Customer Satisfaction Based on Voice Signals

by

Siyamamkela Bomela

Thesis presented in fullfilment of the requirements for the degree of Honours in Computer Science at the University of the Western Cape

Supervisor: Reg Dodds
Co-supervisor: Mehrdad Ghaziasgar

November 2017

Contents

Declaration

I,**Siyamamkela Bomela**, declare that this thesis *"Customer Satisfaction Based on Voice Signals"* is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or qouted have been indicated and acknowledged by means of complete references.

Signature: ..... Date: .....

Abstract

This paper focuses on the requirements that should be implemented for the above mentioned system, the requirements recorded on the first chapter are solely based on the users' perspective. The user will be expected to clearly define what problems' they have been facing that the system will be expected to solve, This will help to clearly outline what the systems capabilities will be, and what its limitations should be.

The first chapter will not require any formal knowledge of IT/Computer Science as it is meant for the users. The programmers point of view and expectations for the system will be disregarded. The implementation and design process will not be discussed.

The second chapter focuses on the programmers view. The programmer will be expected to analyze the requirements the user had in the first chapter. The programmer will define the problem in their own view. After this process the programmer will be expected to come up with a solution for this problem. After this process the programmer will discuss related work then finally come up witrh an implementation plan for creating the system.

The purpose of this document is to assist in the software engineering phase, It will assist the developers and designers to know what is required by the user.

Chapter 1

User Requirements Document

1.1 Introduction

The following are ideas of how the user wants the system to function, what its expected outcomes are. After the users view a short and long description of the problem will be given and the expected limitation of the software.

1.2 Users' view of the problem

Emotion recognition is the ability to tell how a person is feeling based on the response on their face or from the tone of their voice, The aim of the project is to determine the satisfaction of a customer, whether the customer was happy or not with the service that he was offered. The customer satisfaction can be classified into three different categories; positive(which could be happy, ex

d, suprised), nuetral(No emotion detected) or negative(sad, angry, dissapointed).

The user requires a system that will be viable for all customers, whether a sale is done in the store or telephonically. The system should be able to determine what the overall emotions the customer went through during the sale.

The will also be a matter of accuracy of the results, as people have different signals of expressing their emotions. The system willl have to be able to accommodate any user.

1.3 Description of the problem

The user require a system that will be able to determine whether s customer is happy with the service they have obtained. Numerous systems have been implemented for such a service, but these systems are mostly based on facial expressions. They use a camera to look at the changes in the customers face to determine emotion. The limitations of such systems are that the person has to be in the same room as the camera to determine the satisfaction.

Organizations such as call centers rely on voice to voice interaction to provide services or make sales, which means the system of facial expression cannot be used. Thie main problem that this paper aims to to tackle is to create a system that will be able to work for both kind of scenarios mentioned above, the system should accommodate sales made in store and sales that are made through call center sales agents.

The proposed system should be able to determine how satisfied a customer by a service. This will be done by observing the changes that the customer experiences as they interact with the sales personand how their emotions improve as they buy the product .

1.4 Softwares' Limitations

The software can only be implemented on one customer per time. The system will only be used in an environment with minimum people and should be in environment where there is no noise and no busy people moving around constantly.

Chapter 2

Requirements Analysis Document

2.1 Introduction

The Requirements analysis document focuses on the designers point of view. This document will look at how the system is expected to function, what tools will be required for the implementation of the system and what factors should be considered in order to ensure the system functions fully.

The designer will also be expected to analyze the user requirements document in order to determine if any data is required, data that will assist in improving the functionality of the system or data that is required in order for the system to be usable.

2.2 Designers intepretation of the users requirements

Given the above user requirements, the system created will have to determine emotions through the customers voice, When a customer makes a sale the voice will be analyzed to determine their emotion. This system will be able to accomodate the requirements the user had.

The user will have to be in an environment with minimum noise in order for the system to clearly hear their voice, A more detailed

aprroach of the project can be seen as the steps given below:

- A high quality micophone: This will be used to get high quality audio samples from the user to be used when determining their emotions
- A pre-existing database of emotion recognition audio: This will be used for comparing the customers sample audio to that similar in the database
- Python programming language: The system will be built using this language

These are the basics needed in order to implement the project. The system will be designed to work on computers.

2.3 Related Work

2.3.1 First Solution

A paper written by Te Won Lee, Jiucang Hao, Kwokleung Chan and Oh-Wook Kwon from the university of Carlifonia in San Diego, which also focused on Emotion Recognition by Speech Signals was implemented successfully with a high accuracy.

While designing the system, they noticed that the most important aspect that improved the results was the pitch and energy that the user had while using the microphone. They used a hybrid o Quadratic Discriminant analysis algorithm, Support vector machines and MFCC as the base feauture. Their chosen databases was the SUSAS database and the AIBO database. The SUSAS database is text-independent, while the AIBO database is speaker independent.

The SUSAS database when tested came back with an accuracy of 96.3 percent for stressed and nuetral emotions, whilst the AIBO database provided a 42.3 percent accuracy. (Kwon, Hao, Chan & Lee, 2003)

2.3.2 Second Solution

The Indian Institute of Technology Kharagpur wrote a paper on Emotion Recognition through speech. They used two different databases, one was acted emotion speech and the second one was real emotion speech samples. The project used the Guassian Mixture Model. The average classification accurracy obtained was 95 percent. (Rao et al., 2003)

2.3.3 Third Solution

The Hewlett-Packard laboratories wrote a paper on a system called Recognition of Emotions in Interactive voice system. The system was designed to determine when a customer gets annoyed or angry by the Interactive Voice Response system, it would transfer the customer to a real human call center agent.

HP used two databases, one was obtained from actors which imagined a scenario and acted it out. The second database used was obtained from the linguistic department in thr University of Pennsylvania.

They used Support Vector Machines and achieved an overall 77 percent accuracy. (Yacoub, Simske, Lin & Burns , 1990)

2.3.4 Best Solution

All the solutions provided excellent results, most of the excellence of the results mostly depended on which database they use and the method they chose to use. From all these solutions it is quite clear that in order to get great accuracy, one should use a hybrid for the classifiers and obtain a reliable database.

The best solution is the first solution. Their system was implemented the following way:

1. Feature Extraction: They selected the log energies, pitch and MFCCs as their base feautures, then selected their frame shift to be 10 miliseconds.

2. Feauture Selection: Identifies feautures and properties of speech that are important for distinguishing different emotions.

3. Classification: In this stage they used their chosen algorithms to process the audio to obtain the emotion outcome

(Kwon, Hao, Chan & Lee, 2003)

The reason for choosing this as the best solution is because it is more inline with the system the user requires and this system was tested for different emotions.

2.4 Testing

To determine whether the system is functioning properly, the following can be looked at

- Test i the system recognises audio when the customer speaks
- Test if it can differentiate between angry, sad or nuetral
- Test if it gives an output.

These are the basics tests that the system should pass to be seen as functional.

Chapter 3

System Design

3.1 Description of System Interface

The chosen interface for the system is a GUI. The GUI will be simply built to allow for easy input and for everyone be able to interact wit the system. After the classification of the audio has been completed, the GUI will then provide the user with the emotion detected and the probability that the emotion detected is the correct one.

The system is composed of the following buttons and options:

- Directory Input line: Input the directory with the file that you would like to classify.
- Get Audio Button: This gets the .wav file that is in the directory input line, then proccesses it and tells you the emotion.
- Record Button: This records audio through the microphone and the tells you the emotion detected after recording.
- Progress label: Informs you of what the system is currently doing. e.g Recording, Proccessing, Complete
- Emotion Label: Informs of the detected label
- LCD Number Display: Shows the percentage of the accuracy of the predicted emotion.
- Exit Button: This allows the user to exits.

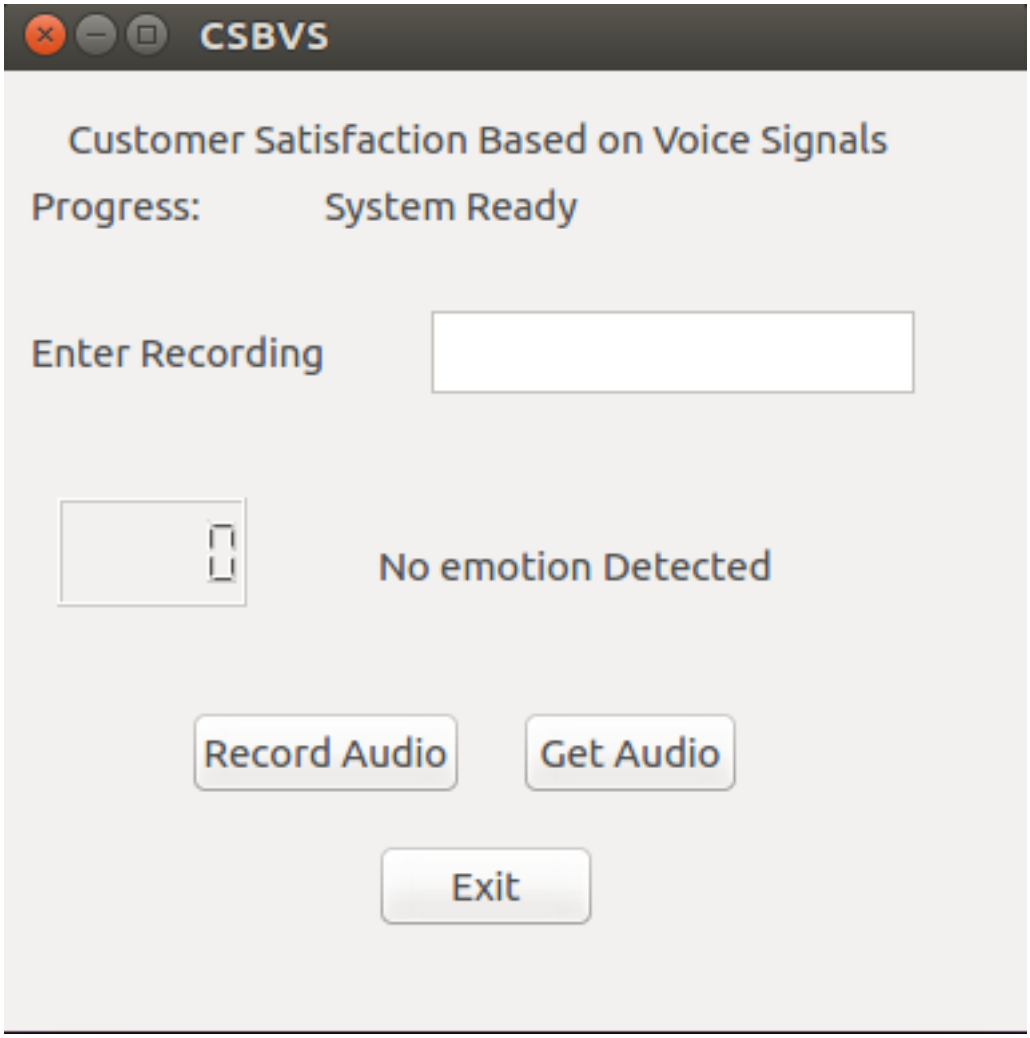

Figure 1: CSBVS User Interface

3.1.1 System Functionality

The system has a microphone that takes a short recording of the user, the audio file is saved in .wav format. If a pre-existing audio file had already been pre-recorded it can also be uploaded on the system. The system takes recordings of four seconds. The audio file also has to be of a mono channel.

The microphone has to be directly infront of a user when recording for best quality in sound to ensure a high accurracy when the audio is being classified.

The system will be able to distinguish between three emotion, which are anger/dissapointment, Happy/Satisfied and Nuetral

3.2 High Level Design

This section deals with how the architecture of the system is structured. A definition of how the each component will work and how the architecture will be structured.

The system is mainly designed to be able to analyze the voice of a user. This will be done through the use of a micropphone that will be connected to the computer. The computer recognizes the microphone connected. The recorded audio is then passed through sound processing algorithms to obtain usable data that the computer can use.

The information is returned as data that is a represantation of the audio file. This recorded audio is then passed on to be tested on the trained data. The trained data is a large sum of a audio files, composed of many different people acting out different emotions, specifically angry/sad, happy/excited and nuetral.

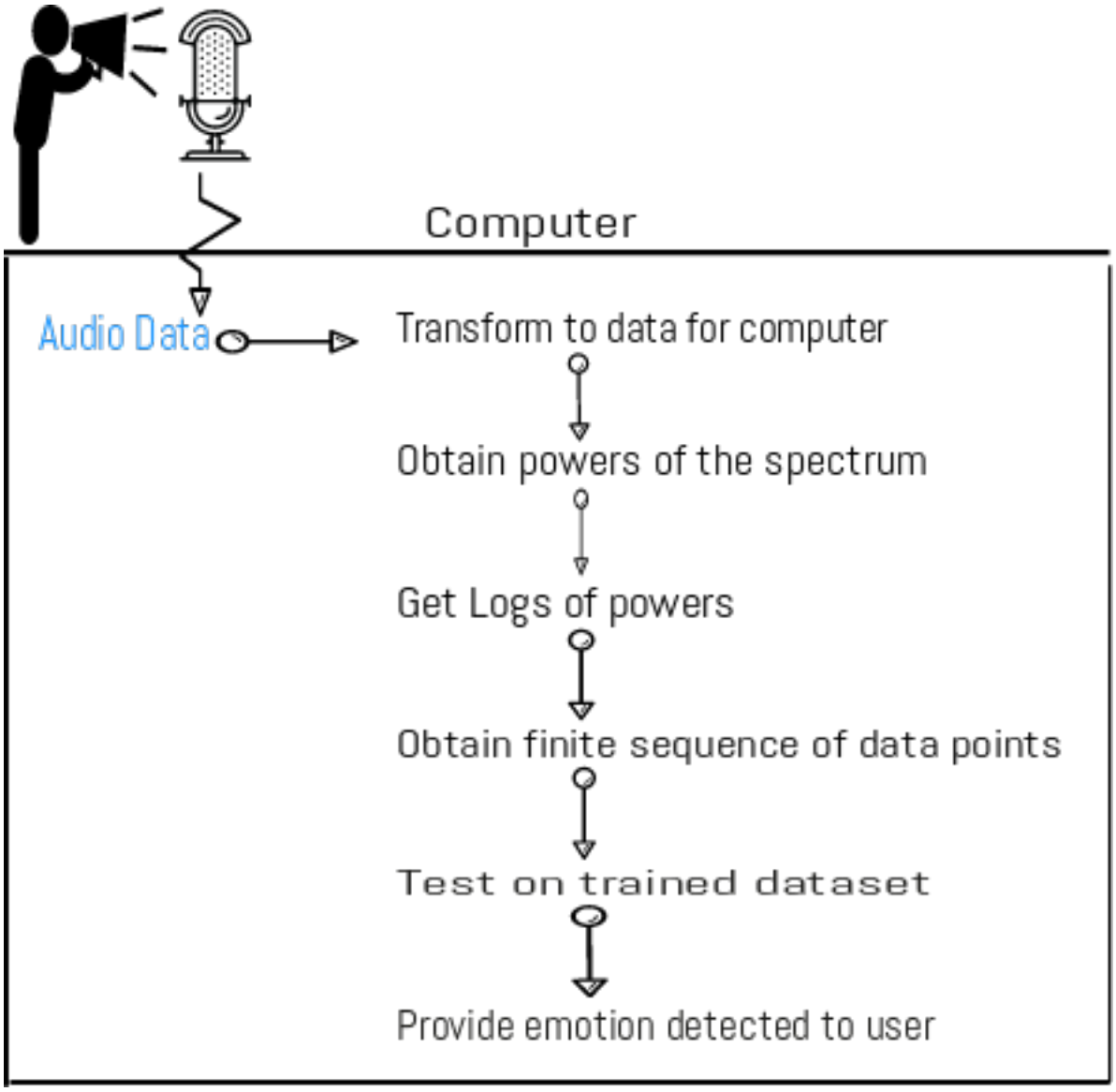

Figure 2: High Level Design

3.3 Low Level Design

This section focuses on the implementation of the system. The methods used and psuedo code for all functions that played a huge role in the implementation of the system.

In order for the system to accurately work, a large sum of audio files have to be trained, the audio files will be of the three emotions we are testing for. The audio will have to be turned to data that the computer can understand and do calculations upon. To do this, Mel Frequency Cepstrum Coefficients(MFCC) will be used. MFCC is a feature extraction method for speech recognition, the goal of MFCC is to mimic the human hearing perception.

3.4 Feature Extraction

3.4.1 Mel Frequency Cepstrum Coefficient

The implementation of the MFCC used in the system was composed of four functions, Fast Fourier transform, Mel Scale, Log and Discrete Cosine Function. The use of these functions gave the feature that were used to be trained and to be tested.

These functions work in a linear method, each function depends on data from the previous method to accuratelly get the feature, The feature data given by the MFCC is a three dimension vector. The three dimensional vector can be manipulated to suit any format that the chosen classifier to be used to be able to understand the data. The audio file being proccessed by MFCC must be noise free as far as possible, as any other unneccessary noise on the audio might affect the accuracy of the testing.

The steps to follow when implementing MFCC:

1. Frame the signal given into short frames.
2. Calculate the periodogram estimate of the power spectrum
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.

The process is shown below as a schematic diagram.

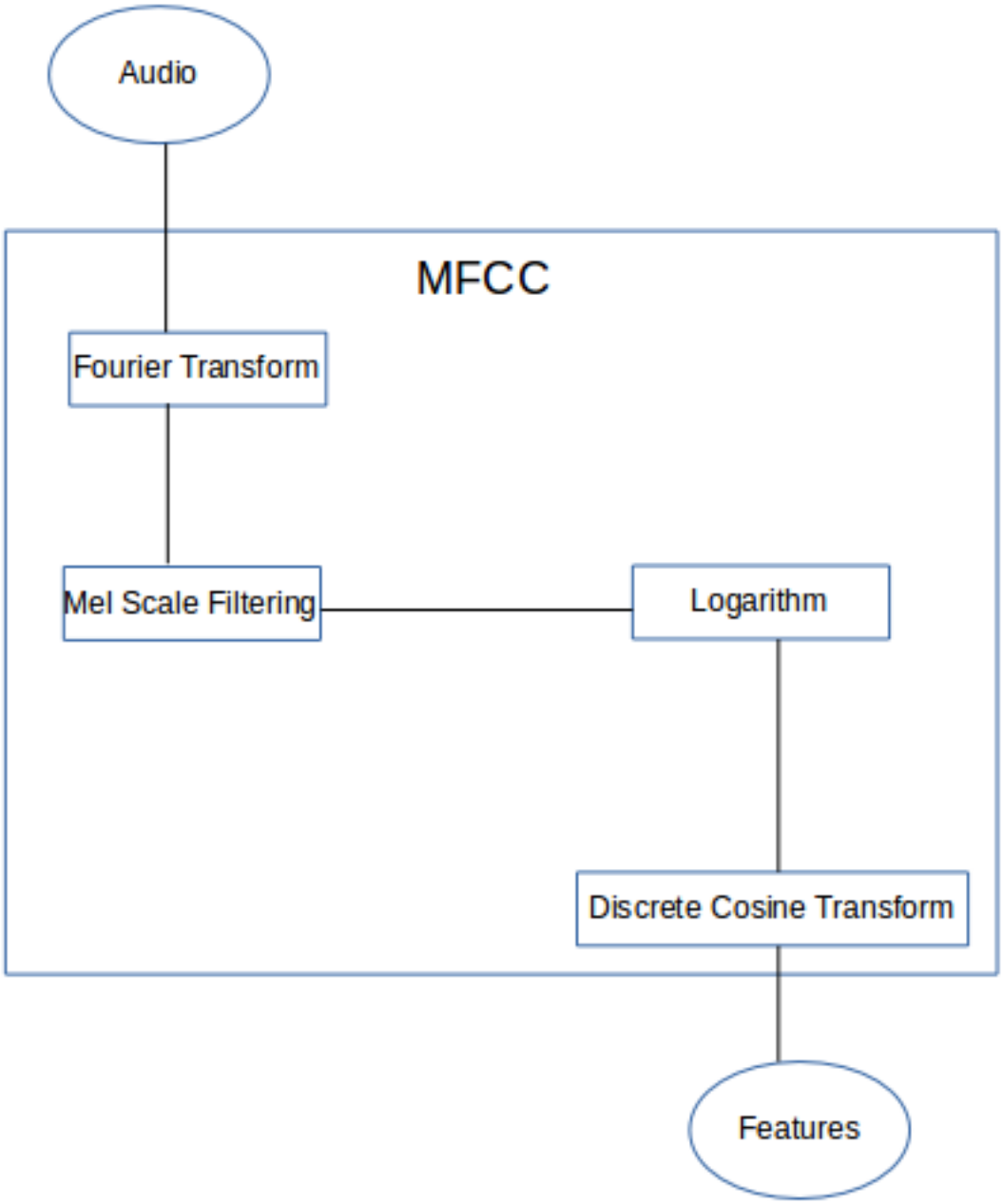

Figure 3: How Mel Frequency Cesptrum Coefficient works

Fast Fourier Transform(FFT)

The FFT is the first function that takes in the audio file and converts it to data that can be understood and proccessed by the computer. The data returned from this function is magnitude frequency response for each frame. The FFT algorithm does this by calculating the Discrete Fourrier Transform(DFT). The DFT calculation is done as follows

$$c_{t,x}^{l} = \left| \sum f_j \, exp\left[-i \, 2\pi \frac{jx}{N} \right] \right|, x \in (0, (N/2 - 1)) \tag{3.1}$$

FFT is not only limited to using DFT. DFT is one of the algorithms of FFT that require a large amount of computing power.

Mel Scale Filtering

This function is used to obtain the mel frequency spectrum. To do this, every element returned in the FFT function above will go through the algorithm for this function. The formula for doing this is:

$$m = 2595 \log_{10} \left(1 + \frac{c_{t,x}^{l}}{700} \right) \tag{3.2}$$

Logarithm

The next step is to get the log of the powers of all the frequencies. The aim of this is to mimic the human perception of loudness. The formula for doing this is:

$$c_{t,x}^{1} = \log(m) \, where \, x \in (1, N_m) \tag{3.3}$$

Discrete Cosine Function

After the logarithm has been taken. The next step is to remove the speaker depedency by obtaining the Cepstral Coefficient. The formula for doing this is given as:

$$c_{t,x}^{2} = c_{t,x}^{1} \left[\frac{k(2j-1)\pi}{2N_d} \right] where x \in (0, N_c < N_d) \tag{3.4}$$

All the above functions worked in finding the final data that was needed for classification.

3.5 Feature Classification

3.5.1 Support Vector Machines

Inorder for the system to work properly, the system will use SVM to train the data. Training data is a proccess of obtaining pre-existing data, which would be an audio set for the system. All the different emotions will be trained and labeled as a certain value during training. Happiness audio files could be labeled as 1, sadness 0 and nuetral as 2.

but before any audio file is passed through the SVM, it has to be in understandable format for the SVM. Inorder to do so every audio file that comes to be classified has to pass the feature extraction phase.

Once this proccess is done the system will be able to test any file that satisfies the requirements required by system for testing.

References

[1] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee, Emotion Recognition by Speech Signals ,Institute for Neural Computation University of California, San Diego, USA

[2] Sherif Yacoub, Steve Simske, Xiaofan Lin, John Burns , Recognition of Emotions in Interactive Voice Response Systems , HP Laboratories Palo Alto

[3] K. Sreenivasa Rao, Tummala Pavan Kumar, Kusam Anusha, Bathina Leela, Ingilela Bhavana and Singavarapu V.S.K. Gowtham, Emotion Recognition from Speech, School of Information Technology, Indian Institute of Technology Kharagpur

[4] Todor Ganchev, Nikos Fakotakis, George Kokkinakis, Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, University of Patras

[5] Christopher J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition