frequencies higher than the Nyquist frequency will influence the samples in a way that is misinterpreted by the interpolation process.

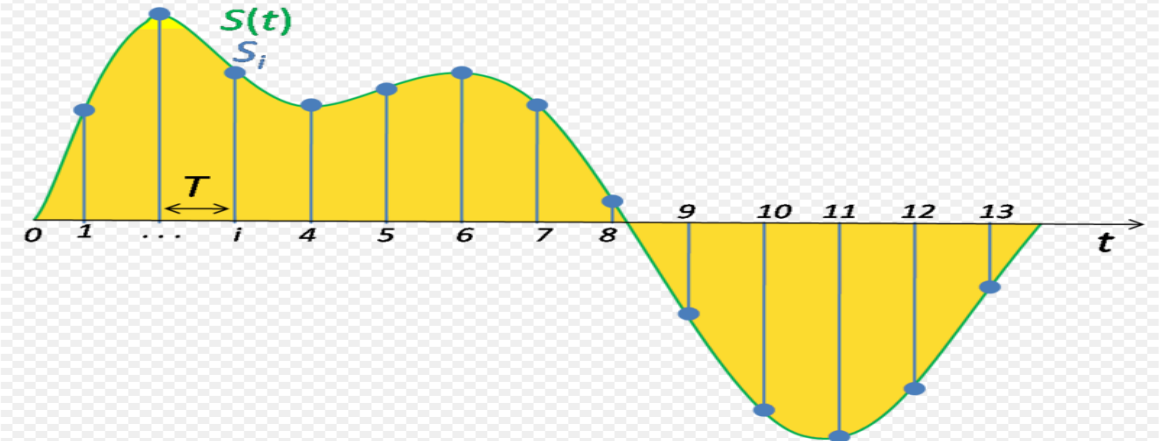


Figure 5.1 Sampling of signal

5.1.3 FEATURE EXTRACTION AND SELECTION FROM EMOTIONAL SPEECH

After normalization of emotional speech signal, it is divided into segments to form their meaningful units. Generally these units represent emotion in a speech signal. The next step is the extraction of relevant features. These emotional speech features can be classified into different categories. One classification is long term features and short term features.The short term features are the short time period characteristics like formants, pitch and energy. And long term features are the statistical approach to digitized speech signal. Some of the frequently used long term features are mean and standard deviation. The larger the feature used the more improved will be the classification process.

After extraction of speech features only those features which have relevant emotion information are selected. These features are then represented into n- dimensional feature vectors. The prosodic features like pitch, intensity, speaking rate and variance are important to identify the different types of emotions from speech. The purpose of speech feature extraction is to convert the speech waveform, using digital signal processing (DSP) tools, to a set of features (at a considerably lower information rate) for further analysis. This is often referred as the signal-processing front end.

The speech signal is a slowly timed varying signal (it is called quasi-stationary). An example of speech signal is shown in Figure 2. When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal.

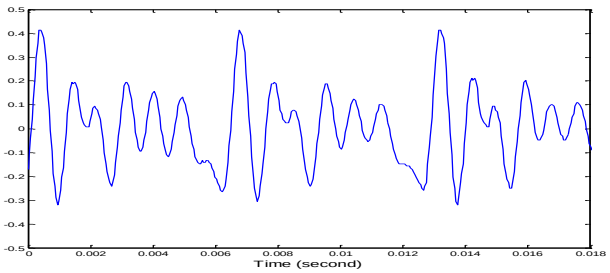


Figure 5.2 Example of speech signal

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and will be described in this paper.

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the mel-frequency scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The process of computing MFCCs is described in more detail next.

Speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Therefore proper choice of feature vectors is one of the most important tasks. There are many approaches towards automatic recognition of emotion in speech by using different feature vectors. Feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated. Over the entire length of the utterance, while the short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently. Short time features uses interrogative phrases which has wider pitch contour and a larger pitch standard deviation. Most common features used by researchers are: The value of pitch frequency can be calculated in each speech

frame and the statistics of pitch can be obtained in the whole speech sample. These statistical values reflect the global properties of characteristic parameters. Emotional contents of a utterance is strongly related with its voice quality. The voice quality can be numerically represented by parameters estimated directly from speech signal. The acoustic parameters related to speech quality are: (1) Voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level; (2) voice pitch; (3) phrase, phoneme, word and feature boundaries; (4) temporal structures.

Peak Value Detection:-

Maximum Stem on Input Voice Signal (Low Pass Signal) using “max” command process on Feature_Peak=max(low_pass_signal)

Minimum Value Detection:-

Minimum Stem on Input Voice Signal (Low Pass Signal) using “min” command process on

Feature_Minimum=min(low_pass_signal)

Mean (Average) Value Detection:-

Minimum Stem on Input Voice Signal (Low Pass Signal) using “min” command process on

Feature_Average=mean(low_pass_signal)

- **Mel-Frequency Cepstrum Coefficients Processor**

A block diagram of the structure of an MFCC processor is given in Figure 5.3. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to

mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

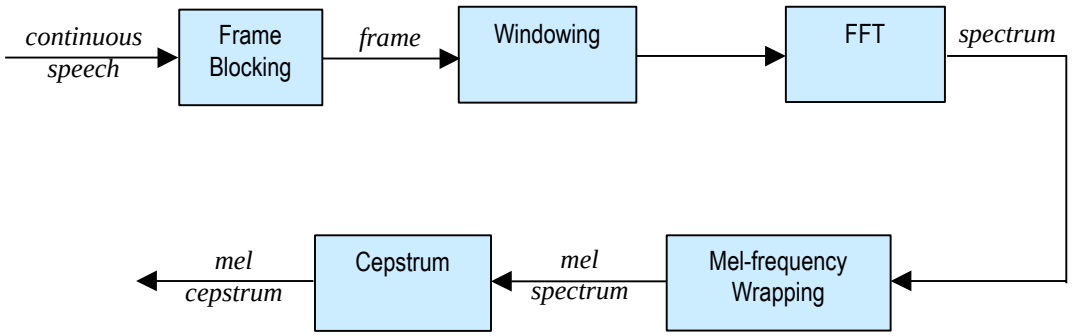


Figure 5.3 Block diagram of the MFCC processor

- **Frame Blocking**

In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

- **Windowing**

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal

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

Typically the Hamming window is used, which has the form:

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

- **Fast Fourier Transform**

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1$$

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, $F_s$denotes the sampling frequency.The result after this step is often referred to as spectrum or periodogram.

- **Mel – Frequency Wrapping**

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

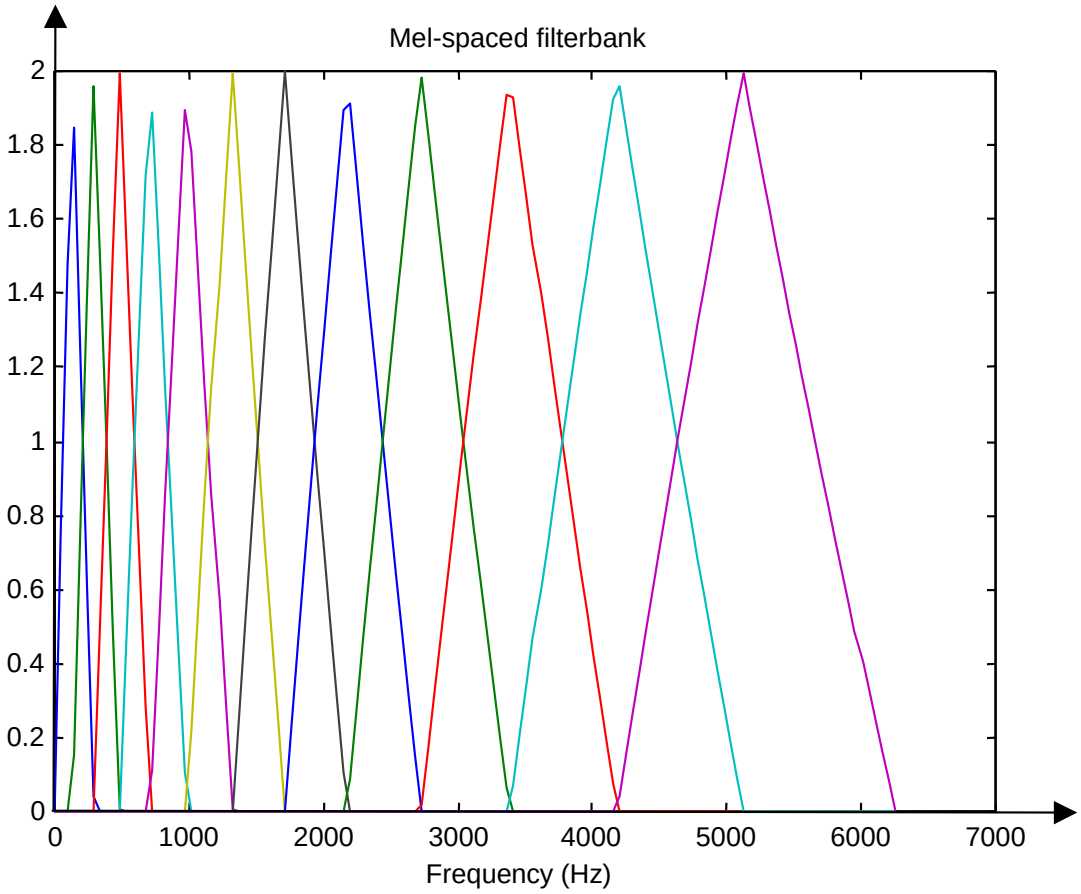


Figure 5.4 An example of mel-spaced filterbank

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel-scale (see Figure 4). That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, K, is typically chosen as 20. Note that this filter bank is applied in the frequency domain, thus it simply amounts to applying the triangle-shape windows as in the Figure 4 to the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

- **Cepstrum**

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,...,K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad n = 0,1,...,K\text{-}1$$

Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information.

By applying the procedure described above, for each speech frame of around 30msec with overlap, a set of mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. This set of coefficients is called an acoustic vector. Therefore each input utterance is transformed into a sequence of acoustic vectors. In the next section we will see how those acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.

5.1.4 DATABASE FOR TRAINING AND TESTING

The database created is used for training, testing and development of feature vector. A good database is important for desired result. Various databases are available created by speech processing community. The databases can be divided into training data set and testing data set. The famous

databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database, TIMIT. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

The database used in SER is classified into 3 types. Type 1 is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion, e.g. DES, EMO-DB. Type 2 is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come from real-life applications for example call-centers. Type 3 is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated.

Emotional databases have evolved much during the last decade. Older studies on emotion recognition have relied almost exclusively on small and heterogeneous databases collected by the researchers for their own personal use, for a review, see Ververidis & Kotropoulos (2006). Typically, the contents and used annotations in the small databases have a large degree of variance making the comparison or fusion of databases hard. Douglas-Cowie et al. (2003) presented a guideline for future databases to address these issues and guide the database collection processes. The main considerations for a database development have been identified to be as follows: scope, naturalness, context, descriptors and accessibility. For a more thorough discussion about the identified considerations, see Douglas-Cowie et al. (2003). The more modern databases have not only adopted the presented guidelines, but also incorporated many advances in emotional models, e.g. dimensional model annotations. Furthermore, the modern databases have logically included the multi-modal aspects of emotion. Thus, the more modern larger databases, e.g. the HUMAINE Database (Douglas-Cowie et al. 2007), or MAHNOB-HCI (Soleymani et al. 2012),

have begun including not only speech, but also a full range of multi-modal signals such as audio- 44 visual recordings from audio-visual stimuli, gestures, and physiological biosignals (i.e. ECG, skin conductance, respiration, eye tracking, etc.). The database used in this thesis is of the older variety, limiting its usability in the future, due to the fact that Finnish emotional speech data has not yet been extensively collected using the more modern collection guidelines and multimodal scope. Nevertheless, the MediaTeam Emotional Speech Corpus (Seppänen et al. 2003) is still currently the largest corpus of emotional Finnish speech.

5.1.5 CLASSIFIERS TO DETECT EMOTIONS

A classifier is a key element of machine learning. A multitude of classifiers has been developed. Linear classifiers, e.g. Naive Bayesian (NB), Linear Discriminant Classifier (LDC), or Perceptron (Duda et al. 2001), are typical examples of classic statistical classification methods. Further advances in statistical classification have made nonlinear classification methods available. Typical methods capable of nonlinear classification are k-Nearest Neighbour (kNN) and various advanced Neural Network (NN) methods.

The linear classifiers can also be extended to nonlinear classification, e.g. using the kernel trick (Aizerman et al. 1964). Using the kernel trick, a linear Support Vector Machine (SVM) (Vapnik & Lerner 1963) is generalized to nonlinear classification (Boser et al. 1992) making SVM a very effective framework for classification. Many classifiers have been found effective in emotion recognition from speech, typically using mostly prosodic features and a label based class approach (typically 4–6 basic emotions). LDC approaches have been found effective with feature selection (McGilloway et al. 2000, Lee et al. 2001).

Classification And Regression Trees (CART) performed well for the classification of frustration in Ang et al. (2002). Oudeyer (2002) found the CART approach also efficient when combined with a meta-optimisation method. Simple

NB classifiers exhibited poor performance, but when combined with a good feature selection method, good performance was reached (Dellaert et al. 1996, Oudeyer 2002). Various kNN classifiers were all found performing well when a suitable feature selection method was utilised (Dellaert et al. 1996, Lee et al. 2001, Yu et al. 2001, Oudeyer 2002). The NN type of classifiers has not been found effective. In McGilloway et al. (2000), a generative vector quantisation approach led to over learning problems, which resulted in a poor performance. In Oudeyer (2002), radial basis function neural networks and voting perceptron type of classifier setups did not perform well, either.

Linear SVMs were found performing poorly, but when using polynomial kernels, a good performance was observed (McGilloway et al. 2000, Oudeyer 2002). In Nwe et al. (2003), a Hidden Markov Model (HMM) based solution was found very effective. It can be seen from the literature that emotion recognition from speech has the 50 typical dualistic solution property common in classification tasks. A simple classifier can be effective when the features are selected and the input nonlinearities are handled first with a more robust method (e.g. manifold modelling or nonlinear transformation).

Another solution is to use a classifier that is particularly well suited for the problem. For example, one can use an SVM that has the feature space problem solved by using the correct kernel or use a solution that is intrinsically well aligned with the information structure of the input features, e.g. an HMM that robustly incorporates the segmental nature of speech.

SVM is used as a classifier according to their specific usage based on selected features. Emotions are predicted using classifiers and selected feature vectors to predict emotion from training data set and the development data set.

For the training data sets the emotion information are known whereas for testing data set the emotion information are unknown. When performing analysis of complex data one of the major problems comes from the number of variables

involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still the data with sufficient accuracy.

Typically, in speech recognition, we divide speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage.

Emotion recognition from speech using pattern recognition techniques is a relatively recent approach that has become possible with the increase in computational resources. Traditional label based classification solutions have been developed using multiple state-of-the-art classifier techniques. Reasonably good results around 50%–80% correct classification for 4–6 basic emotions (e.g. neutral, sad, angry, happy, fear, disgust) have been attained (Dellaert et al. 1996, McGilloway et al. 2000, Oudeyer 2002, Bosch 2003). Label based classification attempts are, however, prone to semantic confusion in the truth data and other problems in data collection (e.g. methodological issues in obtaining some types of emotion) (Douglas-Cowie et al. 2003) limiting their performance. More robust approaches using dimensional models of emotion, e.g. Zeng et al. (2005), Nicolaou et al. (2011), are very recent. For a state-of-the-art review of emotion recognition and classification techniques from traditional approaches to recent advances, see Cowie et al. (2001), Ververidis & Kotropoulos (2006), Zeng et al. (2009), and/or Calvo & D'Mello (2010).

- **SVM CLASSIFIER**

Sampling can be done for functions varying in space, time, or any other dimension, and similar results are obtained in two or more dimensions.

For functions that vary with time, let s(t) be a continuous function (or "signal") to be sampled, and let sampling be performed by measuring the value of the continuous function every T seconds, which is called the sampling interval. Thus, the sampled function is given by the sequence**:**

S(nT), for integer values of n.

The sampling frequency or sampling rate f_s is defined as the number of samples obtained in one second (samples per second), thus $f_s = 1/T$.

Reconstructing a continuous function from samples is done by interpolation algorithms. The Whittaker–Shannon interpolation formula is mathematically equivalent to an ideal low pass filter whose input is a sequence of Dirac delta functions that are modulated (multiplied) by the sample values. When the time interval between adjacent samples is a constant (T), the sequence of delta functions is called a Dirac comb. Mathematically, the modulated Dirac comb is equivalent to the product of the comb function with s(t). That purely mathematical function is often loosely referred to as the sampled signal.

Most sampled signals are not simply stored and reconstructed. But the fidelity of a theoretical reconstruction is a customary measure of the effectiveness of sampling. That fidelity is reduced when s(t) contains frequency components higher than $f_s/2$ Hz, which is known as the Nyquist frequency of the sampler. Therefore s(t) is usually the output of a low pass filter, functionally known as an "anti-aliasing" filter. Without an anti-aliasing filter, frequencies higher than the Nyquist frequency will influence the samples in a way that is misinterpreted by the interpolation process.

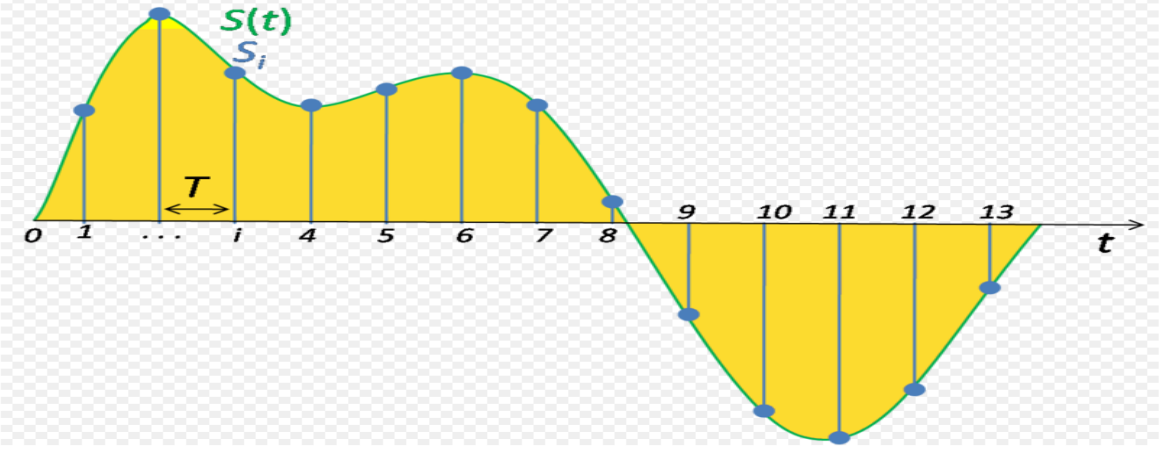


Figure 5.5 Sampling of signal

- **Feature selection and transformations**

Feature selection and transformation techniques are used to find, or remodel, features that are relevant to the pattern recognition problem. In selection techniques, to improve model performance, the relevant and efficient features are kept while the irrelevant and noisy features are excluded. Feature selection is a typical way to implement supervised learning. Optimal feature selection can be performed by brute force, i.e. an exhaustive search of all alternatives. Methods using cost functions that guarantee global optimum, e.g. Branch and Bound style methods, can also be implemented to produce optimal solutions at the expense of computational complexity (Duda et al. 2001). For practical applications, however, optimal solutions are often computationally too inefficient and suboptimal search strategies, i.e. heuristics, are used instead. Many methods have been attempted in emotion recognition research.

Simple sequential selection algorithms, e.g. Promising First Selection (PFS) or Sequential Forward Search (SFS) (Dellaert et al. 1996, Lee et al. 2001), have been found effective. Very similar Sequential Backward Search (SBS) (Oudeyer 2002) has also been used successfully. A more generalized Sequential Forward Floating Search (SFFS) (Pudil et al. 1994) is therefore a good candidate. Evolutionary computing motivated Genetic Algorithms (GA)

(McGilloway et al. 2000) can be used, but they are not ideal for decision type objective functions, nor when using a low amount of features, i.e. around 50–200, typical in emotional speech analysis. In transformations, the existing features are transformed into a new set of features.

The transformed set of features typically satisfies some desired property. For example, the new transformed feature set is an orthogonal set of features, which is the case in Principal Component Analysis (PCA). The target dimensionality of the transformations is almost always a lower dimension than the starting feature space to counteract the curse of dimensionality. Although transformations are usually defined as unsupervised methods, they can also implement supervised learning, e.g. Linear Discriminant Analysis (LDA). Nonlinear transformations are often used in supervised learning. Using the kernel trick (Aizerman et al. 1964), an extension of PCA to nonlinear spaces, i.e. Kernel PCA (KPCA), is straightforward.

KPCA can be seen as a general framework for nonlinear transformations and thus many nonlinear techniques can be reduced to special cases of KPCA (Maaten et al. 2009). Neural Networks (NN) (Haykin 1994) can also be used to learn nonlinear transformations, e.g. using a General Regression Neural Network (GRNN) method. Manifold modelling techniques can also be seen as a type of 47 feature transformations. Both selection and transformation methods can be used at the same time. It is important to note that neither approach gives guaranteed increases in performance as implied by the NFL theories. Transformations and selection are, however, great tools for implementing desired models in feature extraction to simplify subsequent classifier and decision methods

6. SIMULATION RESULTS

6.1 Angry Signal Features

Peak Point: 0.5543

Mean Signal:--0.9941

Minimum level Signal: 1.9137e-005

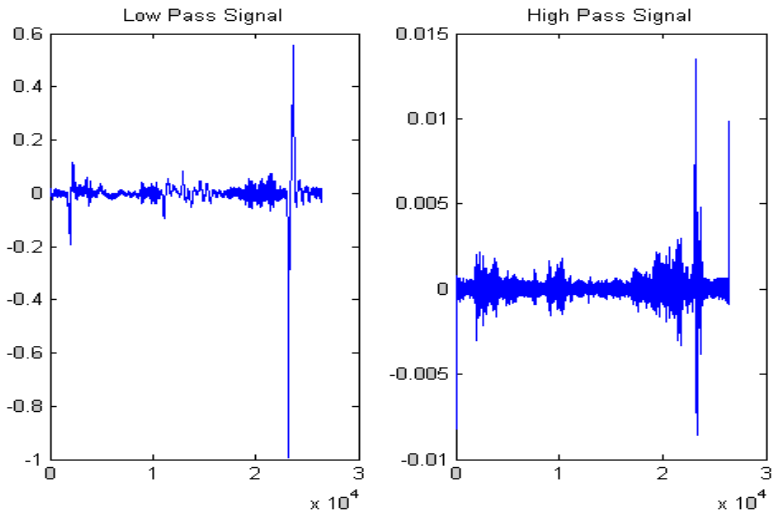


Figure 6.1 Angry Signal Features

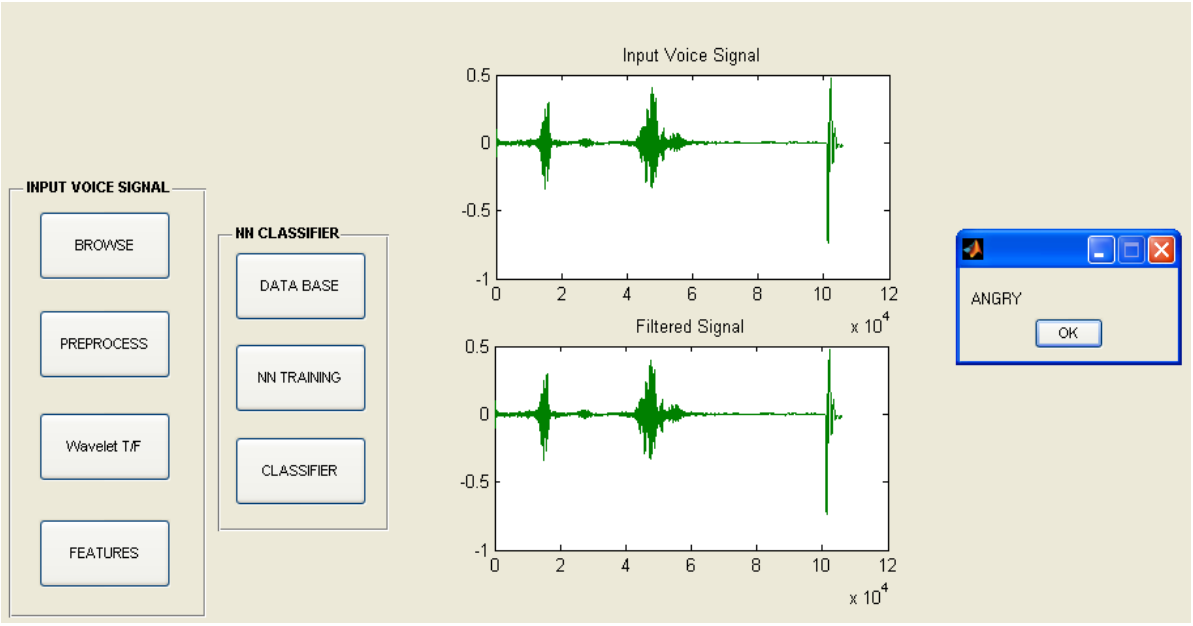


Figure 6.2 Angry voice signal

6.2 Happy Signal Features

Peak Point: 0.2268

Mean Signal:- -0.9305

Minimum level Signal: -9.1273e-004

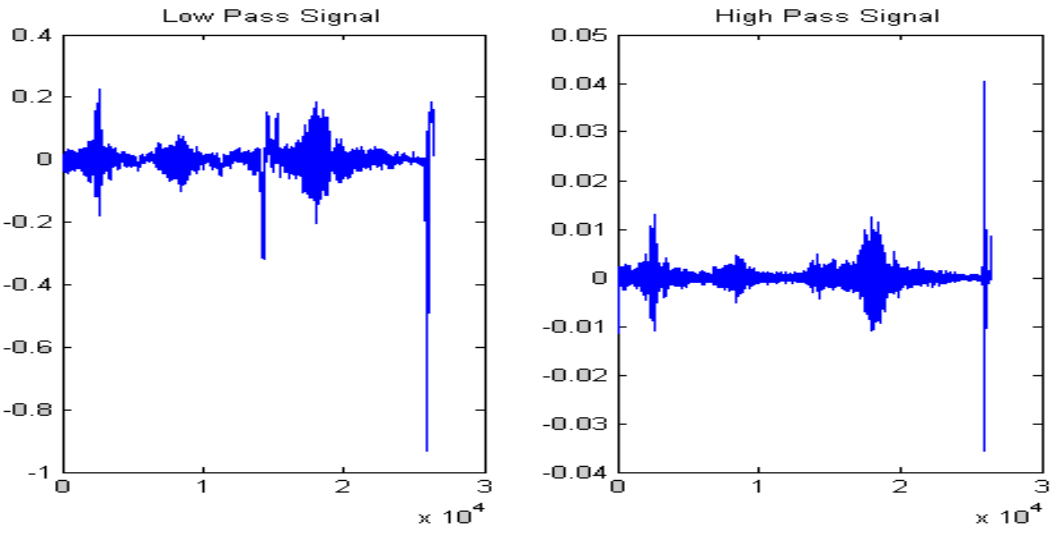


Figure 6.3 Happy Signal Features

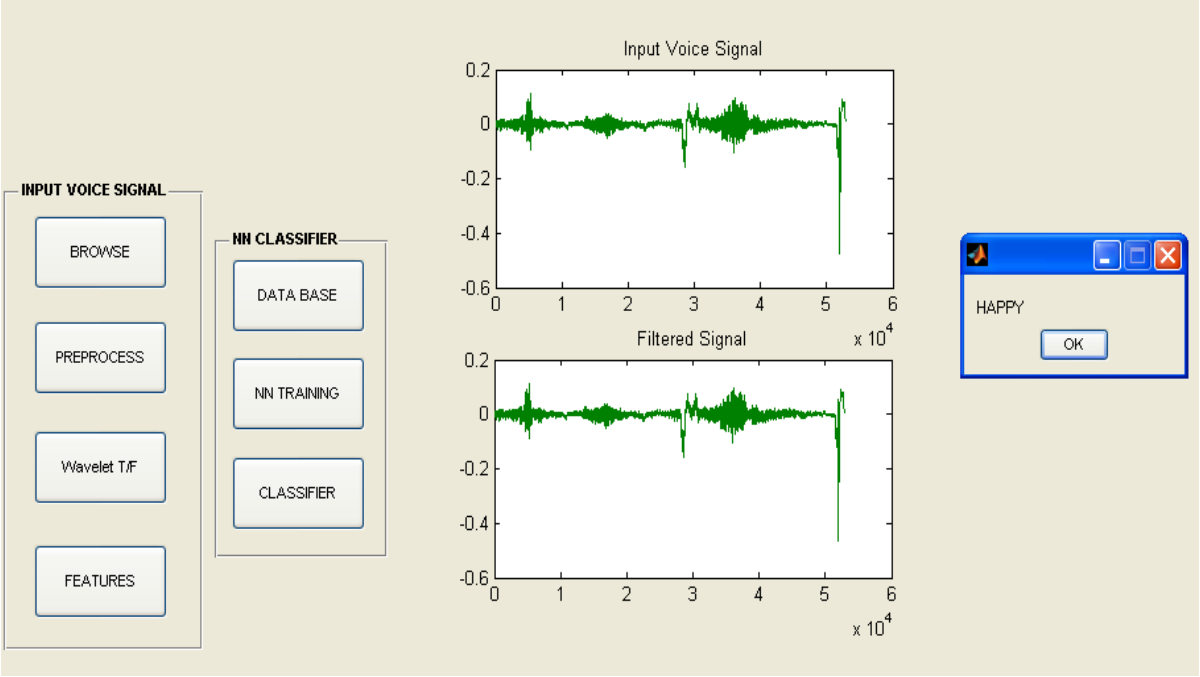


Figure 6.4 Happy voice signal

6.3 Sad Signal Features

Peak Point: 1.0781

Mean Signal: -1.1281

Minimum level Signal: 1.1351e-005

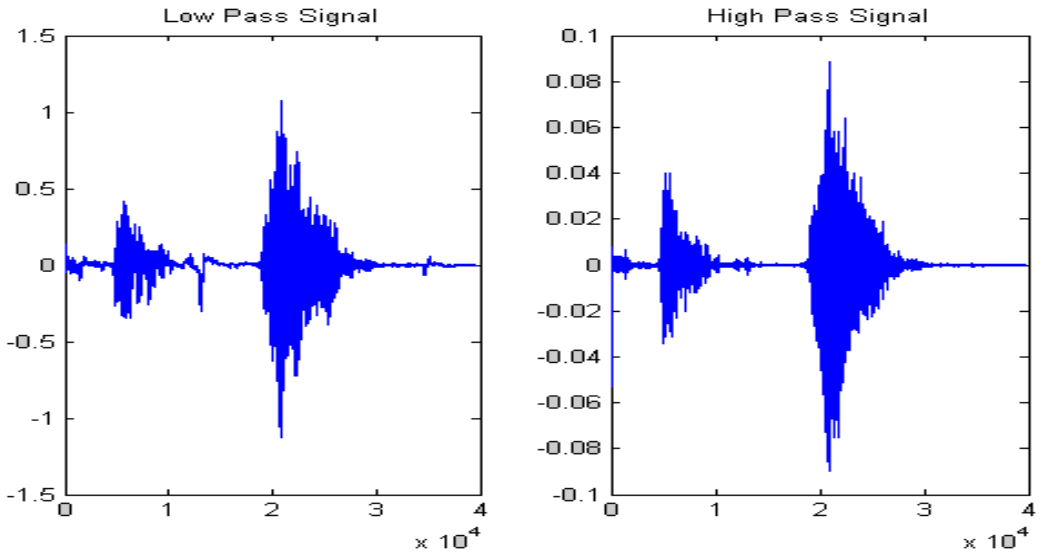


Figure 6.5 Sad Signal Features

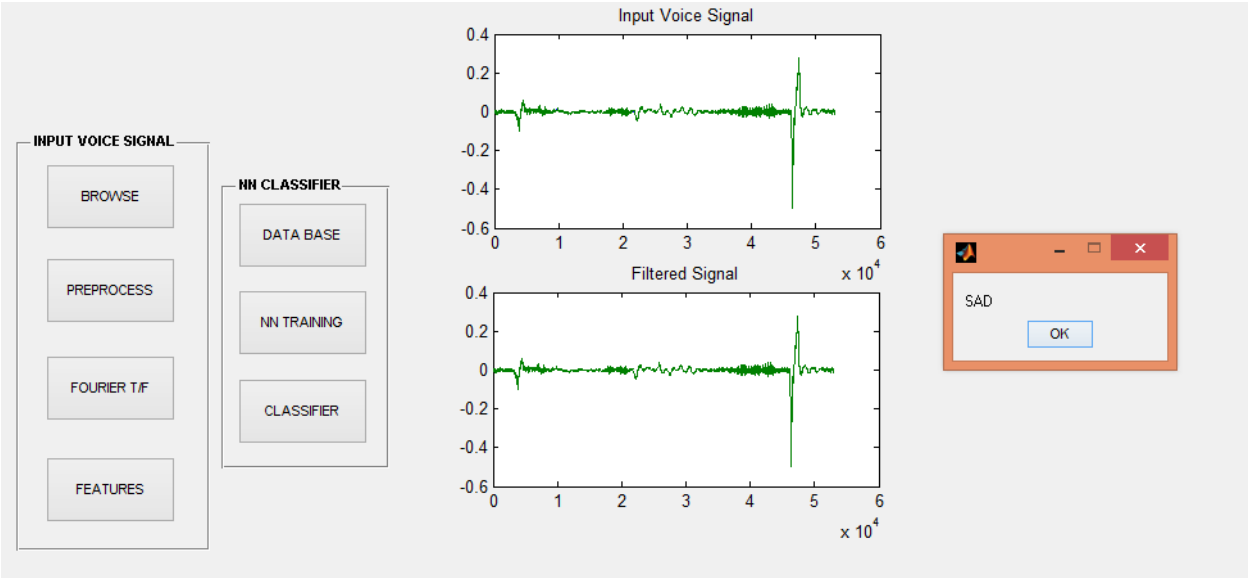


Figure 6.6 Sad voice signal

6.4 Neutral Signal Features

Peak Point: 0.7076

Mean Signal: -1.5265

Minimum level Signal: -4.0754e-005

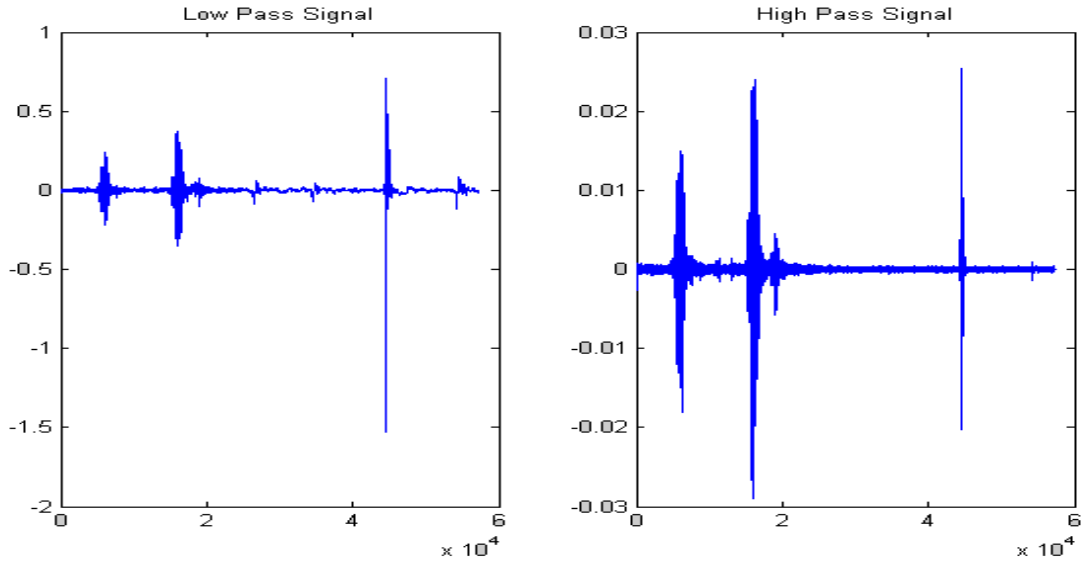


Figure 6.7 Neutral Signal Features

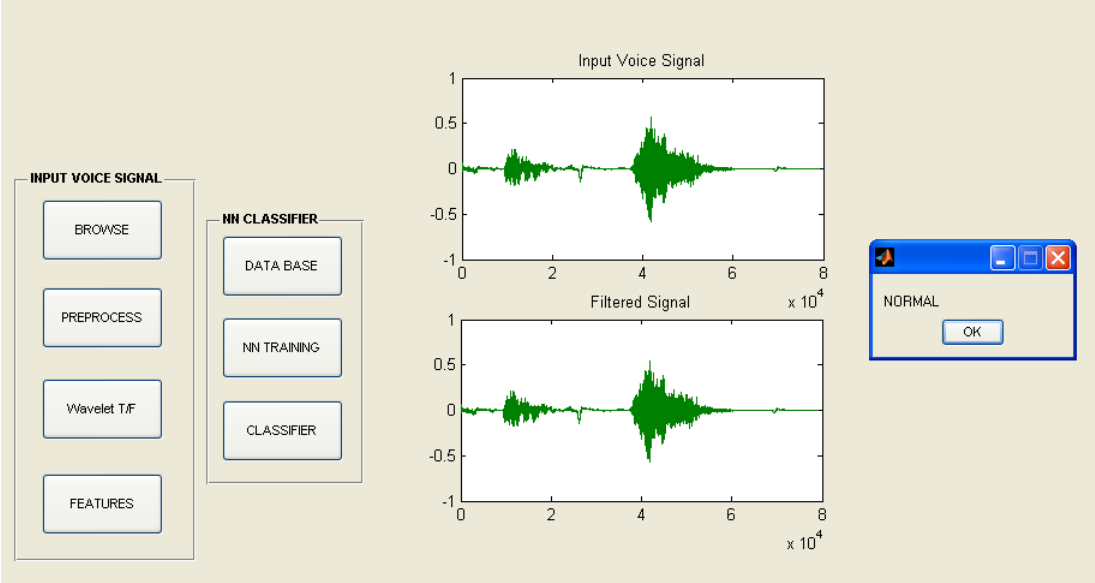


Figure 6.8 Neutral voice signal

7. CONCLUSION

Thus the Tamil emotional corpus was created using emotion dialogues from different Kollywood movies. Initially all files were cut in .mp3 format and then converted into (.wav) and the emotions were categorized as happy, sad, anger, fear and neutral. The MFCC values were extracted for every emotion and 60% of the values were used for training the SVM'S and the remaining 40% were used for testing the SVM classifier. The output accuracy rate was good. The system was also able to detect emotions in real time. So that the machine can understand which emotion the user was in. We can develop it into the Psychological Analysis, Robotics and also for the business development. The efficiency of the engine can be further improved by increasing the training data.The system can be improved by increasing the number of speakers and increasing the training data.

REFERENCES

[1] Kunxiawang , Ning An, Bing Nan Li , Yanyong Zhang, ”Speech EmotionRecognition Using Fourier Parameters”,Intelligent Computing and Intelligent Systems (ICIS), Vol. 6, pp. 69-75,Jan 2015.

[2] Md. Touseef Sumer, “Salient Feature Extraction For Emotion Detection Using Modified KullbackLeibler Divergence”, Internationl Journal of Research in Engineering and Applied Science (IJREAS),Vol. 2, pp. 60-69, Jan 2014.

[3] VidhyasaharanSethu, EliathambyAmbikairajah and Julien Epps, “On The Use Of Speech Parameter Contours For Emotion Recognition”, EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2,pp. 36-46, Oct 2013.

[4] BiswajitNayak, MitaliMadhusmita and Debendra Ku Sahu, “Speech Emotion Recognition using Different Centred GMM”, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 3, pp. 55-64. Sep 2013.

[5] Akshay S. Utane and Dr. S.L .Nalbalwar, “ Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model” International Journal of Advanced Research in Computer Science and Software Engineering,Vol.4, pp.34-45, April 2013.

[6] Stavros Ntalampiras and Nikos Fakotakis, “Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition”, IEEE Transactions On Affective Computing, Vol.3, pp. 167-175, January-March 2012.

[7] Chung-Hsien Wu and Wei-Bin Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels”, IEEE Transactions On Affective Computing, Vol. 2, pp.10-19, January-March 2011.

[8] Yuan Yujin, Zhao Peihua and Zhou Qun, “Research of speaker recognition based on combination of LPCC and MFCC”, Intelligent Computing and Intelligent Systems (ICIS), Vol. 12, pp. 269-275, Oct 2010.

[9] Carlos Busso, Sungbok Lee and Shrikanth Narayanan,”Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection” ,IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, pp. 582-596, May 2009.

[10] Daniel Neiberg, KjellElenius and KornelLaskowski,” Emotion Recognition in Spontaneous Speech Using gmms”, Vol. 2,pp.56-65,2008 .

[11] Mohammed E. Hoque, Mohammed Yeasin and Max M. Louwerse, “Robust Recognition of Emotion from Speech”,Intelligent Virtual Agents, Vol.12, pp. 42-53, 2006.

[12] Chul Min Lee and Shrikanth S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs”, IEEE Transactions On Speech And Audio Processing, Vol. 13, pp. 293-303, March 2005.

APPENDIX 1

CODING

```
functionvarargout = gui_final(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @gui_final_OpeningFcn, ...
                   'gui_OutputFcn',  @gui_final_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
ifnargin&&ischar(varargin{1})
gui_State.gui_Callback = str2func(varargin{1});
end

ifnargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
gui_mainfcn(gui_State, varargin{:});
end
functiongui_final_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;
a=ones([200 450]);
axes(handles.axes1);imshow(a);
axes(handles.axes2);imshow(a);
guidata(hObject, handles);
functionvarargout = gui_final_OutputFcn(hObject, eventdata, handles)
varargout{1} = handles.output;
functioninp_voice_Callback(hObject, eventdata, handles)
cdDatasamples
file=uigetfile('*.wav');
```

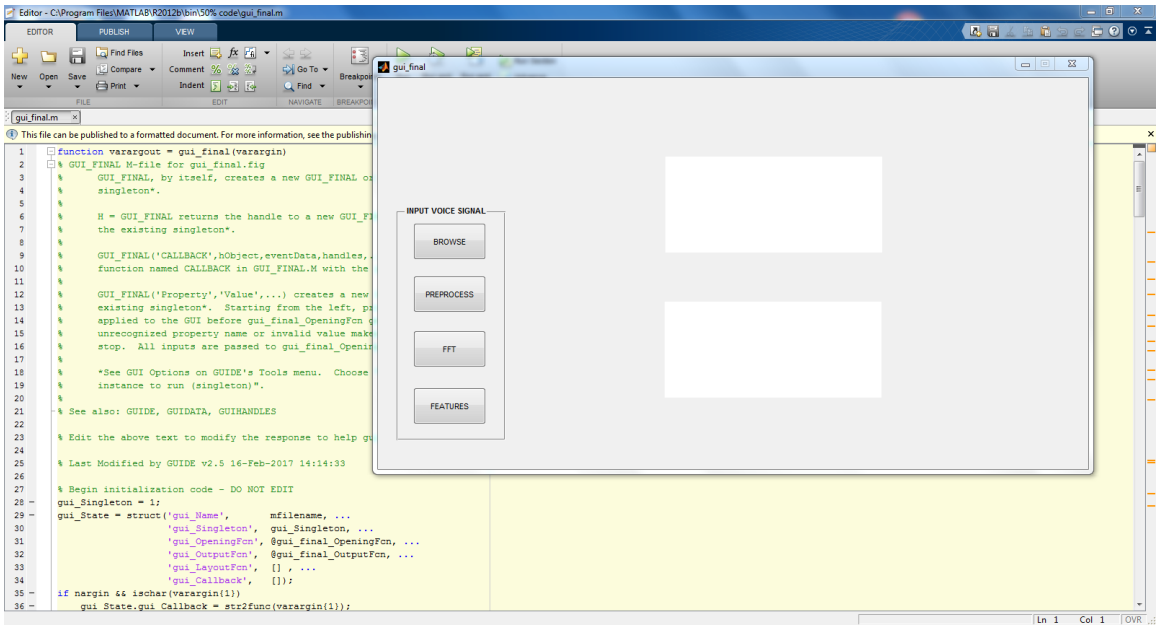
```
inp=wavread(file);
[ speech, fs, nbits ] = wavread(file);
cd ..
wavplay(inp,44200);
axes(handles.axes1);
plot(speech);title('Input Voice Signal');
handles.speech=speech;
handles.fs=fs;
handles.nbits=nbits;
handles.file=file;
guidata(hObject, handles);
functionpre_process_Callback(hObject, eventdata, handles)
speech=handles.speech;
filt_sig=medfilt2(speech,[3 3]);
wavplay(filt_sig,44200);
axes(handles.axes2);
plot(filt_sig);title('Filtered Signal');
handles.filt_sig=filt_sig;
guidata(hObject, handles);
functionfft_Callback(hObject, eventdata, handles)
filt_sig=handles.filt_sig;
[rows cols] = size(filt_sig);
fft_sig =fft(filt_sig,[rows cols]);
figure;
plot(fft_sig);title('FFT Signal');
handles.fft_sig=fft_sig;
guidata(hObject, handles);
functionfeatures_Callback(hObject, eventdata, handles)
inp=handles.speech;
fs=handles.fs;
nbits=handles.nbits;
```

```
fft_sig=handles.fft_sig;
     f1=max(max(fft_sig));
     f1 = abs(f1);
    f2=min(min(fft_sig));
    f2 = abs(f2);
     f3=mean(mean(fft_sig));
     f3 = abs(f3);
    f5=mean(mean(abs(medfilt1(fft_sig))));
    f5 = abs(f5);
    f6=std2(fft_sig);
    f6 = abs(f6);
     p= hist(inp);
     f7= -sum(sum(p.*log2(p)));
     f7 = abs(f7);
    f8=entropy(inp,256);
    f8 = abs(f8);
   [f9, t] = FeatureTimeZeroCrossingRate(inp, 42100, 256,256);
   f9=mean(f9);
   f9 = abs(f9);
%    f10=sum(sum(fft_sig));
%   f10 = abs(f10);
```

APPENDIX- 2

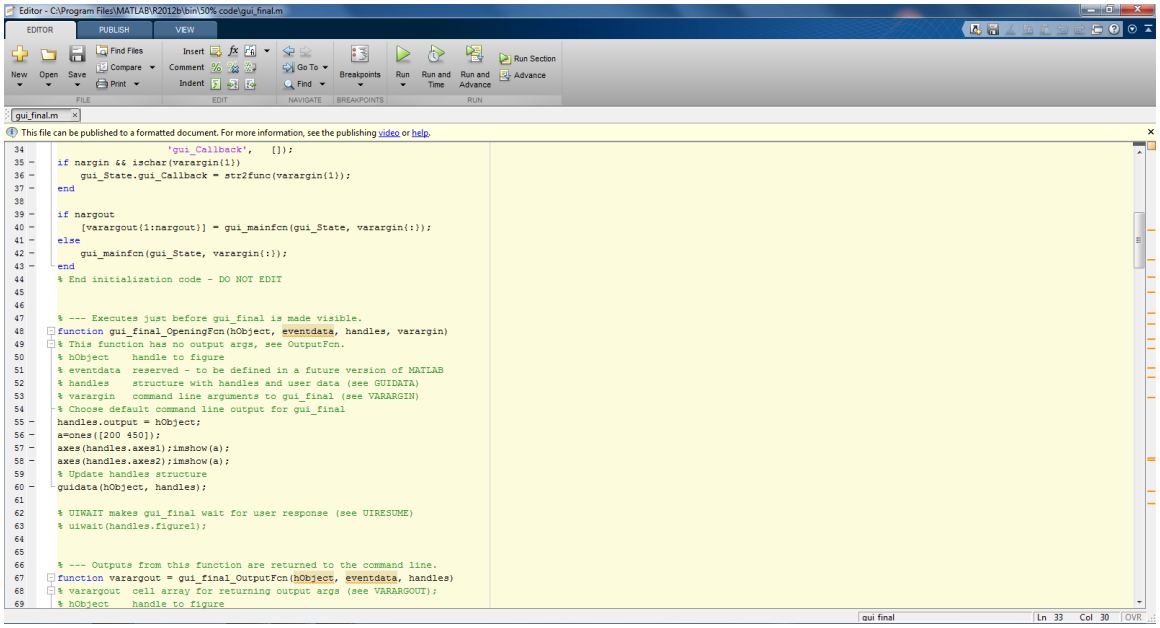
SAMPLE SCREENSHOT

System to detect emotion



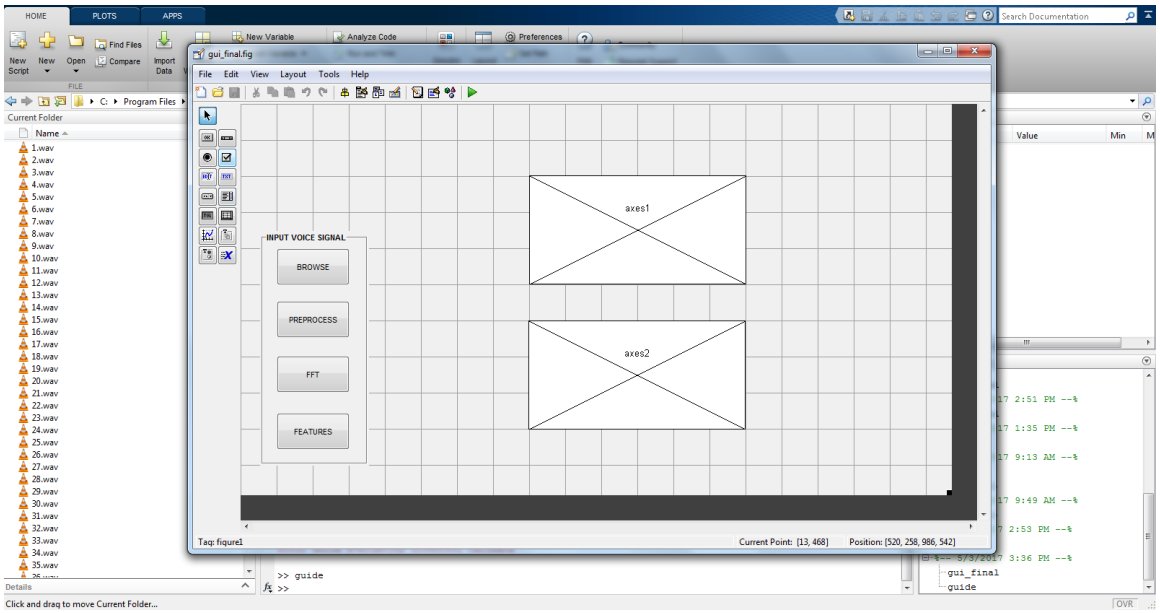
A2.1

Coding page in MATLAB



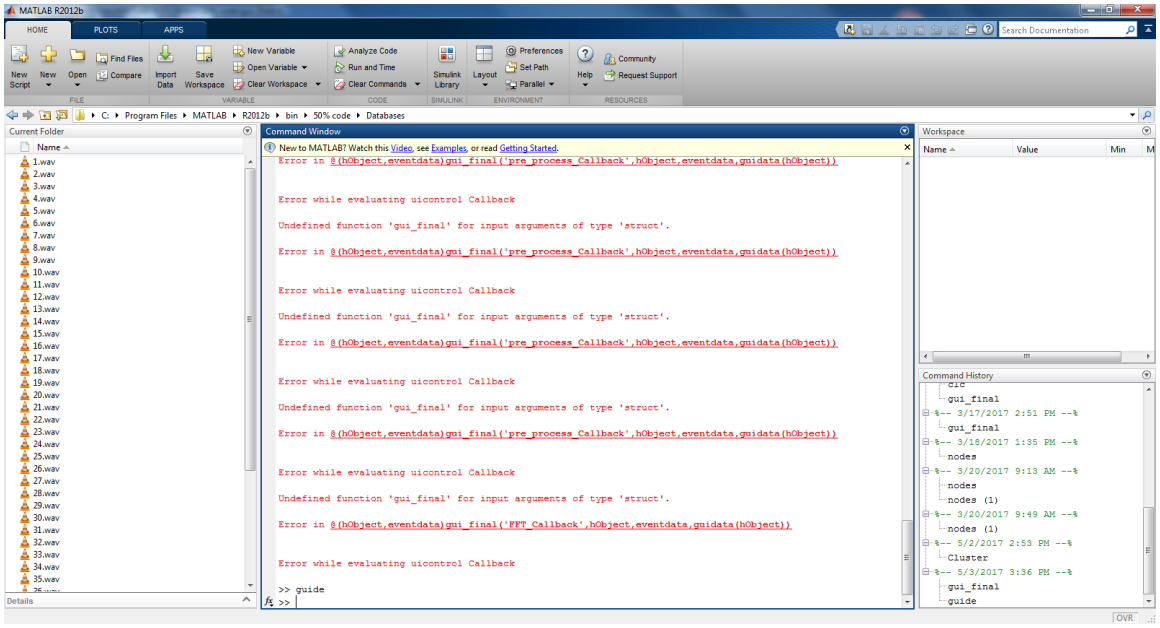
A2.2

Graphical User Interface for designing in MATLAB



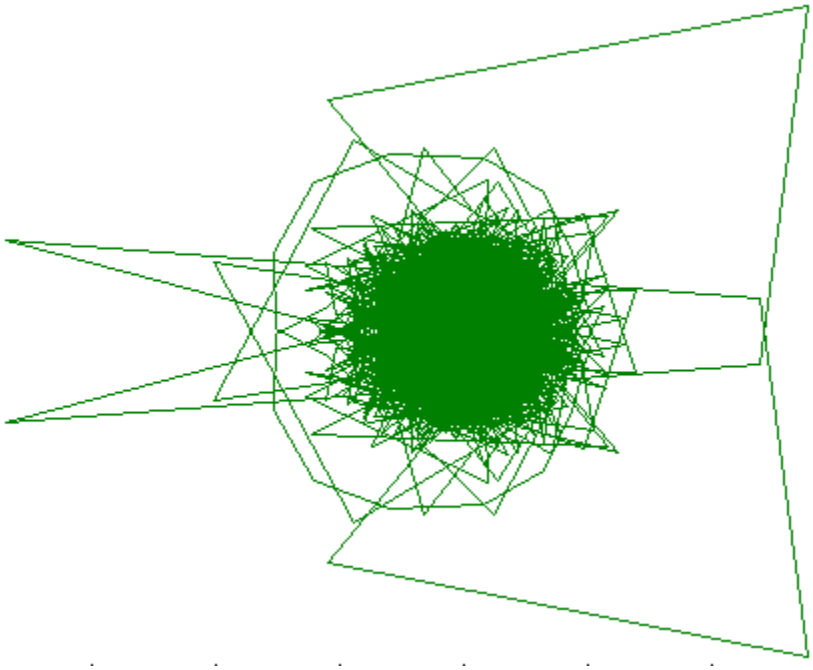
A2.3

Command Window of MATLAB



A2.4

Fast Fourier Transform of speech signal



A2.5

TECHNICAL BIOGRAPHY

Mr. ARUN GOPAL.G (130071601012) born on 09th January 1996, in Chennai, Tamil Nadu. Completed my schooling in SitaDevi Garodia Hindu Vidyalaya Matriculation Higher Secondary School, secured 92.33% in the Higher Secondary Examination and pursuing B.Tech Computer Science and Engineering at B.S. Abdur Rahman Crescent University. My area of interests includes playing and watching Cricket. My e-mail ID: aarunkarthii@gmail.com and contact number: 8056012274.

TECHNICAL BIOGRAPHY

Mr. CHRISTY XAVIER RAJ.K (130071601021) born on 31th October 1995, in Pudukkottai, Tamil Nadu. Completed my schooling in Vairams Matriculation Higher Secondary School, secured 87% in the Higher Secondary Examination and pursuing B.Tech Computer Science and Engineering at B.S. Abdur Rahman Crescent University. My area of interests includes playing and watching Football. My e-mail ID: sterlingkriz@gmail.com and contact number: 9500580692.