Description: Description: Z:\forms\greenwich.jpg

October University for Modern Science & Arts

Faculty of Engineering

Department of Electrical Communication and Electronics Systems

Real time Speech Emotion Recognition for Archetypal Emotions

A Graduation Project

Submitted in Partial Fulfillment of B.Sc. Degree

Requirements in Electrical and Computer Engineering

Part II

Prepared By

Fatima Fouda 121649

Yosif Atef 120723

Supervised By

Dr. Noha Younis

2015–2016

Table of Contents

[List of Tables i](#_Toc453513272)

[List of Figures i](#_Toc453513273)

[List of Equations ii](#_Toc453513274)

[List of Abbreviations iii](#_Toc453513275)

[Acknowledgment iv](#_Toc453513276)

[ABSTRACT v](#_Toc453513277)

[Chapter I 1](#_Toc453513278)

[INTRODUCTION 1](#_Toc453513279)

[1.1 Problem definition 1](#_Toc453513280)

[1.2 Applied systems 1](#_Toc453513281)

[1.3 Main objectives 2](#_Toc453513282)

[1.4 Technology involved 2](#_Toc453513283)

[1.5 Report organization 3](#_Toc453513284)

[Chapter II 4](#_Toc453513285)

[Review of Literatures 4](#_Toc453513286)

[2.1 History 4](#_Toc453513287)

[2.1.1 Emotion Recognition in Speech Using Neural Networks [3] 4](#_Toc453513288)

[2.2 Current alternative systems 6](#_Toc453513289)

[2.2.1 Speech Emotion recognition using Fourier Parameters [5] 6](#_Toc453513290)

[2.2.2 Speech Emotion Recognition with prosody, quality and derived features using SVM classifier [6] 7](#_Toc453513291)

[Chapter III 11](#_Toc453513292)

[Proposed System: 11](#_Toc453513293)

[Real time Speech Emotion Recognition for Archetypal Emotions 11](#_Toc453513294)

[3.1 Main features 11](#_Toc453513295)

[3.2 Used Corpora 11](#_Toc453513296)

[3.3 Methodology 12](#_Toc453513297)

[3.4 System description 12](#_Toc453513298)

[3.4.1 Frame blocking and Windowing 13](#_Toc453513299)

[3.4.2 FFT 13](#_Toc453513300)

[3.4.3 MFCC 14](#_Toc453513301)

[3.4.4 LPC 16](#_Toc453513302)

[3.4.5 LFPC (I need the LFPC paper to replace these references) 17](#_Toc453513303)

[3.4.6 KNN 18](#_Toc453513304)

[3.5 Flow chart 20](#_Toc453513305)

[3.6 Testing method 22](#_Toc453513306)

[3.7 Summary 22](#_Toc453513307)

[Simulation and Testing 23](#_Toc453513308)

[4.1 Simulation Results 23](#_Toc453513309)

[4.1.1 Simulation result of MFCC – KNN based design 23](#_Toc453513310)

[4.1.2 Simulation result of LFPC – KNN based design 24](#_Toc453513311)

[4.1.3 Simulation result of LPC & KNN based design 26](#_Toc453513312)

[4.1.4 Simulation result of the integrated system 28](#_Toc453513313)

[4.2 Simulation discussion 29](#_Toc453513314)

[4.3 Problems encountered 30](#_Toc453513315)

[4.4 Summary 30](#_Toc453513316)

[Chapter V 31](#_Toc453513317)

[Cost Analysis 31](#_Toc453513318)

[Chapter VI 32](#_Toc453513319)

[Time Plan 32](#_Toc453513320)

[Chapter VII 34](#_Toc453513321)

[Conclusion 34](#_Toc453513322)

[7.1 Conclusion 34](#_Toc453513323)

[References 35](#_Toc453513324)

[Appendix A 37](#_Toc453513325)

# List of Tables

[Table 2.1: Comparison of the different neural network structures (30 subjects) 6](#_Toc451353393)

[Table 2.2 Comparison between CASIA, EMODB, and ESDB 7](#_Toc451353394)

[Table 3.1: Filter order to efficiency[5] 14](#_Toc441109180)

[Table 3.2: First five formant frequencies for anger and happiness states 16](#_Toc441109181)

[Table 4.1: Accuracy rates for each emotion compared by database 30](#_Toc454375470)

[Table 4.2: Accuracy rates for Arabic database 30](#_Toc454375471)

[Table 4.3: Human Testing For Arabic Database 30](#_Toc454375472)

[Table 4.4: Conventional and Modified KNN accuracy rates 31](#_Toc454375473)

[Table 4.5: Comparison between different systems 32](#_Toc454375474)

[Table 5.1: Component comparison by cost 31](#_Toc441109807)

[Table 6.1 Task Distribution and Time Plan 32](file:///D:\grad\Final_term%20Grade%202_mod.docx#_Toc451353355)

[Table 6.2 Gantt Chart 33](file:///D:\grad\Final_term%20Grade%202_mod.docx#_Toc451353356)

# List of Figures

[Figure 2.1: Processing Flow 4](#_Toc451353325)

[Figure 2.2: OCON Emotion Recognition Phase 5](#_Toc451353326)

[Figure 2.3: Steps for feature extraction [6] 8](#_Toc451353327)

[Figure 2.4: Types of SVM [6] 9](#_Toc451353328)

[Figure 2.5: Emotion recognition accuracy in Karbi language [6] 9](#_Toc451353329)

[Figure 3.1: Proposed system hierarchical block diagram 11](#_Toc454375481)

[Figure 3.2: Proposed System block diagram 13](#_Toc454375482)

[Figure 3.3: MFCC block diagram 14](#_Toc454375483)

[Figure 3.4: Mel-filter bank response. 15](#_Toc454375484)

[Figure 3.5: LFPC block diagram [18] 17](#_Toc454375485)

[Figure 3.6: LFPC Analysis example [18] 18](#_Toc454375486)

[Figure 3.7: Decision making based on distance [20] 18](#_Toc454375487)

[Figure 3.8: Decision construction 20](#_Toc454375488)

[Figure 3.9: Main algorithm used to design the proposed ASER system 21](#_Toc454375489)

[Figure 3.10: Proposed system flow chart 21](#_Toc454375490)

[Figure 4.1: Extracted MFCC Coefficients of high and low arousal for SAVEE training database. 23](file:///D:\grad\Last\attachment%20(2).docx#_Toc454375491)

[Figure 4.2: Extracted MFCC Coefficients of high and low arousal for TESS training database. 24](#_Toc454375492)

[Figure 4.3: Extracted LFPC Coefficients of Neutral and Sadness emotions for SAVEE training database. 25](#_Toc454375493)

[Figure 4.4: Extracted LFPC Coefficients of Neutral and Sadness emotions for TESS training database. 26](#_Toc454375494)

[Figure 4.5: LPC coefficients of anger and happiness emotions for the same phrase, uttered in Anger and Happiness. 27](#_Toc454375495)

[Figure 4.6: LPC extracted coefficients from the TESS training database for happiness and fear emotions. 27](#_Toc454375496)

[Figure 4.7: LPC coefficients of the same phrase uttered in anger state vs. Happy State from TESS 28](#_Toc454375497)

[Figure 4.8: LPC for happiness and anger emotions extracted 28](#_Toc454375498)

[Figure 4.9: Binary decision tree Flow chart 29](#_Toc454375499)

[Figure 4.10: User friendly GUI 33](#_Toc454375500)

# List of Equations

[Equation 3.1: Hamming window 13](#_Toc441109852)

[Equation 3.2: Frame data times hamming window 13](#_Toc441109853)

[Equation 3.3: FFT 14](#_Toc441109854)

[Equation 3.4: Frequency to Mel Scale 15](#_Toc441109855)

[Equation 3.5: Inverse Mel Scale 15](#_Toc441109856)

# List of Abbreviations

**ACON: A Class-in-One Neural Network**

**ASER: Automatic Speech Emotion Recognition**

**ATR: Advanced Telecommunications Research Institute International**

**DCT: Discrete Cosine Transform**

**F0: Fundamental Frequency**

**FEZ: Fundamental Frequency + Energy+ Zero Crossing Rate**

**FFT: Fast Fourier Transform**

**FIR: Finite Impulse response**

**FP: Fourier Parameters**

**FPGA: Field Programmable Gate Array**

**GMM: Gaussian Mixture Model**

**HCI: Human Computer Interaction**

**KNN: K-Nearest Neighbor**

**LFPC: Log Frequency Power Coefficient**

**LPC: Linear Prediction Coefficient**

**LVQ: Learning Vector Quantization**

**MESDNEI: Multilingual Emotional Speech Database of North East India**

**MFCC: Mel-Frequency Cepstral Coefficient**

**OCON: One-Class-One Neural Network**

**SER: Speech Emotion Recognition**

**SUB: Sub-band Based Cepstral Parameter**

**SVM: Support Vector Machine**

**ZCR: Zero Crossing Rate**

# Acknowledgment

First and foremost, we thank Dr. Noha Younis for her mentoring role, and for her relentless efforts in pushing us to create a better project.

We also thank the faculty of Mass Communication at 6th of October University for Modern Sciences and Arts for their sincere cooperation. We appreciate the many hours of the recording studios at our disposable, the student actors who spent their valuable time recording the database, and the audio engineer who helped compile and segment the utterances.

# ABSTRACT

Emotion, which is the most normal way to express feeling, is a common sense for human beings. Understanding adults' emotion using machines would result in a new era of communication with machines. However, despite the marvelous progress in current Automatic Speech Emotion Recognition (ASER), normal interaction with machines is still far from being applied efficiently, because of the low accuracy of most speech emotion recognition. Also most speech emotion algorithms are complicated and can't be used in portable applications.

The main aim of the proposed system is designing an automatic speech emotion recognition system for the archetypal emotions. The main objectives are combining several speech features, of low computational complexity, to achieve real time performance and improve the ASER accuracy. Also the system targets two different languages for the applied database; that are Arabic and English. The proposed system should be also user independent.

The main algorithms of the proposed system are designed based on signal processing techniques for speech recognition. The K-Nearest Neighbor classifier (K-NN) was used and its algorithm was modified to improve the accuracy rate. The system was designed and tested using MATLAB then it was compared with the most applied current systems. Yet, the range of accuracy achieved is between (51% - 95.8%) for the English databases used for Happiness, Anger, Sadness, Fear and Neutral, based on the selected speech features and the modified K-NN classifier**.** The Arabic database’s accuracy range is between (51% - 81.8%) for Happiness, Anger and Sadness.

# Chapter I

# INTRODUCTION

## 1.1 Problem definition

Automatic Speech Emotion Recognition (ASER) is extracting speaker’s voice and analyzing it for extracting one's emotional feeling state. The ASER technology draws attention, due to believing that it can be used in various applications. The most essential of these applications are Human Computer Interaction (HCI), and human-human interaction. However, other uses for ASER systems vary from criminals' confession, investigations, and psychological cases to health care systems. To extract the data required to recognize the speech emotions, speech should be analyzed and different speech features need to be extracted in order to guarantee sound classification.

It is found that continuous features such as: pitch features, energy features, formants features deliver an important data in analyzing emotional speech. There are also spectral features (features that depend on frequency domain processing), such as: Mel-Frequency Cepstral Coefficient (MFCC), Linear Prediction Coefficient (LPC), and Log Frequency Power Coefficient (LFPC) that play an important role in ASER.

Some disadvantages of previous systems include limited database, and a high computational complexity which render the system immobile.

## 1.2 Applied systems

Although, the great effort exerted, still there are a lot to do to improve the current system accuracy. For instance, in [[1](#Vau12)] an ASER system is presented that is based on Fusion technique. The maximum accuracy rate, achieved is 77 %. Figure 1.2 depicts the accuracies resulted from the application of three different classifiers.

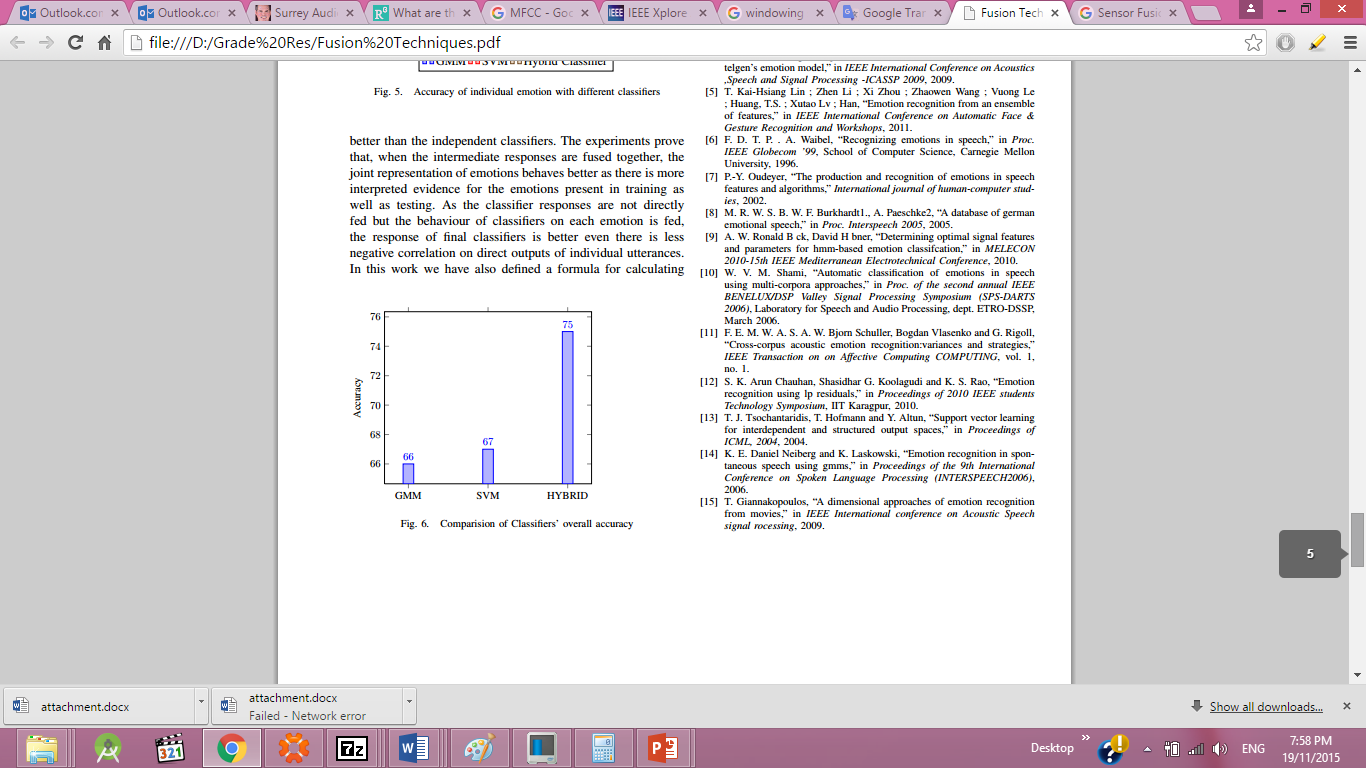


Figure 1.1: The resultant accuracy of the ASER system based on the classifier type [[2](#Kri13)]

Another ASER system is presented in [[2](#Kri13)], where the Sub-band Based Cepstral Parameter (SUB) feature is used together with the Gaussian Mixture Model (GMM) as a classifier to recognize five different emotions. The maximum accuracy rate achieved is 70%.

## 1.3 Main objectives

The main objectives of the proposed system are to design a real time system desktop or web application by reducing the computational complexity, to improve the ASER system’s accuracy by mixing different speech features that can be represented by simple algorithms. Although it’s not our main objective but the proposed system’s reduced power consumption makes it applicable to hardware implementation for psychological analysis and diagnosis.

Another important objective is to design a global system that is compatible with two different languages, English and Arabic. It’s also hoped to achieve a general purpose system that doesn't depend on particular database type.

## 1.4 Technology involved

Signal processing is used to extract the required speech features and to design the emotion classifier. The system is designed and tested in Part1, using MATLAB. Selecting the platform on which the system will be developed and implemented, in Part 2, depends mainly on the application. As an example, if the system is dedicated to psychological or health care systems, then it is better to implement it on a Hardware platform, such as FPGAs, Raspberry Pi, or an Android based device. However, if the system is dedicated for an application such as call center services to measure the satisfaction level of customers, on one side and to recognize the way the employee interact with customers on the other side, then in this case the best platform to develop the system will be a desktop, or web application. After thoughtful consideration, it has been decided to apply the system to call centers, and customer service sectors in general; hence, the platform chosen is a desktop application to suit this particular application.

## 1.5 Report organization

First, the proposed system is introduced in Chapter1. A review of literature including previous and current applied systems is presented in Chapter 2.

In Chapter 3, the paper explores in depth the proposed system, along with all of its components and techniques.

In Chapter 4, the simulation and testing results are provided and discussed, while the perceived cost analysis of the design is estimated and proposed in Chapter 5. In Chapter 6, the time plan of the work is summarized, whereas the conclusion and the future plan are discussed in Chapter 7.

# Chapter II

# Review of Literatures

## 2.1 History

Only dating back a few decades, speech emotion recognition is a relatively new field. The following delineates one of the first systems in this field

### 2.1.1 Emotion Recognition in Speech Using Neural Networks [[3](#Nic99)]

This system is designed by the Advanced Telecommunications Research Institute International (ATR) in Japan. As shown in Figure 2.1. The system relies on several feature extraction techniques, and uses a neural network as a classifier.

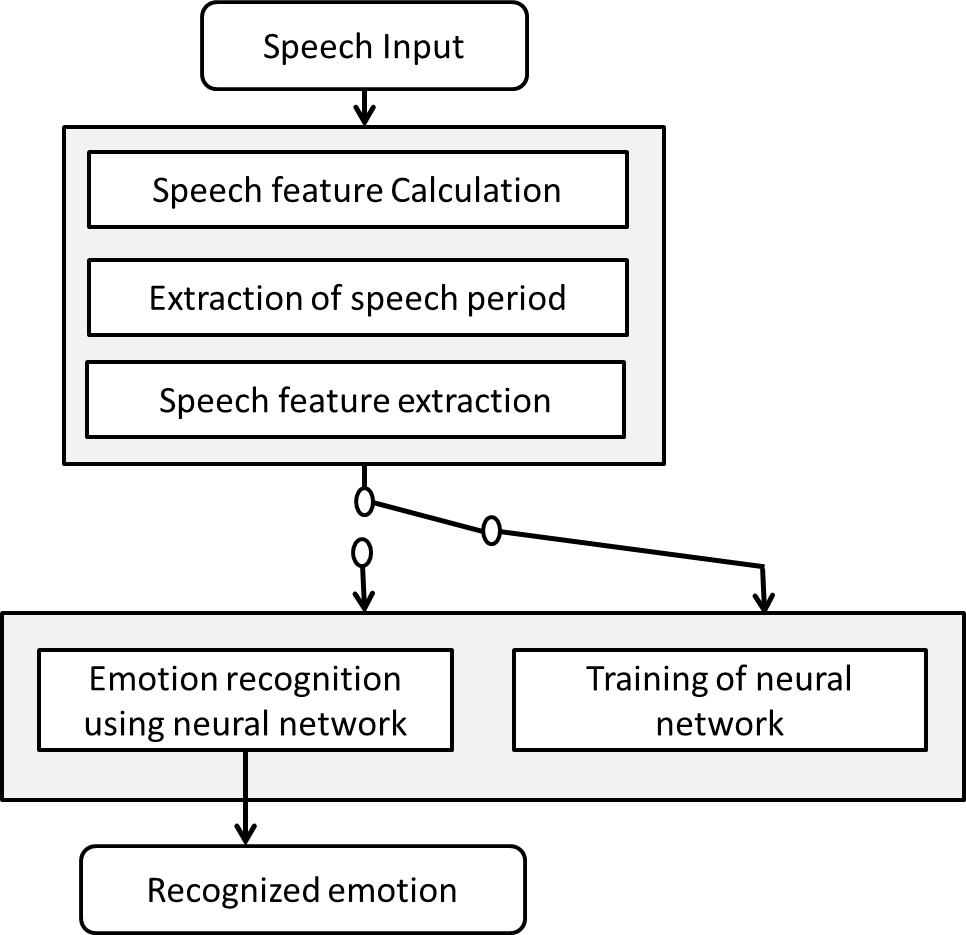


Figure 2.1: Processing Flow

An example of a neural network-based ASER system is presented in [[3](#Nic99)]. The system aims to identify eight different emotions: joy, teasing, fear, sadness, disgust, anger, surprise, and neutral. The system relies on a specially collected speech database based on a large number of phoneme balanced Japanese words. These utterances are collected from 100 radio actors, 50 of whom are male, and 50 female. Since, this accounts for a substantially large database, the system is both speaker independent and context independent.

The speech features used are a) speech power b) pitch c) 12 LPC parameters and d) Delta LPC parameter. These features are then flattened into a one-dimensional vector, which is then fed into the neural network.

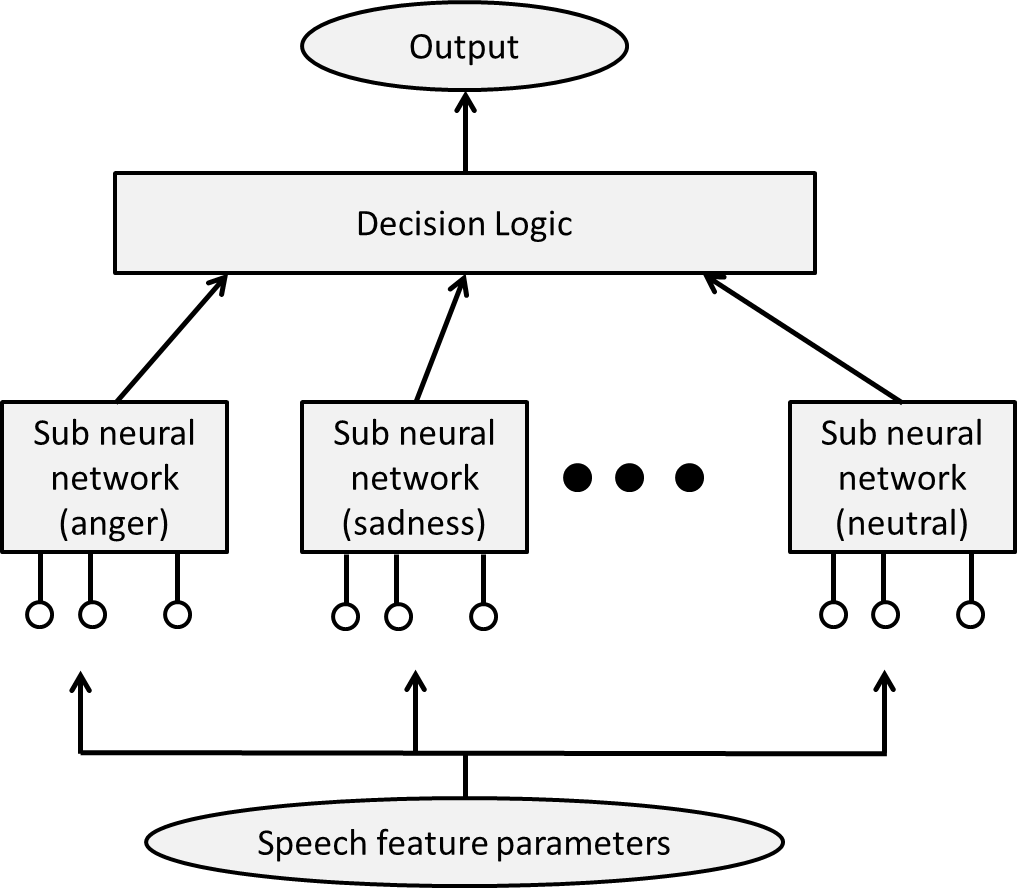


Figure 2.2: OCON Emotion Recognition Phase

The classifier stage of the system is diagramed in Figure 2.2. This type of network is called a one-class-one neural network (OCON), because each one of the eight emotions has its own sub-neural network. The advantage of such network is that it allows for the modification of each individual sub-network without disturbing the entire network. The output — an analog value between 0 and 1— of each sub-neural network is the likelihood of the input features to this particular emotion. The decision logic module then decides the emotion with the “best” likelihood. The authors also note that negative emotions, like anger and sadness were easier to recognize than positive emotions, like joy.

The system results in a 50% accuracy rate. A comparison against the Class-in-One Neural Network (ACON) and the single layer neural network, utilizing a learning vector quantization method (LVQ), was carried out. The comparison is demonstrated in Table 2.1.

Table 2.1: Comparison of the different neural network structures (30 subjects)

|  |  |  |
| --- | --- | --- |
| **Network** | **Closed results** | **Open results** |
| **OCON** | 56.93 | 52.87 |
| **ACON** | 71.99 | 57.18 |
| **LVQ** | 72.23 | 33.32 |

The paper concludes by stating that experimenting with different neural network topologies may result in a higher accuracy. On the other hand, neural networks are a type of machine learning algorithms that possess a very high computational complexity [[4](#Orp94)], and are often difficult if not impossible to translate into a portable device.

## 2.2 Current alternative systems

Advance in understanding the human voice and its features, as well as, advance in current machine learning algorithms, has led current ASER systems to devise innovative techniques. The following is an explanation of two such systems.

### 2.2.1 Speech Emotion recognition using Fourier Parameters [[5](#Wan15)]

In this system, the Fourier Parameters (FP) are extracted from the input speech, instead of extracting the highly common Fast Fourier Transform (FFT) parameters. The output data is then fed to a Select Vector Machine (SVM) classifier. This added a remarkable improvement to the ASER system.

* **Methodology**

The system uses three different databases German, Chinese and elder Chinese. The authors postulate that spectral features such as amplitude and phase are the best speech characteristics for emotion recognition. Hence, the extracted features are (MFCC), (FP), combination of (FP and MFCC), Zero Crossing Rate (ZCR), fundamental frequency (F0) and energy. The SVM classifier then differentiates between eight basic emotions: Happiness, Sadness, Anger, Fear, Boredom, Surprise, Neutral and Anxiety

* **Advantages**

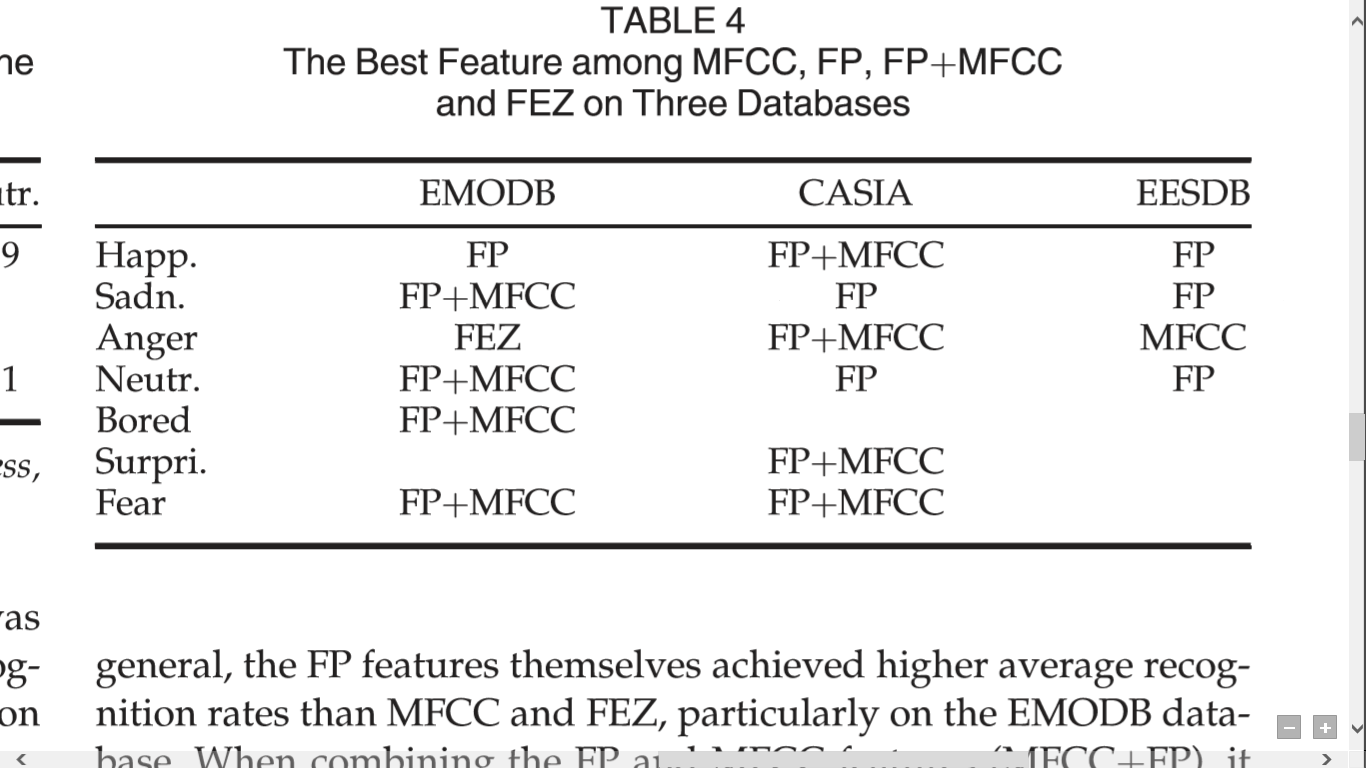
The first obvious advantage is that the system combines three different databases. This translates into an astoundingly large number of wave files (9600) that are used to train the SVM classifier. The paper also successfully identifies feature-emotion pairs, that is, which feature better recognizes which emotion. This distinction enriches the literature review required ASER research.

* **Disadvantages**

The system is deficient because it only focuses on spectral features. Moreover, as shown in Table 2.4 the system is not fixed and not standardized across all emotions. That is, some features are better suited to recognize certain emotions and are not optimum for other emotions. Table 2.4 illustrates the effective features used to give best result using (FP), (MFCC), (ZCR) or F0+Energy+ZCR (FEZ). Therefore, it cannot be implemented as a portable device, due to the irregularities of the system.

Table 2.2

Comparison between CASIA, EMODB, and ESDB [[5](#Wan15)]



### 2.2.2 Speech Emotion Recognition with prosody, quality and derived features using SVM classifier [[6](#Sam15)]

This system uses a large number of features along with a Multilevel SVM classifier to identify seven discrete emotional states: anger, disgust, fear, happy, neutral, sad and surprise. The corpus used is “Multilingual Emotional Speech Database of North East India” (MESDNEI) [[6](#Sam15)]. The average accuracy achieved by the combination of features is 82.26%.

* **Methodology**

The system uses a basic Finite Impulse response (FIR) filter for the preprocessing phase. The system then extracts four prosody features: Pitch, ZCR, Short-term Energy, and Log-entropy. For quality features, six are extracted: first three formant frequencies, Spectral Roll-off, Spectral flux, and Spectral centroid. Moreover, the system uses 14 Mel Frequency Cepstral, Coefficients, 12 Linear Predictive Coding Coefficients, and 13 Mel-Energy spectrum Dynamic Coefficients. As shown the block diagram in Figure 2.3, for each feature class, the system then takes four statistics (mean, standard deviation, max and range) to form a single feature vector for each utterance. This makes a feature vector of 196 features for each sample.

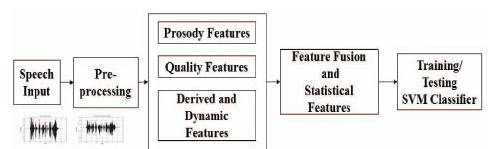
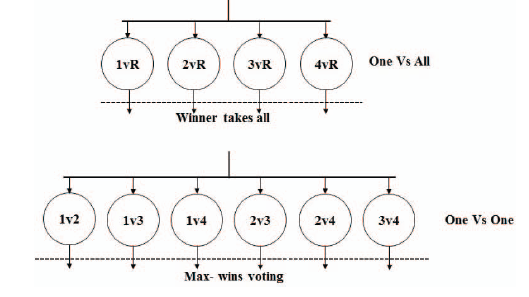


Figure 2.3: Steps for feature extraction [[6](#Sam15)]

There are different types of Multilevel SVMs. Figure 2.4 illustrates two of them. The first scheme is called a “one vs all” SVM, which means that each emotional state has its own SVM and that the input vector is compared by each one. The Second scheme is called a “one vs one” SVM, and in it one SVM compares pairs of emotional states. The decision process implemented by the system follows a “winner takes all” approach which means that as soon as the SVMs release their outputs, the one with the highest likelihood is chosen as the emotional state. This is different from a “max wins voting” approach, where individual SVMs must first cast their votes, and then a state is chosen.



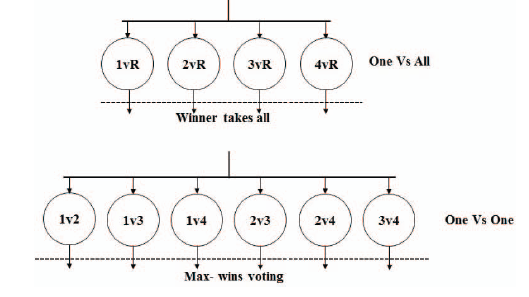


Figure 2.4: Types of SVM [[6](#Sam15)]

All five languages in the MESDNEI corpus are tested. Figure 2.5 demonstrates the accuracy rates for obtained from the Karbi language.

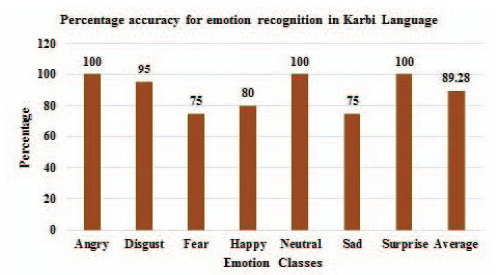


Figure 2.5: Emotion recognition accuracy in Karbi language [[6](#Sam15)]

* **Advantages**

The system’s most notable advantage is that it was successful in extracting a very large number of speech features — compared to current systems. The system’s use of a SVM as the classifier is another advantage, as it has a relatively low computational complexity and therefore will result in less power consumption. This last advantage makes it easier to later transform the system into a portable device.

* **Disadvantages**

The system faced a common disadvantage among ASER systems, and it’s the difficulty in distinguishing between anger and happiness. Another disadvantage is that although the system uses a multilingual corpus comprising of five languages, each language is trained and tested separately, and that is why the classifier performs well across all five. However, one would postulate that were the system to be trained by all languages at once, the results might be lower. Lastly, the system makes use of a huge number of features; this would entail substantial computation, and complicates the design requirements to a certain degree.

# Chapter III

# Proposed System:

# Real time Speech Emotion Recognition for Archetypal Emotions

## 3.1 Main features

The proposed system aims to extract spectral features of MFCC, LFPC, and LPC that deal with frequency domain and temporal characteristics: energy of signal, maximum amplitude and minimum energy. The implemented classifier is KNN. The proposed design aims to be speaker independent as well as context independent. The system, also, pushes the limitations of current systems, and attempts to establish a more generalized system, that performs well across different database, and thus cultures, and backgrounds. To achieve this, the system exploits multiple English corpora, while also working with an Arabic corpus that has been created specifically for this system. The range of frequency taken (0 KHZ to 4 KHZ) is based on speech maximal range and the 12th order is used for the LPC, which is the most efficient order for SER [[7](#Den03)]. Figure 3.1 demonstrates the top level of the proposed system. The proposed system design will be explained at length in the following sections. Input speech enters through a standard microphone, it’s preprocessed digitally before being accepted by the proposed system, where it’s analyzed and finally the output emotion is displayed.

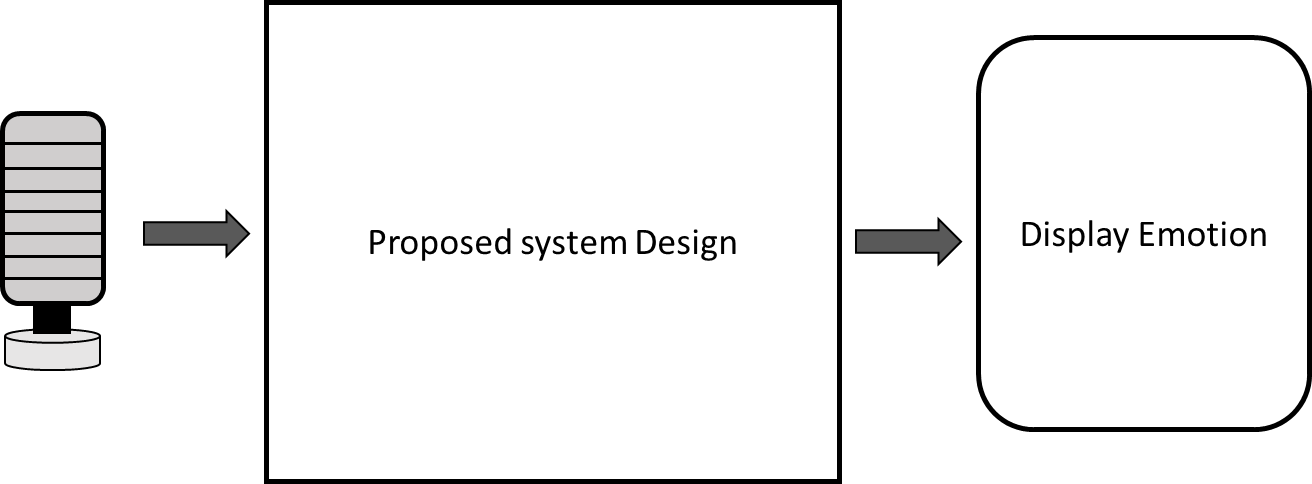


Figure 3.1: Proposed system hierarchical block diagram

## 3.2 Used Corpora

Two English speech emotion databases have been utilized by the proposed system. The first one Surrey Audio-Visual Expressed Emotion (SAVEE) database is a males-only corpus [[8](#Sur15)]. The database consists of recordings from 4 male actors who simulated 7 different emotions: anger, disgust, fear, happiness, surprise, sadness, and neutral. The total number of utterances is 480 British English ones. The sentences for each emotion were phonetically balanced and derived from the standard TIMIT corpus [[9](#Gar93)]. The second database Toronto Emotional Speech Set (TESS) is a female-only corpus [[10](#Tsp10)]. It also takes advantage of actors to portray the seven same archetypal emotions as [1]. Two female actors, one old and one young, spoke a list of 200 target words in the form of Say the word \_\_\_\_\_.” The database includes a total of 2800 stimuli.

In an attempt to make the proposed system more culturally pervasive, steps have been taken towards creating a multi-dialect Arabic database. The Arabic dialects are Egyptian, Levantine, and Gulf. Student actors from 6th of October University for Modern Sciences and Arts have been recruited to simulate the seven basic emotions. The sentences used for the utterances were chosen from Almeman and Lee’s multi dialect Arabic speech parallel corpora [[11](#Alm13)]. After the student actors finished recording their utterances, the database has been compiled; however, it consist of three different delicate which are; Lebanon, Egyptian and Gulf with number of 2 Males actors whit each with 20 different sentences.

## 3.3 Methodology

The system uses signal processing methods, with a focus on human speech processing, due to the nature of the project. Feature extraction techniques are implemented, namely: MFCC, LFPC, and LPC. More features are expected to be tested in the near future. The second highly important methodology utilized is machine learning. To establish the categorization between the sets of emotions, the proposed system resorted to using the K-Nearest Neighbors (KNN) classifier. The KNN distinguishes between six different archtypal emotions: happiness, anger, sadness, fear, neutral, and surprise.

## 3.4 System description

The proposed system follows the logical flow of feature extraction to classification and then decision making. Figure 3.2 depicts the block diagram of the proposed system. The first two blocks (Frame blocking, and FFT) are the preprocessing phase of the speech signal. The next three blocks (MFCC, LPC, and LFPC) are the responsible for feature extraction. The last two blocks (Feature processing, and K-NN) are in charge of arranging the extracted features, and the subsequent emotion classification.

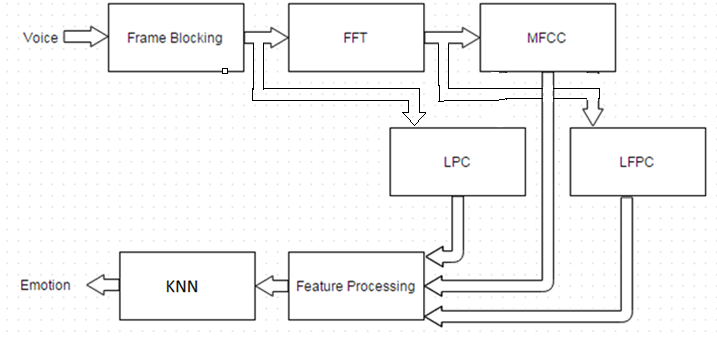


Figure 3.2: Proposed System block diagram

### 3.4.1 Frame blocking and Windowing

The purpose of this block is to take the input signal of the speech signal and split it one second intervals. Then this signal data of one second goes under second processing to divide it into small frames.

The frame duration is (~30ms) done with the hamming window process to minimize as much as possible the spectral distortion, with (3.1) [[12](#mat16)], where N is the order of the filter, which is equal to the filter length – 1.

(3. 1)

The framed data is then multiplied by the hamming window using (3.2). W(n) is the hamming window , Xi(n) is the data of the framed signal and Yi(n) is the output data of the windowed signal.

(3. 2)

### 3.4.2 FFT

In this procedure the output from the previous stage is taken as an input for this stage. FFT is used to transform the input time domain data to frequency domain using (3.3). The signal is transformed into its corresponding frequency domain due to the rich spectral features that could be extracted by working in that particular domain.

(3. 3)

### 3.4.3 MFCC

This feature is commonly used in signal processing and speech recognition due to its mobility and flexibility [[3](#Nic99)] [[5](#Wan15)] [[6](#Sam15)]. The audio of speech signal is constantly changing, due to its changing nature, the signal should be divided into number of frames; these frames have optimum duration from 20ms to 40ms. This step has been done in the previous stages of the signal preprocessing. The MFCC internal instruction is shown in Figure 3.3 [[13](#Tiw10)].

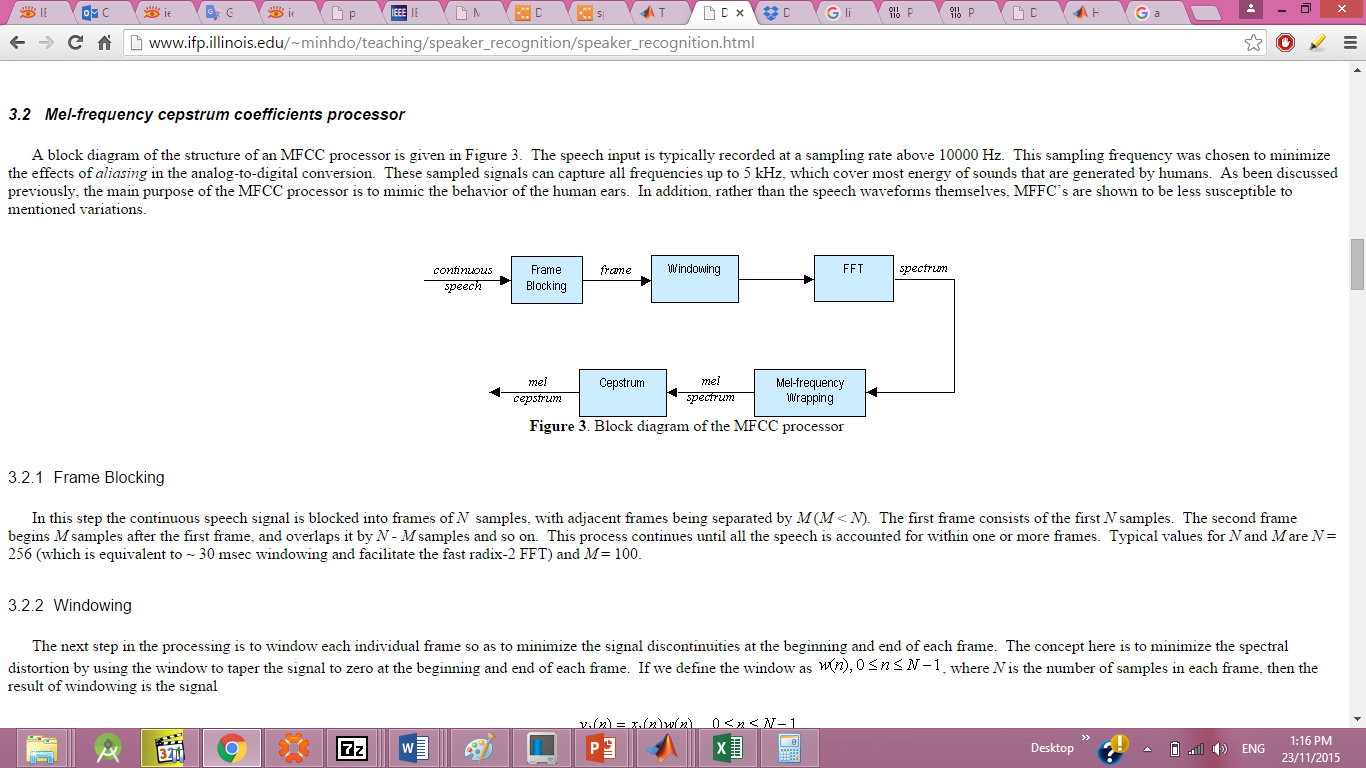
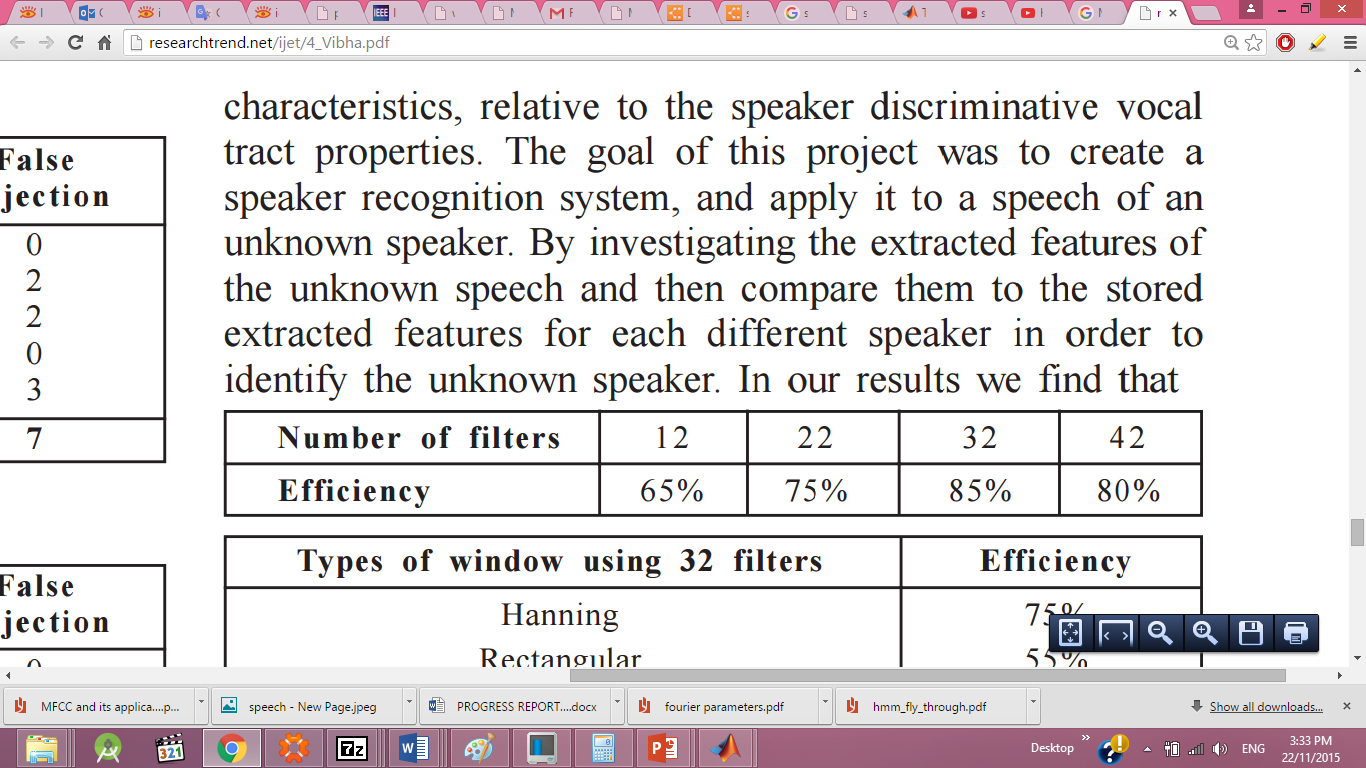


Figure 3.3: MFCC block diagram

The number of Filter bank used are 32 due to its high efficiency as described by comparing using 12, 22, 32 and 42 MFCC filters in illustrated in Table 3.1.

Table 3.1

Filter order to efficiency [5]



The main advantage of using the MFCC is that this feature simply gives good data and results that are capable of differentiating between the two categories (Sadness, Anger, and Fear) and (Sadness and Neutral) with high accuracy; it also neglects some noise, but still is not entirely immune to noise.

* **Mel-Frequency Wrapping**

Research concerned with the human voice proves that the speech perception of the human voice frequency doesn’t follow a linear scale. The Mel-frequency scale is a linear scale below 1000Hz and is logarithmic scale above 1000Hz.This aptly represents the tone of the human voice. To achieve this, the input signal is passed through the Mel-filter bank shown in

Figure 3.4.

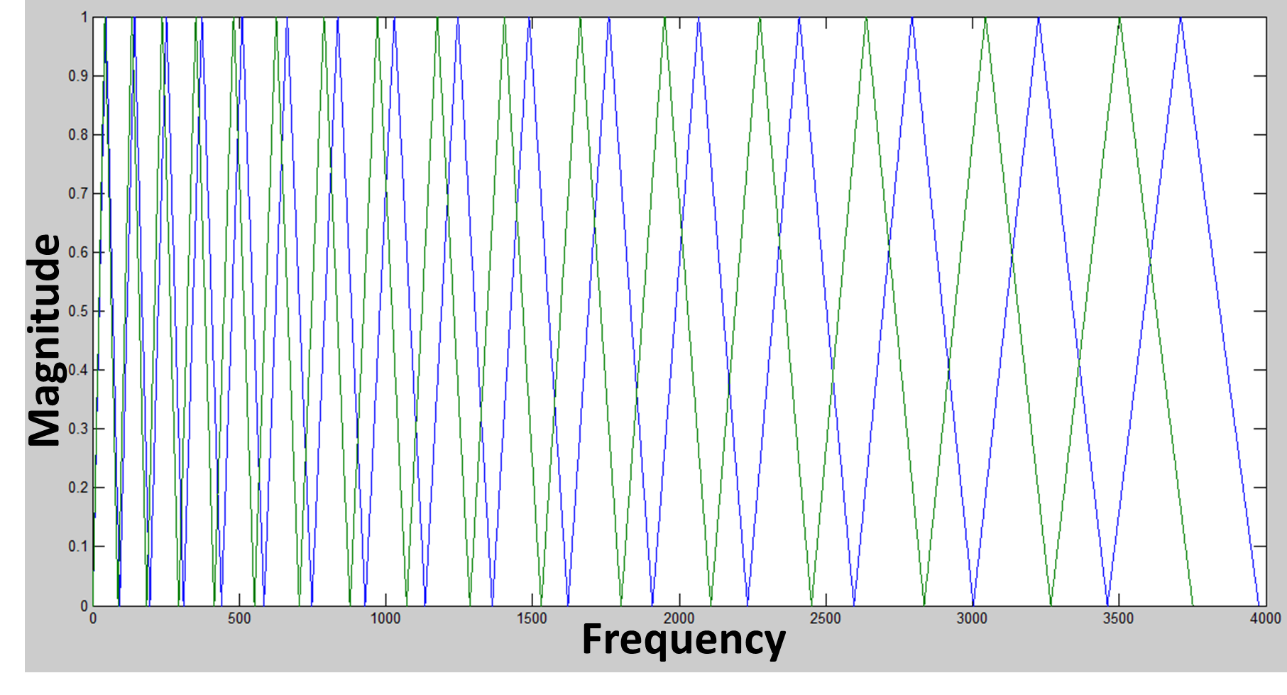


Figure 3.4: Mel-filter bank response.

Using (3.4) we convert from frequency domain to Mel scale.

(3. 4)

For vice versa, to obtain frequency from Mel scale (3.5) is used.

(3. 5)

### 3.4.4 LPC

LPC is a process utilized for the most part as a part of sound preparing and speech handling. It is a tool that uses the information of a linear predictive model to extract the spectral envelope of a digital signal of speech in compressed form [[14](#dat)] [[15](#Gol00)]. It is a standout amongst the most capable speech examination strategies and a standout amongst the most valuable systems for encoding great quality speech at a low bit rate and gives to a great degree precise assessment of speech parameters.

The order used is the 12th order. This means that the output is a polynomial of the 12th order. As the LPC order increases, the envelope resembles more closely the shape of the signal. The equation in 3.6 represents the polynomial used to extract the LPC coefficients.

**(3.6)**

Another important feature that could be extracted from the LPC is formant frequencies [[16](#Sne93)]. The use of format frequency extraction as a means for speech emotion recognition systems has been exploited before [[17](#Bag12)]. Formant frequencies are the roots of the LPC polynomial. Table 3.2 shows a comparison of the first 4 formant frequencies of two emotional states. It is obviously difficult to distinguish between two different emotions with the formant technique, due to the very small difference between formant frequencies between the two emotions Anger and Happiness; therefore, in order to not add unnecessary complexity to the system, it was decided to resort to LPC analysis alone.

Table 3.2

First five formant frequencies for anger and happiness states

|  |  |  |
| --- | --- | --- |
|  | **Anger** | **Happiness** |
| **Formant frequency f1 in Hz** | 721.3 | 832.6 |
| **Formant frequency f2 in Hz** | 1390.8 | 1525.9 |
| **Formant frequency f3 in Hz** | 2200.1 | 2239.9 |
| **Formant frequency f4 in Hz** | 2935 | 2946.3 |

# 3.4.5 LFPC

LFPC is another essential feature for the proposed system. It outlines the spectral shape of the signal. As shown in the diagram in Figure 3.5, LFPC analysis consists of the basic preprocessing blocks of hamming, windowing, FFT and filter bank, and then proceeds to log the result of the output to get the power. This process is applied to each frame of the signal [18].

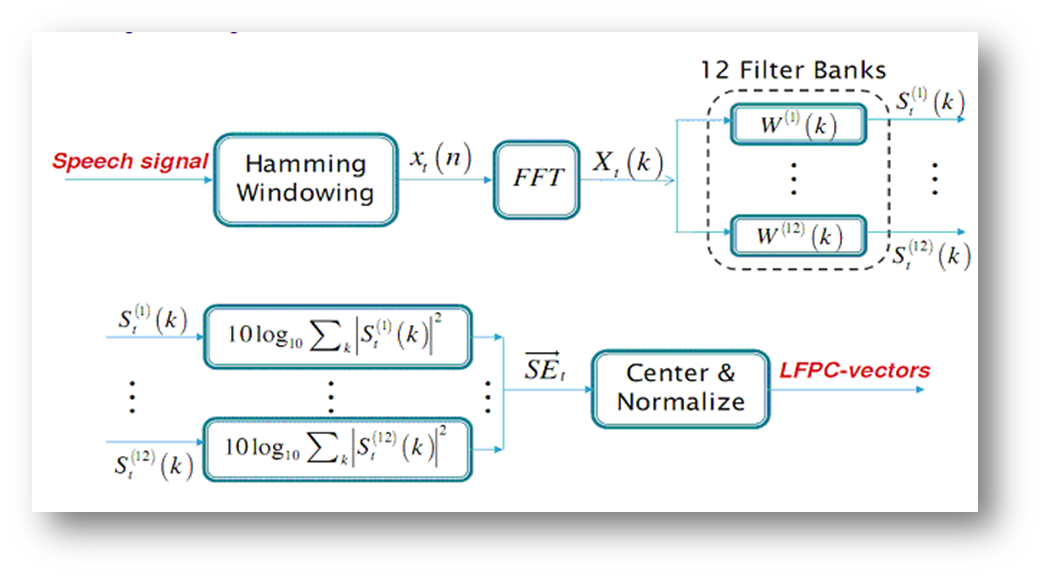


Figure 3.5: LFPC block diagram [18]

Figure 3.6 illustrates the different LFPCs of the same sentence spoken in three different emotions. The LFPC shows that it can differentiate between high arousal emotions, but there are some intersections that cause errors for the system. Applying this technique for the low arousal emotions gives better results and less intersection that gives higher accuracy by testing it on Matlab.

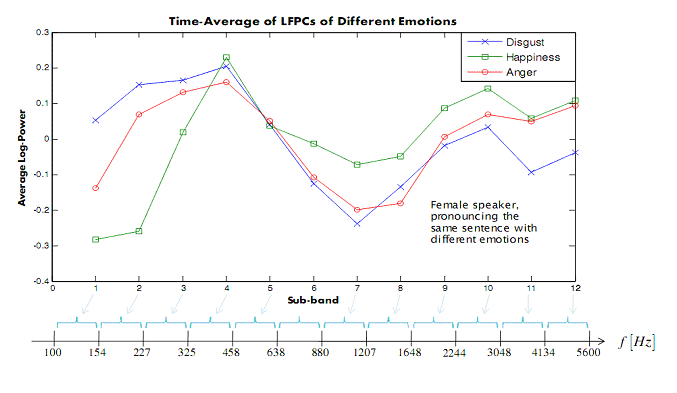


Figure 3.6: LFPC Analysis example [18]

## 3.4.6 KNN

K-nearest neighbors has been utilized for pattern analysis and for statistical approximation since the late 1970’s, when it was used for clustering [19]. Figure 3.7 represent KNN is a straightforward algorithm that takes stock of all previous cases (training phase) and classifies test cases on a similarity basis.

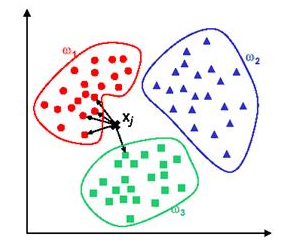


Figure 3.7: Decision making based on distance [20]

KNN uses a measure of “closeness” to classify a case. Figure 3.8 demonstrates how ‘xj’ is tested against its two closest clusters, and then determined to be belonging to ‘w1’. In KNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors [20].

KNN is an instant-based classifier, because all computation is done at the time of the classification. Training the system is merely a matter of ordering data in an appropriate fashion [21]. Some advantages of the KNN classifier are that it’s analytically tractable, it has a simple implementation, that it outputs nearly perfect results in the large samples, and that it takes advantage of local data [21].

In a regular KNN, the feature vector for each emotional state is assigned a cluster, and then any unknown instance is compared against the whole set to determine its closest neighbor, hence class.

After some manipulation and modification to LPC and LFPC for Male section, rather than by taking the resultant LPC and LFPC coefficients for 1sec for the trained utterances, it has taken for every 30ms. The differences are discussed in the following chapter. This technique allows the system to provide more space and more number of utterances that can be taken in the future that would increase emotions accuracy.

Such changes haven’t been applied on the LPC but also on the KNN classifier. The changes are; 30ms resultant coefficient have been considered and taken as a pattern and as trained data, but also the new modified KNN have been provided with new feature in classifying ,as it can compare an income new pattern with the stored pattern for two or more arrays of data that contain large number of database, and it can provide the system with the similarity percentage betwwen the new income pattern and the trainingdatabase.in simulation chapter demonstrates how this modification has been observed to increase the system accuracy substantially, by applying different utterances with different emotions; Sadness, Fear, Anger , Happiness and Neutral shows highest accuracy achieved by both conventional K-NN and modified K-NN.

## 3.5 Flow chart

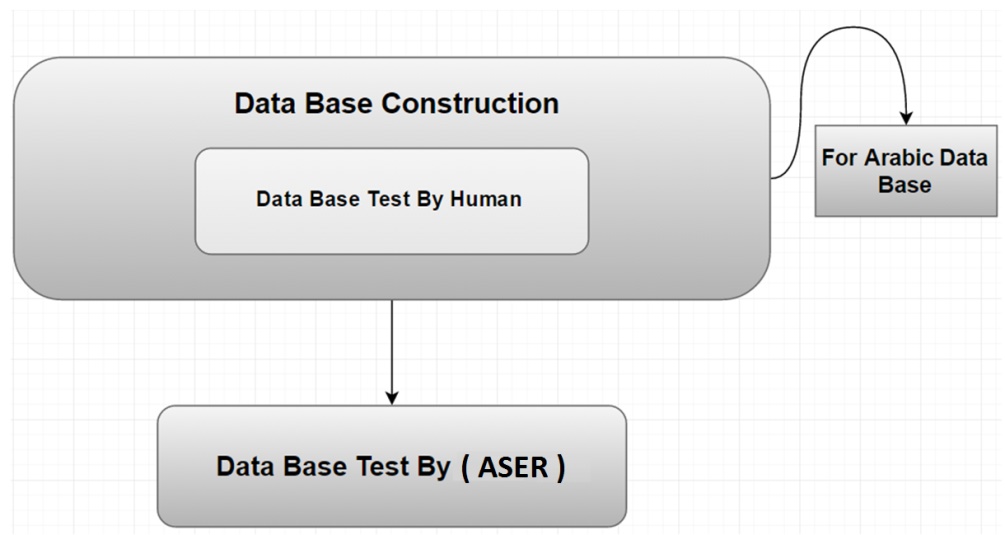
 Speech emotion recognition systems’ flow charts can be distinguished from one another by what’s included within each stage. Other than that, the flow charts are pretty much the same. Figure 3.8 clarifies the main idea of the ASER system, where the database is used to train and test such ASER system. Since one of the main objectives of the proposed system is to be compatible with two different languages, English and Arabic, the Arabic data base is first constructed then tested by human. Afterword, the flow is preceded normally where the database is used to train and test the required ASER system.

Figure 3.8: Decision construction

The machine learning algorithm technique is based on three major aspics which are; the database, speech feature extraction, and the classifier. Figure 3.9 shows how the database is used to train and test the proposed ASER system, where the extracted features are MFCC, LPC, and LFPC, while the Decision Making Classifier stage is comprised of the KNN classifier. On the other hand Figure 3.10 illustrates a detailed flowchart of the software algorithm, already prepared for the proposed system.

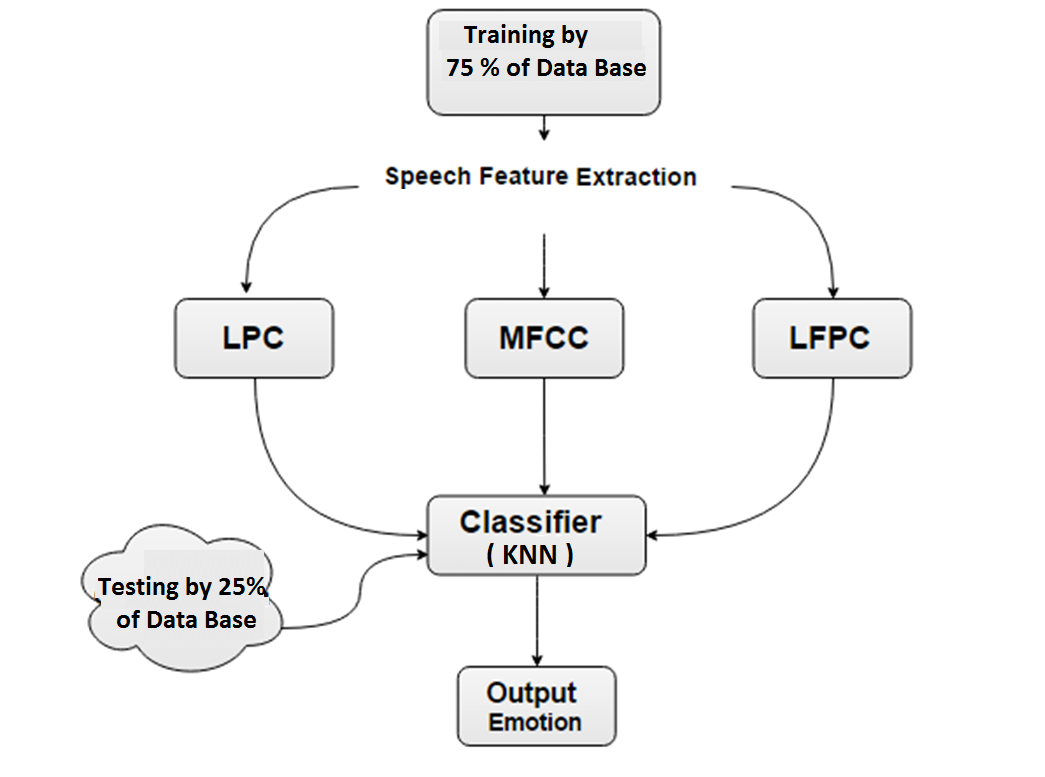


Figure 3.9: Main algorithm used to design the proposed ASER system

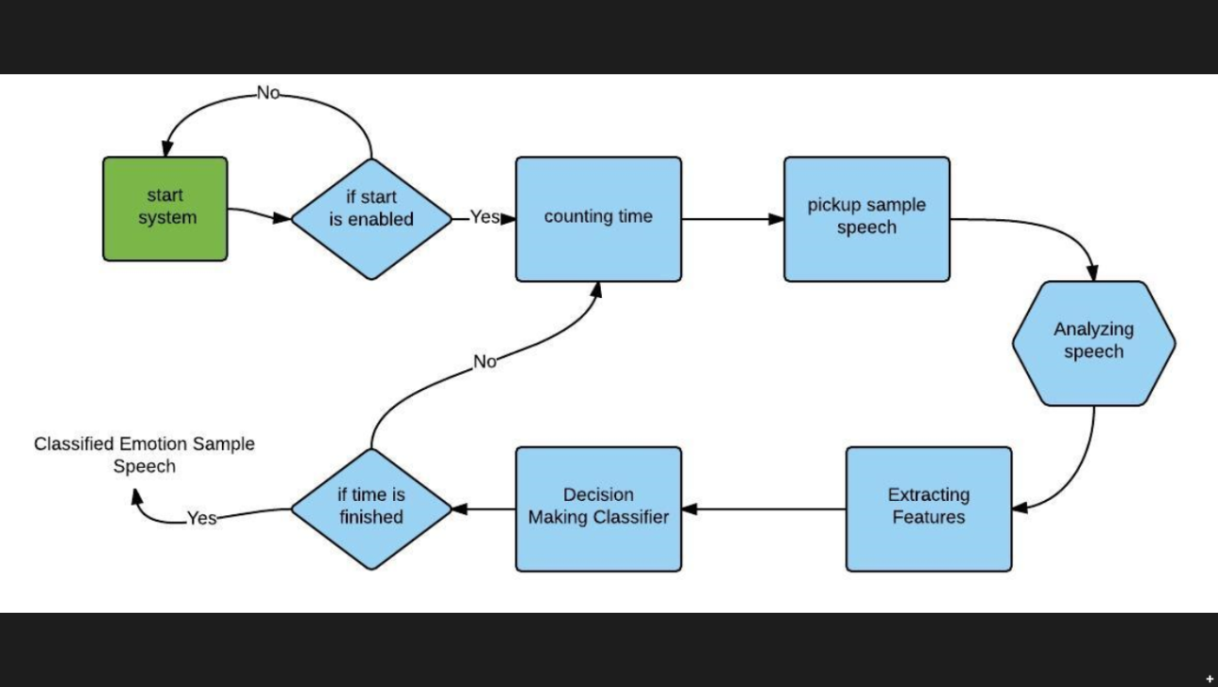


Figure 3.10: Proposed system flow chart

## 3.6 Testing method

The proposed system has been tested on MATLAB, due to its versatile capabilities with signal processing.

## 3.7 Summary

So far, all the components of the proposed system were discussed and explained in length. The main components for the speech feature extraction were MFCC, LPC, and LFPC. The classifier used is KNN.

In the next three chapters simulation and testing, cost analysis and project time plan are discussed.

**Chapter IV**

# Simulation and Testing

As previously mentioned, MATLAB is used to process the speech samples, to handle the feature extraction, and to compute the accuracy rates. In this chapter the simulation results of the developed system are demonstrated and discussed. Also the main problems encountered while testing are clarified.

4.1 Simulation Results

In this section the simulation results of each individual part of the system is demonstrated, highlighting the main role of each part in classification, then the simulation result of the overall integrated system is presented and discussed.

### 4.1.1 Simulation result of MFCC – KNN based design

The SAVEE male- based data base is applied on the MFCC feature with the modified KNN classifier, where 75% of the Data base samples are used to train the system while the rest is used to test it. Figure 4.1, illustrates all MFCC coefficients, extracted from the training database for high and low arousal, where the MFCC coefficients represent the spectral envelope of the utterances. The MFCC coefficients, for the remaining database, which represents the testing data, was also extracted, and then the KNN Classifier was applied to test the accuracy of the MFCC-KNN based ASER system. Simulation results show that the accuracy reached 71.1% and 90% for high and low arousal, respectively

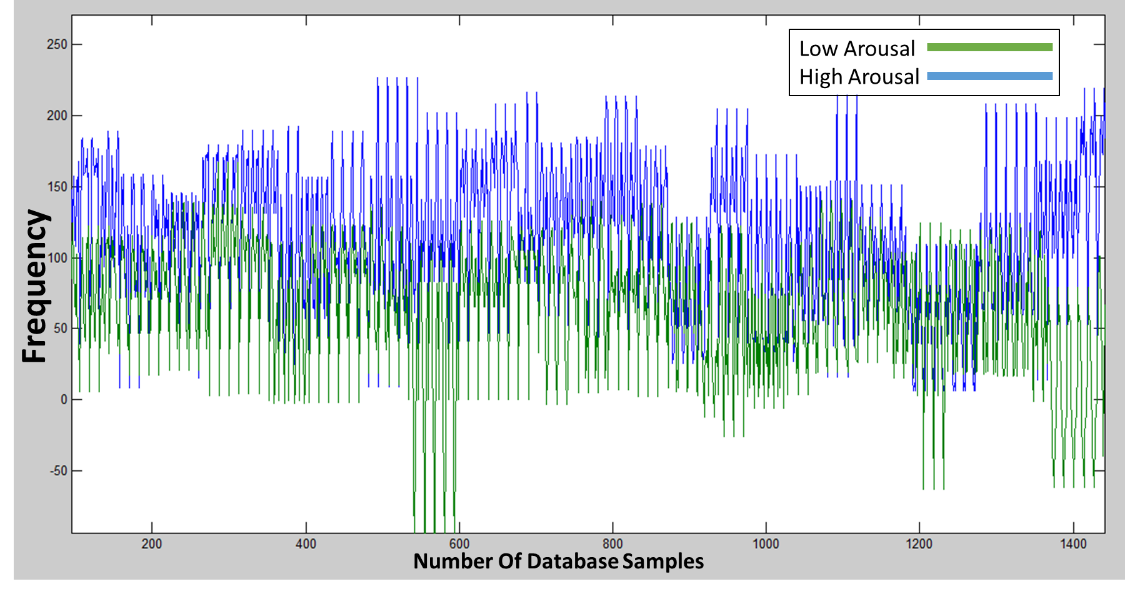


Figure 4.1: Extracted MFCC Coefficients of high and low arousal for SAVEE training database.

Similarly, the MFCC was also applied on the TESS female-based database along with the modified KNN classifier, where 75% of the database samples were used to train the system while the rest was used to test it. Figure 4.2, illustrates all MFCC coefficients, extracted from the training database for high and low arousal. The MFCC coefficients, for the testing database was also extracted, and then the KNN Classifier was applied to test the accuracy of the MFCC-KNN based ASER system. Simulation results show that the accuracy reached 87.5 % and 95.8 % for high and low arousal, respectively

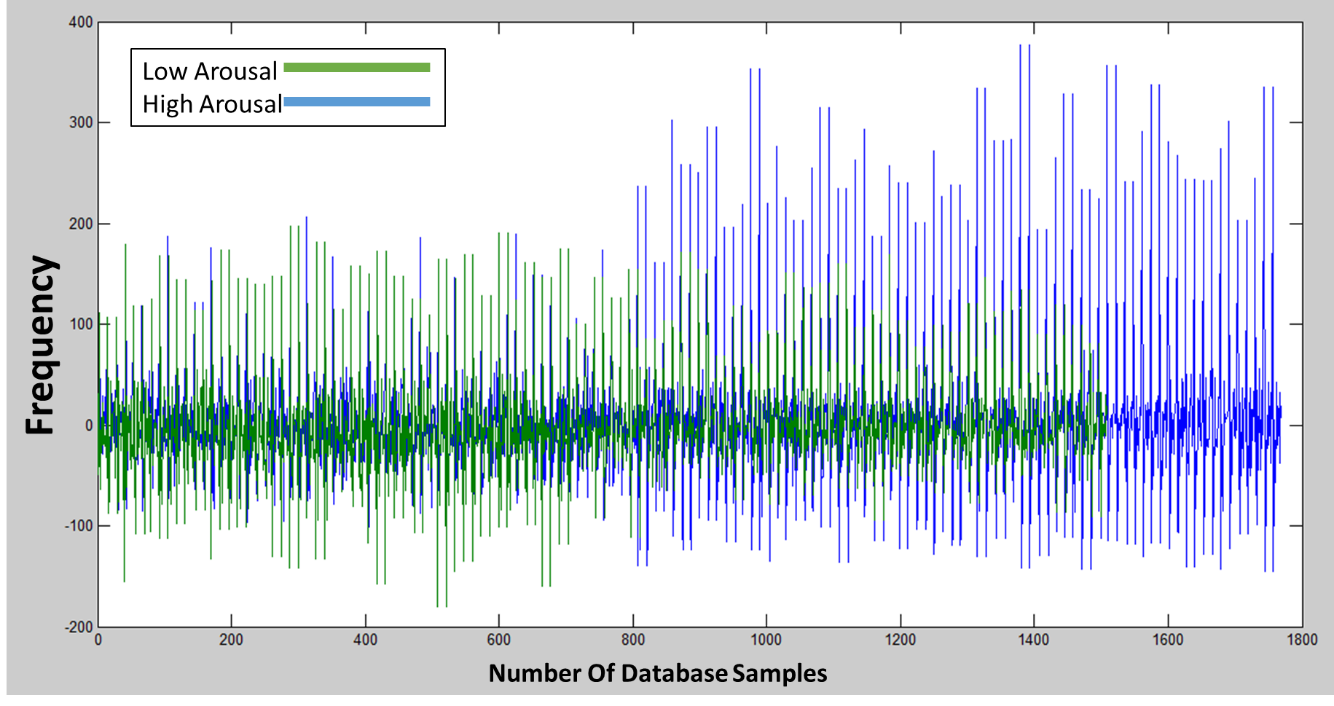


Figure 4.2: Extracted MFCC Coefficients of high and low arousal for TESS training database.

This shows that the MFFC feature can be successfully used to distinguish between high activation emotions such as Anger/Happiness/Fear and low activation emotions such as Sadness/Neutral. However it failed to distinguish between emotions of the same activation level such as Anger, Happiness and Fear, because they almost have the same range of frequencies, and their MFCC coefficients are so far similar.

### 4.1.2 Simulation result of LFPC – KNN based design

The SAVEE male- based data base is trained and tested by extracting the LFPC speech feature, with the modified KNN classifier, where 75% of the Data base samples are used to train the system while the rest is used to test it. Figure 4.3, illustrates all LFPC coefficients, extracted from the training database for two different emotions of the same arousal that are neutral and sadness (Low arousal emotions), where the LFPC coefficients represent the power of the utterances. The LFPC coefficients, for the remaining database, which represents the testing data, was also extracted, and then the modified KNN Classifier was applied to test the accuracy of the LFPC-KNN based ASER system. Simulation results show that the accuracy reached 100% and 86.5% for Neutral and Sadness emotions, respectively

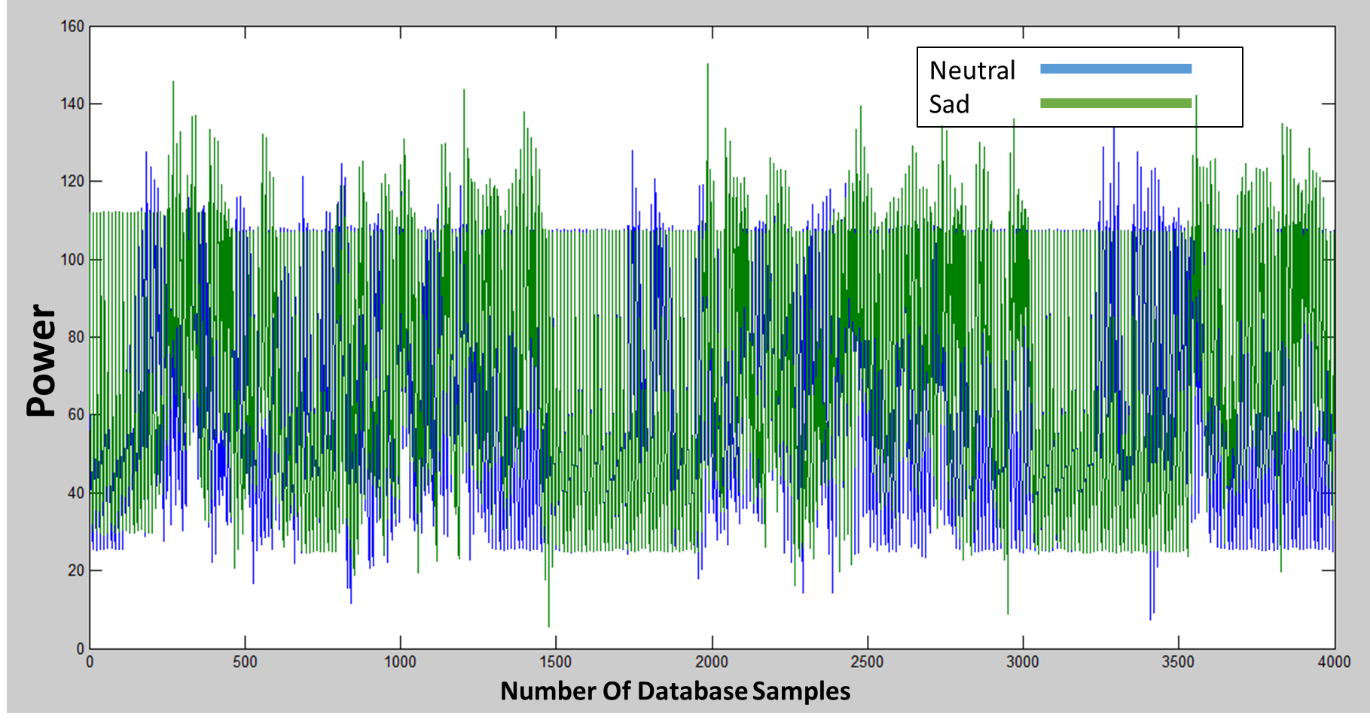


Figure 4.3: Extracted LFPC Coefficients of Neutral and Sadness emotions for SAVEE training database.

In comparison, LFPC was extracted from the TESS database and applied to the modified KNN classifier, where 75% of the database samples are used to train the system while the rest is used to test it. Figure 4.4, illustrates all LFPC coefficients, extracted from the training database for Neutral and Sadness emotions. The LFPC coefficients, for the testing database was also extracted, and then the modified KNN Classifier was applied to test the accuracy of the LFPC-KNN based ASER system. Simulation results show that the accuracy reached 100%, and 100 % for Neutral and Sadness emotions, respectively.

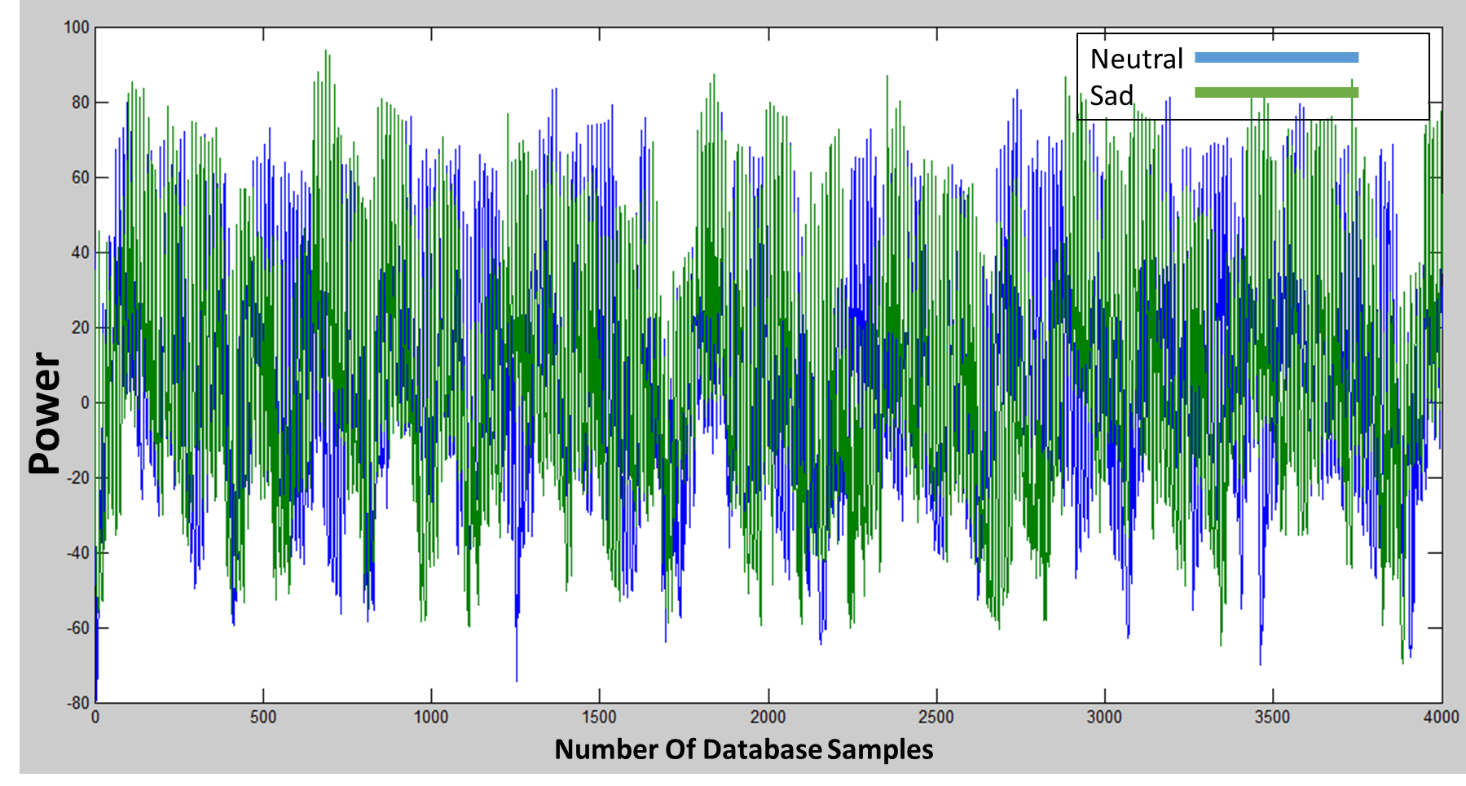


Figure 4.4: Extracted LFPC Coefficients of Neutral and Sadness emotions for TESS training database.

### 4.1.3 Simulation result of LPC & KNN based design

The same way, the LPC was extracted from the SAVEE database and applied to the modified KNN classifier, where 75% of the database samples are used to train the system while the rest is used to test it. Figure 4.5 demonstrates the LPC simulation results of two words from the same utterance, performed in two different emotions, Anger and Happiness that belong to the same arousal (high arousal). The LPC coefficients, for the testing database was also extracted, and then the modified KNN Classifier was applied to test the accuracy of the LPC-KNN based ASER system. Simulation results show that the accuracy reached 80 %, and 90% for Anger and happiness, respectively.

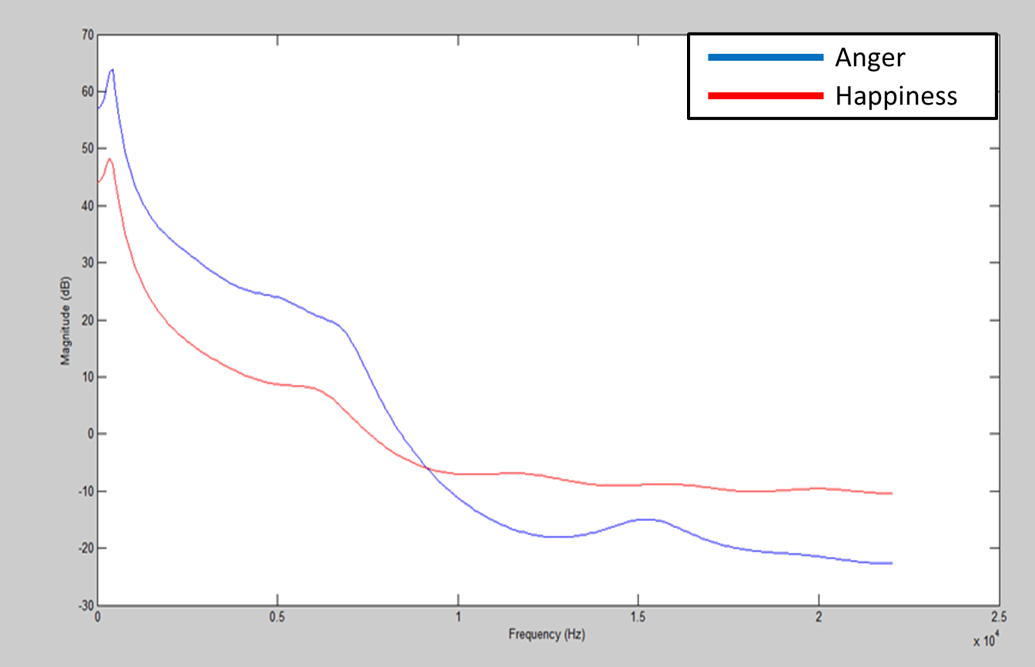


Figure 4.5: LPC coefficients of anger and happiness emotions for the same phrase, uttered in Anger and Happiness.

The same feature was used to differentiate between Fear and Happiness with accuracy reached 73.3% and 86.67%. Figure 4.6 demonstrates the simulation results of the LPC coefficients, extracted from the training database for Happiness and fear emotions.

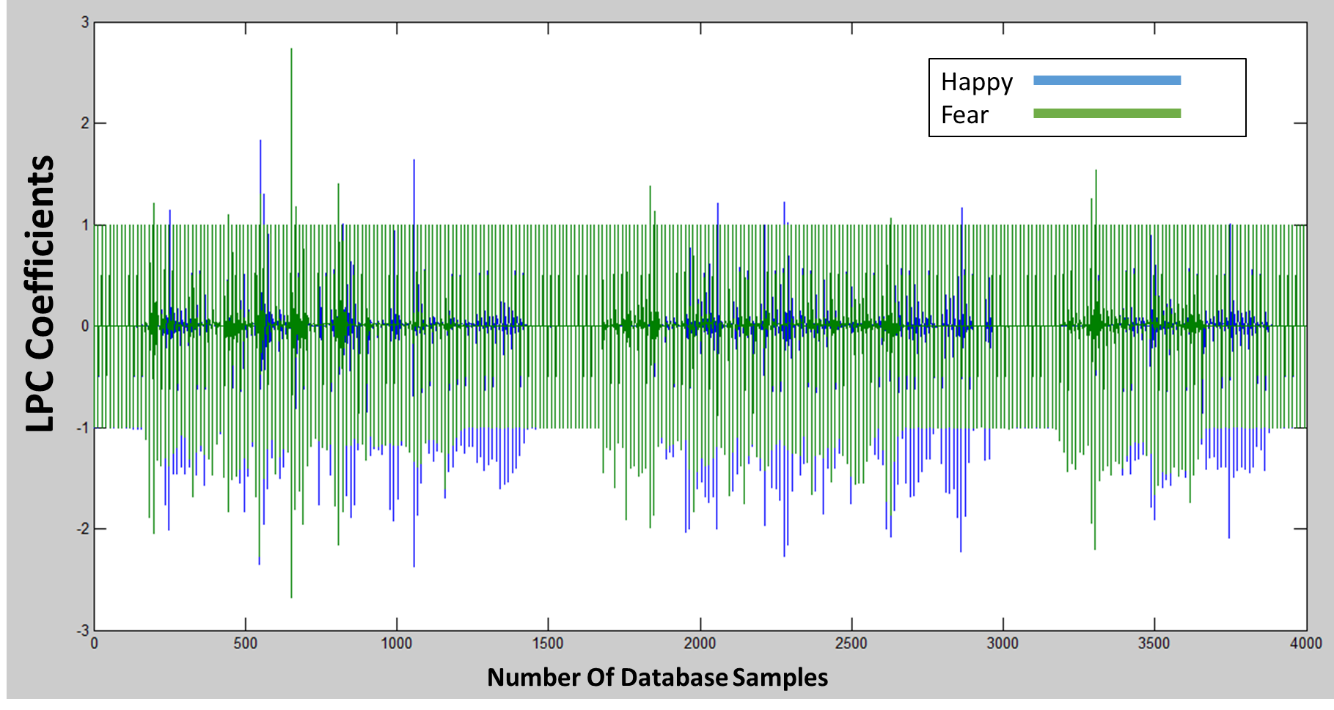


Figure 4.6: LPC extracted coefficients from the TESS training database for happiness and fear emotions.

Likewise, LPC was extracted from TESS database. After applying the modified KNN classifier, the accuracy for anger and happiness, respectively, reached 100% and 100%.

Figure 4.7 shows the LPC spectral envelope of the same phrase uttered in the anger state vs the happy state.

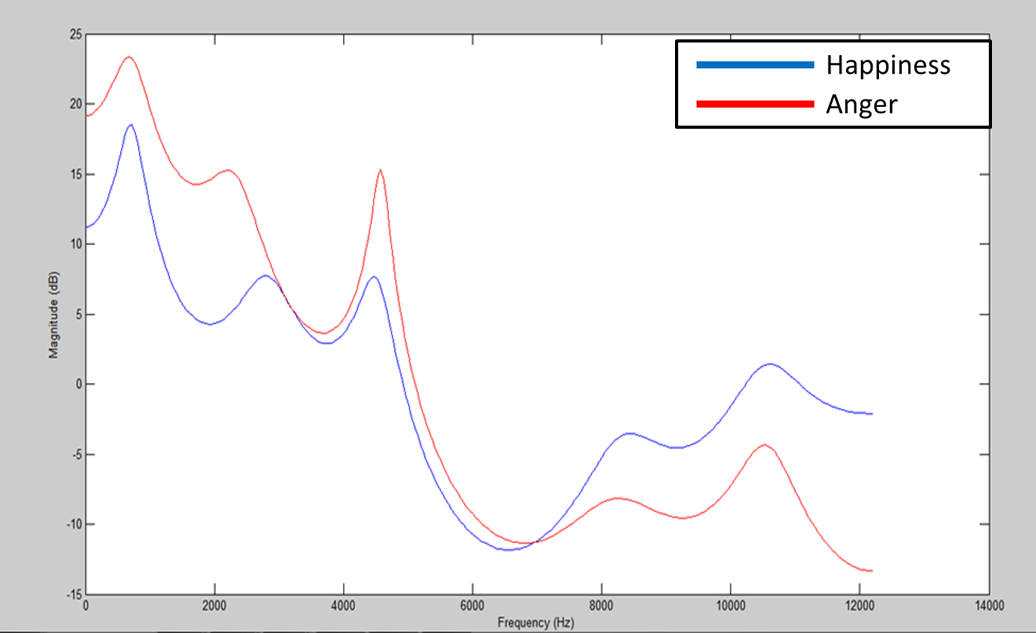


Figure 4.7: LPC coefficients of the same phrase uttered in anger state vs. Happy State from TESS

The same feature used to differentiate between Fear and happiness with accuracy reached 100% and 87.5% for Female TESS database. Figure 4.8 demonstrates the simulation results and gives description between samples of some date, trained in the system.

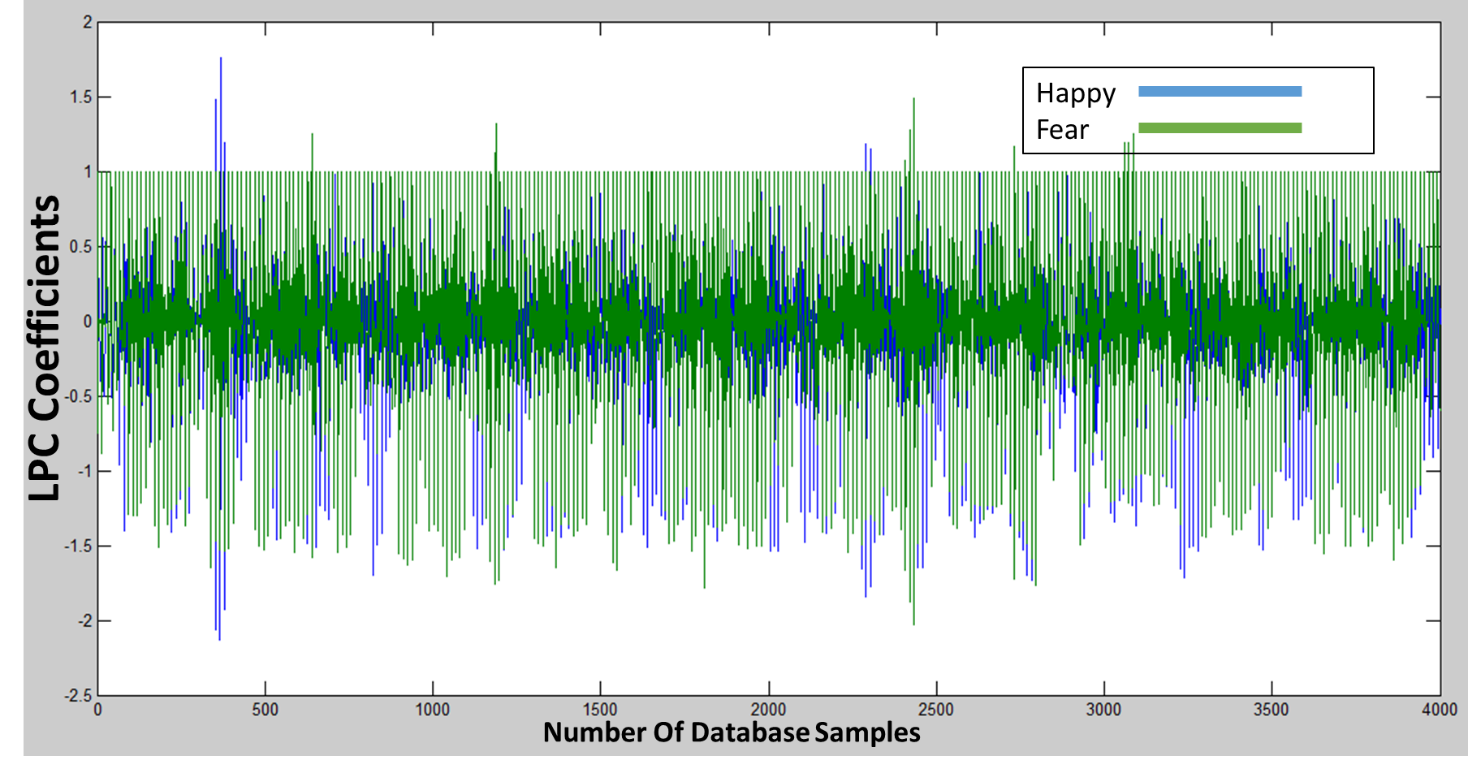


Figure 4.8: LPC for happiness and anger emotions extracted

From TESS

### 4.1.4 Simulation result of the integrated system

Figure 4.9 shows the flow chart that explains the decision tree used as MFCC is extracted first then according to the KNN classifier decision whether LFPC or LPC features are used to test the system.

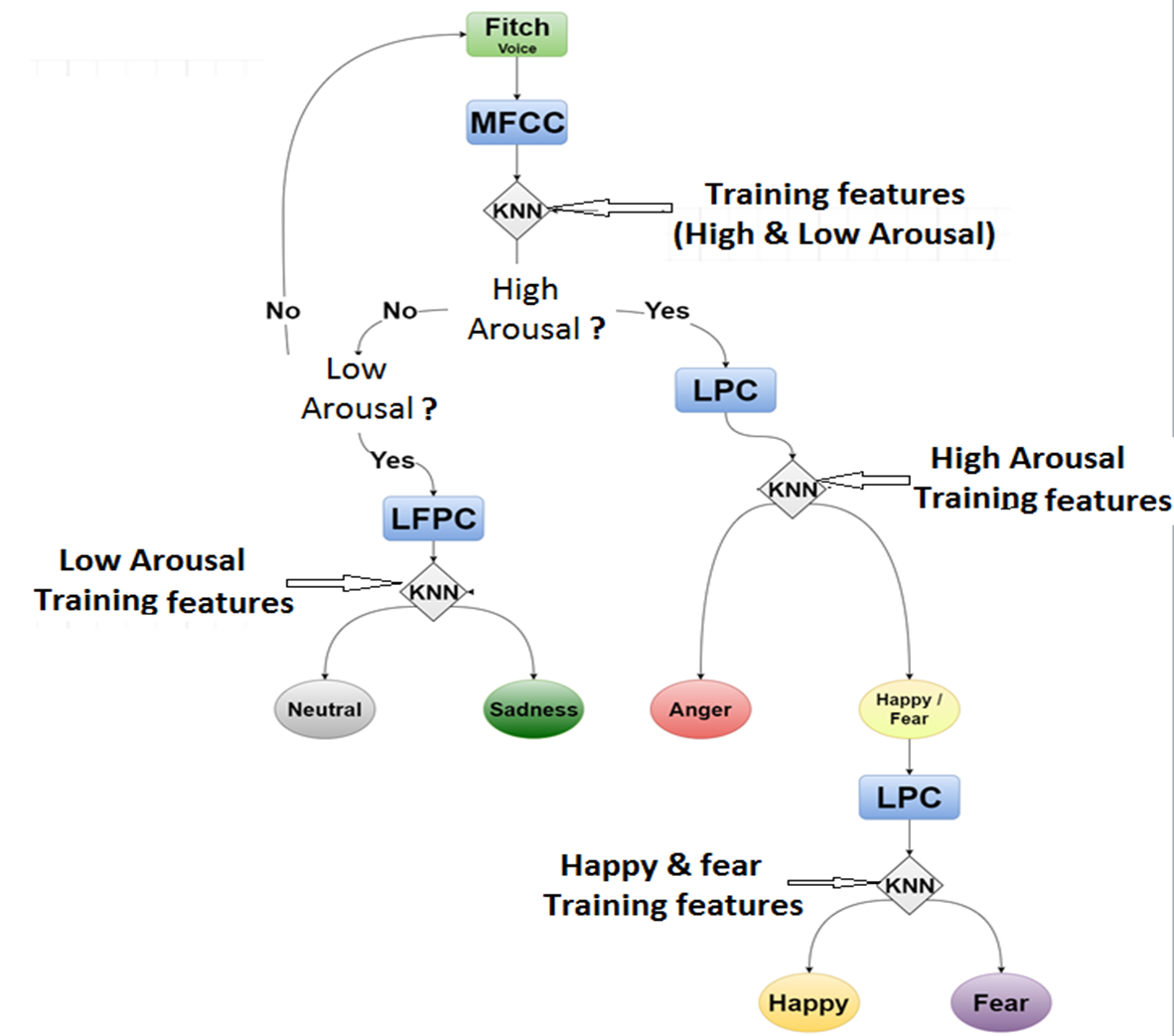


Figure 4.9: Binary decision tree Flow chart

After combining the three features MFCC, LFPC and LPC for SAVEE and TESS database, and creating the top level architecture database with the modified KNN classifier, the resultant accuracy for each emotion is estimated and demonstrated in Table 4.1.

Table 4.1: Accuracy rates for each emotion compared by database

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Database** | **Anger** | **Happiness** | **Sadness** | **Fear** | **Neutral** |
| **SAVEE** | 56.88% | 55.4% | 78% | 51% | 90% |
| **TESS** | 87.5% | 87.5% | 95.8% | 87.5% | 95.8% |

Testing the new created Arabic database which are represented by two Males with Egyptian, Gulf Dialect and Lebanon with number of 300 utterances gives accuracy illustrated in Table 4.2.

Table 4.2: Accuracy rates for Arabic database

|  |  |  |  |
| --- | --- | --- | --- |
| Nationality/emotion | Happy | Anger | sad |
| Lebanon | 82% | 58% | 64% |
| Gulf Dialect | 80% | 70% | 75% |
| Egyptian | 62% | 51% | 75% |

Table 4.3 shows how does the human testing have been done for estimating the Arabic database as it is not certificated as the English database.

Table 4.3: Human Testing For Arabic Database

**Egyptian Utterance**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Emotion/sentence | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | total |
| Happiness | 3/5 | 2/5 | 4/5 | 3/5 | 3/5 | 5/5 | 2/5 | 2/5 | 5/5 | 2/5 | 62% |
| Anger | 2/5 | 3/5 | 4/5 | 2/5 | 4/5 | 1/5 | 2/5 | 3/5 | 4/5 | 1/5 | 52% |
| Sadness | 4/5 | 3/5 | 3/5 | 4/5 | 4/5 | 5/5 | 4/5 | 3/5 | 4/5 | 3/4 | 75% |

**Lebanon Utterance**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Emotion/sentence | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | total |
| Happiness | 3/5 | 4/5 | 4/5 | 5/5 | 4/5 | 5/5 | 4/5 | 4/5 | 5/5 | 3/5 | 82% |
| Anger | 1/5 | 3/5 | 4/5 | 3/5 | 4/5 | 2/5 | 2/5 | 3/5 | 4/5 | 3/5 | 58% |
| Sadness | 4/5 | 3/5 | 2/5 | 4/5 | 3/5 | 4/5 | 3/5 | 3/5 | 3/5 | 3/5 | 64% |

**Gulf Utterance**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Emotion/sentence | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | total |
| Happiness | 5/5 | 4e/5 | 4/5 | 5/5 | 4/5 | 3/5 | 4/5 | 4/5 | 4/5 | 3/5 | 80% |
| Anger | 3/5 | 3/5 | 4/5 | 3/5 | 4/5 | 4/5 | 3/5 | 5/5 | 4/5 | 3/5 | 70% |
| Sadness | 5/5 | 3/5 | 3/4 | 4/5 | 3/5 | 5/5 | 4/5 | 3/5 | 3/5 | 4/5 | 75% |

Table 4.4 demonstrates how this modification has been observed to increase the system accuracy substantially, by applying different utterances with different emotions; Sadness, Fear, Anger, Happiness and Neutral shows highest accuracy achieved by both conventional K-NN and modified K-NN.

Table 4.4: Conventional and Modified KNN accuracy rates

|  |  |
| --- | --- |
| **Classifier** | **Achieved Accuracy** |
| **Conventional KNN** | 56% |
| **Modified KNN (Males)** | 90% |
| **Modified KNN (Females)** | 95.8% |

## 4.2 Simulation discussion

From the simulation results it can be deduced that each feature used satisfies its own target with different variation in accuracy; as MFCC can differentiate between two different categories; LFPC can differentiate between emotions from low arousal while LPC differentiates between emotions from high arousal.

It’s also worth mentioning that due to the binary decision tree employed by the system –that is recognition is done in stages – error rates propagate from one stage to the next, ultimately decreasing the accuracy rate of a stand-alone stage.

The proposed system is compared with two other current alternative systems, in terms of the overall combined accuracy of the same emotions applied, the database used, and the extracted speech features, as demonstrated in Table 4.5

Table 4.5: Comparison between different systems

|  |  |  |  |
| --- | --- | --- | --- |
| **System** | **Data base** | **Features Extracted** | **Accuracy** |
| **NN-based** [3] | Specially collected speech database  (Japanese) | MFCC  Energy  F0 | 52% - 62% |
| **SVM-based** [6] | MESDNEI | Pitch  ZCR  Short-term Energy  Log-entropy  first three formant frequencies  Spectral Roll-off  Spectral flux  Spectral centroid  14 Mel Frequency Cepstral Coefficients  12 Linear Predictive Coding Coefficients  13 Mel-Energy spectrum Dynamic Coefficients | 82.6% |
| **Proposed system** | SAVEE  &  TESS | MFCC  LFPC  LPC | 51% - 95.8% |

Figure 4.10 illustrates the User friendly GUI that is prepared to facilitate using the proposed system as a desktop application, by choosing either male or female at the beginning, and then the user starts speaking. Afterword, the program estimates and detects the user’s emotion, based on the training features that are determined, according to the decision tree of Figure 4.9.

****

Figure 4.10: User friendly GUI

## 4.3 Problems encountered

One of the main problems that faced the system is the low number of database utterances available. Also, utterance time is small (~2s) which is not ideal to confidentially determine an emotion. These two problems have been solved by increasing the data taken by the utterances, and modifying the KNN-classifier by arranging the database given by the features, then taking the nearest three stored data corresponding to the new input data to be estimated, and the result of this K-NN will be as a voting system. Also if the new input data hits the same target of one of stored emotions database then it will neglect the voting system and resultant K-NN classifier will be the hitting target database emotion, and taking more time from the user before estimating the resultant emotion.

## 4.4 Summary

Using modified K-NN gives better accuracy, also using the highlighted features: MFCC, LFPC and LPC distinguishes effectively between different emotions without taking long time in estimating. So far, simulating the proposed system on MATLAB achieved maximum accuracy of 95.8% and minimum accuracy of 51%, which is so far comparable to other alternative speech emotion systems. For the Arabic database highest achieved accuracy is 81.8% and minimum accuracy reached 51%.

# Chapter V

# Cost Analysis

|  |  |  |
| --- | --- | --- |
| Component | **Cost per unit** | **Economical cost**  **for 100 units** |
| **Raspberry Pi**  **B+** | 19.99$ | 1450$ |
| **Mobile phone (Android)** | 63.98$ | **-----** |
| **Desktop app.** | ------- | ------- |

Table 5.1: Component comparison by cost

From Table 5.1 it can be noticed that:

Although the cost of the Raspberry Pi is less than a mobile phone, but in fact the processor of an android mobile phones way more capable than Raspberry Pi . However, the majority of people have mobile phones [22] [23] [24]. For a costumer services application, developing the system on a desktop will be more effecient than implementing it on a hardware platform. In this case only the cost of the software program, downloaded on the desktop, should be added.

# Chapter VI

# Time Plan

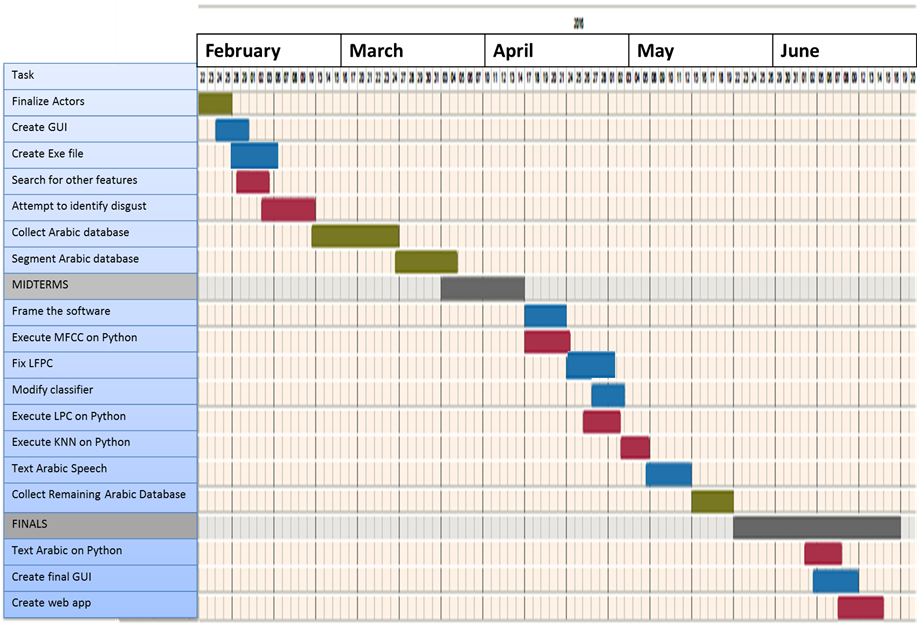
Table 6.1

Task Distribution and Time Plan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task No.** | | | **Task Holder** | **Task** | | **Exact Execution Time** |
| **1** | | | **Yosif** | **Creating an executable file** | | **5 days** |
| **2** | | | **Fatima** | **Modifying software**  **Plat form** | | **5 days** |
| **3** | | | **Yosif** | **Creating real time GUI system on Matlab** | | **4 days** |
| **4** | | | **Fatima** | **Searching for another features** | | **1 week** |
| **5** | | | **Fatima** | **Collecting the New Arabic database from Actors** | | **2 weeks** |
| **6** | | | **Yosif** | **Analysing actors speech through Matlab** | | **2 weeks** |
| **7** | | | **Fatima** | **Testing the new database accuracy** | | **1 week** |
| **8** | | | **Yosif** | **Crating GUI and exe. File** | | **1 week** |
| **9** | | | **Yosif** | **Finalize the Arabic database (Egyptian and Levantine Female)** | | **1 Week** |
| **10** | | | **Fatima** | **Analyzing the Arabic Database** | | **1 week** |
| **11** | | | **Yosif & Fatima** | **Improve the accuracy** | | **2 weeks** |
| **12** | | | **Yosif** | **Create GUI** | | **5 days** |
| **13** | **Fatima** | **Create Executable File** | **4 days** | |

Table 6.2

Gantt Chart



Green- Both

Blue- Yosif

Red- Fatima

Gray- Exams

# Chapter VII

# Future Plan

* Experiment with other classifiers
* Remake the Arabic database with proper professional actors.
* Transfer the system into an opensource platform e.g. Python
* Add text sentiment analysis
* Add social media monitoring and analysis
* Design web app
* Publish paper
* Find customers to beta test

# Chapter VIII

# Conclusion

## 8.1 Conclusion

In conclusion, speech emotion recognition endeavors are highly important due to their many uses. Speech emotion recognition systems can be distinguished from one another by three defining aspects. These three aspects are: Database used, features extracted, and type of classifier. Most previous speech emotion systems have attempted to find the most effective algorithm, by manipulating any or all of these previously stated aspects in order to achieve higher accuracy.

In the proposed system, rather than only using a frequently cited English database, an Arabic data base is also constructed that doesn’t currently exist in speech emotion systems proposed system is optimized for general databases so that the system yields sound results across a broader spectrum. As for the speech features that we are relied on: they are MFCC, LPC, and LFPC. The extraction algorithms of these features are built from scratch in order to have the ability to tweak their fundamental variables to better suit the application specification. For classifier, it has been resorted to using KNN, which is a machine learning algorithm with low computational complexity. After experimenting with the conventional KNN design, the accuracy was between 54-56%. Modifications to the KNN resulted in a much higher accuracy rate of 51% - 95.8%.

Lastly, the proposed system is transformed into a desktop application system with a user-friendly front end that can be easily marketable. To satisfy telemarketing market needs.

# References

|  |  |
| --- | --- |
| [1] | P. Vauki and C. Aravindan, "Improving emotion recognition from speech using sensor fusion techniques," in *TENCON 2012 - 2012 IEEE Region 10 Conference*, Cebu, 2012. |
| [2] | K. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using MFCC and wavelet features," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, Ghazibad, 2013. |
| [3] | J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion recognition in speech using neural networks," in *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on*, Perth, 1999. |
| [4] | P. Orponen, "Computational Complexity of Neural Networks: a Survey," Helsinki, 1994. |
| [5] | K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," *IEEE Transactions on Affective Computing,* vol. 6, no. 1, pp. 69-75, January 2015. |
| [6] | A. Samantaray, K. Mahapatra, B. Kabi and A. Routray, "A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages," in *IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)* , Kolkata, 2015. |
| [7] | L. Deng and D. O'Shaughnessy, peech Processing: A Dynamic and Optimization-Oriented Approach, 1 ed., CRC Press, 2003, p. 49. |
| [8] | Perso.ee.surrey.ac.uk, 2 April 2015. [Online]. Available: http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/. [Accessed 16 January 2016]. |
| [9] | Garofolo, John and e. al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," in *Linguistic Data Consortium*, Philadelphia, 1993. |
| [10] | Tspace.library.utoronto.ca, 2010. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487. [Accessed 16 January 2016]. |
| [11] | K. Almeman, M. Lee and A. Almiman, "Multi Dialect Arabic Speech Parallel Corpora," in *1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013. |
| [12] | mathworks.com, 21 January 2016. [Online]. Available: http://www.mathworks.com/help/signal/ref/hamming.html. |
| [13] | V. Tiwari, "MFCC and its Applications in Speaker Recognition," *International Journal on Emerging Technologies,* vol. 1, no. 1, pp. 19-22, 2010. |
| [14] | data-compression.com, "Speech Compression," 2016. [Online]. Available: http://www.data-compression.com/speech.shtml. |
| [15] | G. B. and M. N., Speech and Audio Signal Processing: Processing and Perception of Speech and Music, 1 ed., New York: Wiley, 2000. |
| [16] | R. Snell and F. Milinazzo, "Formant Location from LPC Analysis Data," *IEEE Transactions on Speech and Audio Processing,* vol. 1, no. 2, pp. 129-134, April 1993. |
| [17] | V. Bageshree, Sathe-Pathak and A. R. Panat, "Extraction of Pitch and Formants and its Analysis to Identify 3 different Emotional States of a Person," *International Journal of Computer Science Issues (IJCSI),* vol. 9, no. 4, July 2012. |
| [18] | Nwe, T. L., Foo, S. W., & Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication,* *41*(4), 603-623. doi:10.1016/s0167-6393(03)00099-2 |
| [19] | L. Rabiner, "Speaker independent recognition of isolated words using clustering techniques," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79,* vol. 4, pp. 574-577, 1979. |
| [20] | byclb.com. [Online]. Available: https://www.byclb.com/TR/Tutorials/neural\_networks/ch11\_1.htm. |
| [21] | fon.hum.uva.nl, May 2008. [Online]. Available: http://www.fon.hum.uva.nl/praat/manual/kNN\_classifiers\_1\_\_What\_is\_a\_kNN\_classifier\_.html. |
| [22] | Amazon.com, 2016. [Online]. Available: http://www.amazon.com/dp/B00ME5XUAG?psc=1. |
| [23] | ebay.com, 2016. [Online]. Available: http://www.ebay.com/itm/XILINX-SPARTAN-6-XC6SLX9-EDA6-1-FPGA-Microblaze-SOPC-Development-System-/271934757904?hash=item3f50931c10:g:kBcAAOSw3yd. |
| [24] | Amazon.com, 2016. [Online]. Available: http://www.amazon.com/Samsung-Factory-Unlocked-Android-SmartPhone/dp/B00D2S3P6I/ref=sr\_1\_cc\_2?s=aps&ie=UTF8&qid=1448561206&sr=1-2-catcorr&keywords=Samsung Galaxy Star S5282. |

# Appendix A

Description: Description: Z:\forms\greenwich.jpg

جامعة أكتوبر للعلوم الحديثة والآداب

كلية الهندسة

قسم هندسة نظم الاتصالات و الالكترونيات

"تمييز المشاعر النمطية من خلال الخطاب في الوقت الحقيقي "

مشروع التخرج

قدم ضمن متطلبات درجة البكالوريوس فى الهندسة الكهربية والحاسبات

الجزء الثاني

أُعد بواسطة

فاطمة فودة 121649

يوسف عاطف 120723

يشرف عليها  
الدكتورة نهى يونس

2015–2016

ملـخص ا لــبحــث

المشاعر , هي الصفة المتباينه بين الناس , منها يحدد مدى صدق نواياك الداخلية و صحة مفهوم الكلام . تحديد و اختيار الشعور لدى الانسان بواسطة الاجهزة سينتج عن ذلك ظهور عصر جديد من التطور في مجال الاتصال و نقل المعلومات. بالرغم من العديد من التطورات الملحوظة في مجال تحديد الشعور , الا ان هذا التقدم لم يرتقى به الى ان يتحدث الانسان بطلاقة و سهولة مع الجهاز. ويرجع هذا الى ضبط الدقة لدى برنامج الجهاز الخاص باستشعار الاحساس الداخلي لدى المرء. بالاضافة الى ذلك ان غالبية الطرق المستخدمة في هذا المجال معقدة , حيث من الصعب تصنيعه على اجهزة محمولة.

في هذا التقرير نود ان نوضح انه يوجد العديد من الطرق المختلفة لتحديد تباين الشعور و الاحساس, يرجع هذا الى انه يوجد العديد من الطرق لاستخلاص ملامح و مميزات الصوت و تصنيفها.من هذا المنطلق من الممكن تطوير برنامج فعال ممزوج بطرق عدة ينتج منه دقة اعلى و اجود و اقل تعقيدا , حيث من الممكن تحقيقه على ارض الواقع . ايضا النظام المقطرح هادف الى استخدام لغتين مختلفتين من الصوت (العربي و انجليزي) كقاعدة بيانات.

في الجزء لاول توظف العماليات الحسابية و التقنية من النظام المقترح باستخدام ال.Matlab كمعالج للاشارات و بمحاكة شكل الناتج و بمقارنة للنظم الحالية على ارض الواقع. ولقد تم قياس دقة البرنامج بنجاح و اظهرت النتائج دقة 95.8%51% - للغة الانجليزية باستخدامMFCC ,LFPC ,LPC كمميزات استخلاص و KNN معالج تفرقة بين خصائص مميزات الصوت لإستخلاص المشاعر و للغة العربية اظهرت النتائج دقة ما بين 51%- 81.8 % .