FYP Report

Topic: Emotion Detection from Voice

By

Name: Dahal Prakash,

Student Number: 1828421

Location: Canvas

Module Name: Project and Professionalism

Email: p.dahal2@wlv.ac.uk

Supervisor,

Mr. Krishna Aryal

Reader,

Mr. Rupak Koirala

Date of submission: 6th May 2019

Award Title:

Declaration Sheet

Presented in partial fulfilment of the assessment requirements for the above award.

This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work. I confirm that the intellectual contents of the work is the result of my own efforts and of no other person.

It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the University a perpetual royalty-free licence to do all or any of those things referred to in section 16(i) of the Copyright Designs and Patents Act 1988. (viz: to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make an adaptation of the work).

Student Name: Prakash Dahal

Student Number: 1828421

Signature: Prakash Dahal Date: 6 May 2019

(Must include the unedited statement above. Sign and date)

Please use an electronic signature (scan and insert)

Table of Contents

Abstract

Communication with machine is common. These days, human speak and communicate with their phone, tablets and computer. While speaking, the voice will be carrying some messages like; noise, environment, emotions and so on. Emotion detection has been a field of research and development for researchers. This report gives the detail information used to complete this project including its aims and objectives, methodology used, language, library, software, etc. The project focus on the emotion of the speaker based on their voice. It detects the emotion of the speaker extracting important features from it like first 13 MFCC and chroma as energy. This system has been developed using python and its library. For the part of training, dataset is used by from SAVEE database which will be trained using K-nearest neural network (KNN), Support Vector Machine (SVM), Artificial Neural Network using sk-learn, tensor-flow and keras.

Acknowledgement

I would like to acknowledge for my Supervisor Mr. Krishna Aryal and reader Mr. Rupak Koirala for suggesting and guiding me throughout the project. Their guidance helped a lot to face and overcome problems. They are the key helpful person to guide me.

I am thankful towards Mr. Mahendra Thapa Sir. He guided me to handle errors and provided base of artificial intelligence which helped me to make further research and understand fast.

I would like to remember Mr. Prakash Gautam Sir. He supported me in required mathematics and AI knowledge. He also supported me in report formatting.

I am also thankful towards my friends to motivate and support me.

Without them, this project would not have been completed. I am happy and thankful for their support, guidance and motivations.

List of Figures

List of Tables

1. Introduction

1.1. General Introduction:

Sound is a mechanical energy wave which is generated by the vibration (i.e. back and forth movement of an object). Sound creates disturbances in the air molecules and the air molecules stars moving back and forth. The back and forth process is called as Oscillation. If guitar string is pulled, then it starts vibrating from its initial position and reaches to the maximum point, goes beyond the initial position and get back to the same position. This vibration decreases slowly.

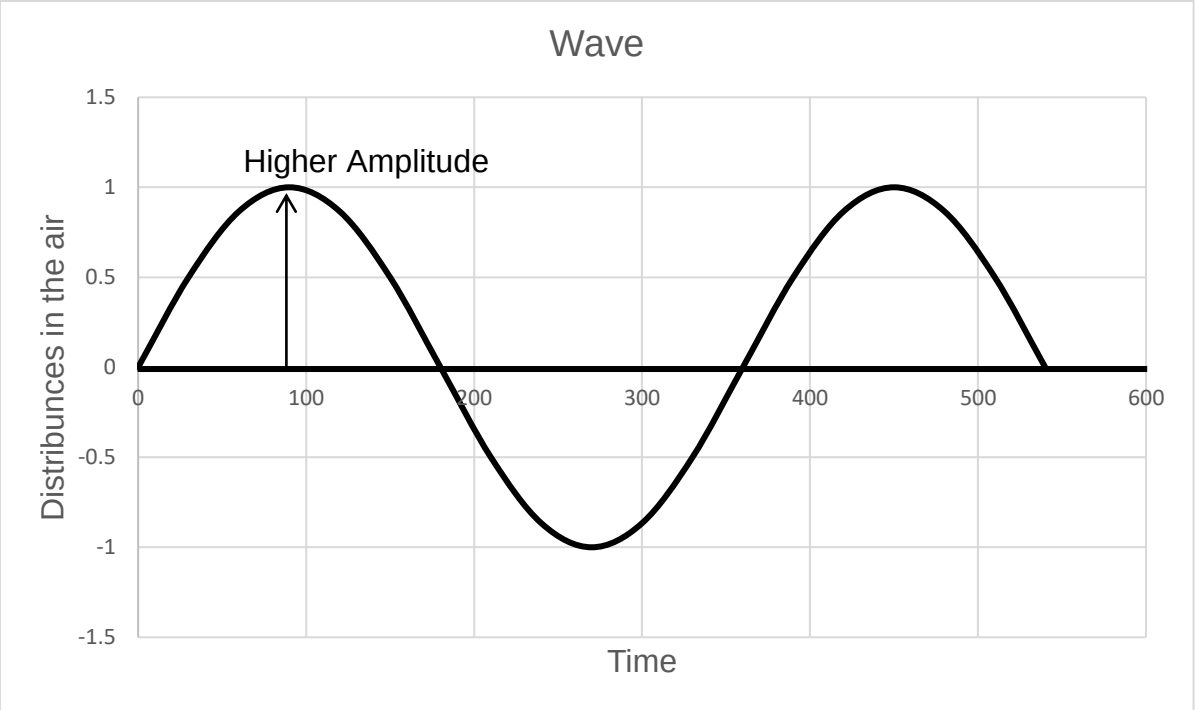


Figure 1: Representation of Wave frequency and it's properties

Here black bold x-axis is an equilibrium point. Maximum displacement from the equilibrium position is called as amplitude. The time taken by an air molecule to get back to its original position is called as a cycle. This cycle represents for the period (T). Period is the time taken by an air molecule to complete one cycle or oscillation. Frequency us measured as one over the period.

$$Frequency\ (f) = \frac{1}{T}\ Hz$$

Where T is a period, and Hz is a unit called as Hertz.

Human can only hear frequency from 20 Hz to 20,000 Hz. Sound below 20 Hz are called infrasonic and above 20 KHz is called ultrasonic (SAMUEL J. LING, 2016).

Human are social beings, so they represent their feelings and emotions in different situations and environments. Facial expressions, speech, body language and text are some key factors for emotion representation. This paper is about the emotion detection from speech factor, therefore standing right to the topic.

Speech of human is associated with different aspects like; vocal track, duration, his/her style, language, culture, community and so on. These aspects are different, so the speech of the speaker is different. But still the speech of human carries different messages which can be processed and analysed. Speech contains different features like; energy, pitch, timing, voice quality and so on which are being used to recognize speech and the speaker. These features also carry the emotional factors like tension, dissatisfaction, disagreement, pain of the speaker which can be used to detect the emotion of the speaker (Sreenivasa Rao Krothapalli, 2013). There is a primary or basic six emotions. They are; Anger, happiness, sadness, surprise, dislike and fear.

Different features can be extracted from the speech. Speech features can be categorized into four parts. Spectral-based features like; Linear Predictor Coefficients (LPC), Mel-frequency Cepstral Coefficients (MFCC), Log Frequency Power Coefficients (LFPC), Continuous speech features like; Pitch, Energy, Formants, Quality-based features like; voice quality, harsh, tense, breathy and Teager Energy Operator (TEO) based features like; TEO-FM-var, TEO-Auto-Env, TEO-CB-Auto-Env can be extracted from the given speech.

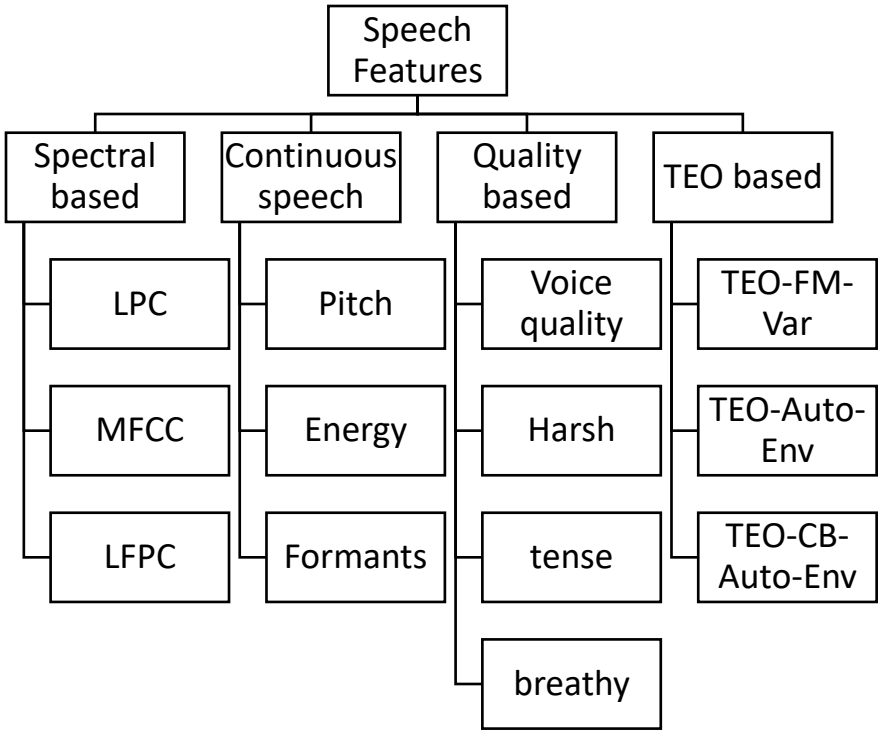


Figure 2: Chart of Speech features Categories (Moataz El Ayadi, 2011).

Spectral-based speech features:

Mostly Spectral features are selected as it represents for the short speech signal which also includes the acoustic features like pitch and energy. Spectral energy distribution has more impact on emotion with speech frequency range. For example; Happiness emotion has high energy at high frequency whereas sadness emotion has small energy even at the high frequency. This is one of the reasons for the researchers to user spectral features for emotion classification. Spectral frequency is extracted in a linear-based way like; ordinary linear predictor coefficient (LPC), one-sided autocorrelation linear predictor coefficients (OSALPC) and so on then, Cepstral-based features was used which is comparatively better than linear-based for emotion recognition. Cepstral based features are MFCC and LPCC.

Cepstral coefficients are derived from linear prediction (LP) or a filter bank approach which is regarded as standard front-end features. Speech system developed on these features have high accuracy, if the speech is given from the clean

environment. Mel Frequency Cepstral Coefficients (MFCC) is considered as one of the standard methods for feature extraction. First 10-12 MFCC are extracted and used in machine learning either for speech recognition or emotion detection. Speech signal is non-linear which can be processed through nonlinear Mel-scale filter bank. So, twenty filter banks are used to compute 8, 13 and 21 MFCC features from a speech frame of 20 ms. Linear Prediction Cepstral Coefficients (LPCCs) holds the emotion related information through the vocal tract features. The main purpose of using LPCC is to consider vocal tract of the speaker while recognizing the emotions (Sreenivasa Rao Krothapalli, 2013).

Continuous speech features:

Continuous speech features include energy and pitch which is preferred by most of the researchers for emotion. Duration and formants are other key points included in it.

Quality-based features:

Most of the researchers believe that the quality of voice is an essential feature for the emotion detection. Quality of voice has influence on positive and negative actions and thought of a person.

TEO based features:

The study of Teager reveals that the speech produced is non-linear and therefore non-linear can give more accurate result than linear. TEO based features are used for the stress detection on speech (Moataz El Ayadi, 2011).

1.2. Project Overview

Emotion recognition by machine has been an interesting research for the researchers and developers. It is difficult task to detect human emotion and it is more challenging if it is only based in the speech factor. Emotion detection by machine is different than that of human. Machine processes information in a numeric way because so far programming is not smart enough to work on frequencies directly and on the other hand, it does not have sensation like of human. Therefore, the extracted features of speech must be converted into numeric format so that the programming can compute it well. In this report, classification of emotion is done using different learning algorithm and then the obtained output is compared to identify which learning algorithm is suitable for this type of problem. SAVEE database is used which is in WAV format. Sad, Happy, Angry and Fear are the four emotions used for the classification. First Thirteen Mel Frequency Cepstral Coefficient (MFCC) features and chroma energy are extracted from all required classes and it is used for emotion classification. K-nearest Neighbour (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used as learning algorithm and these results are compared. KNN is an unsupervised learning algorithm whereas SVM and ANN are supervised learning.

1.3. Problem Domain:

Emotion is a regular expression provided by humans in different situation which gives us some information. These information's can be used in different fields for investigation. Rape cases are increasing in Nepal where if mobile phone of victim can detect the emotional state and take some actions then that can help the victim. Similarly, criminal activities are increasing which can be controlled if the mental state is analysed. Lie investigation can be done which may Call centres can use emotion detection program to know the user emotions and act accordingly. It can be integrated with chatbots which can handle the depressed human. Depression handling can be a best application by detecting the emotion of the speaker.

These are some of the areas which is demanding to know the emotional state of a person. These problems can be solved through emotion detection. Therefore, the development of the emotion detection is necessary.

1.4. Academic Question:

The two major academic questions are given below:

i. How to detect emotion of a person from his/her voice?
ii. Which algorithm is appropriate for the classification of speech emotion?

1.5. Aims and Objectives:

This project aims to analyse the human emotion from voice which is very important for any communication between human and computer.

Some objectives to fulfil these above aims are:

a. Understand the use of speech parameters mostly applicable for speech emotion.
b. Feature extraction from the given audio.
c. Use algorithms to classify emotions.
d. Find the drawbacks of the algorithms for emotion classification by comparing

1.6. Project as a Solution:

Emotion detection system is required in different sectors. To solve these problems this system has been proposed. In this system, user gives the voice as an input and the result is given from the trained and saved model. This project provides 76% of accuracy using KNN algorithm, SVM provides 68.75% of accuracy in 13 MFCC mean and standard deviation data. Accuracy of ANN with 500 epoch and learning rate as 0.00001, hidden layer 1024,521,128,4 and optimizer as gradient descent gives accuracy of 69.44%. This accuracy can be changed by changing the hyper-parameters. After getting better accuracy, this system can be used in different areas which can solve the problems of emotion detection.

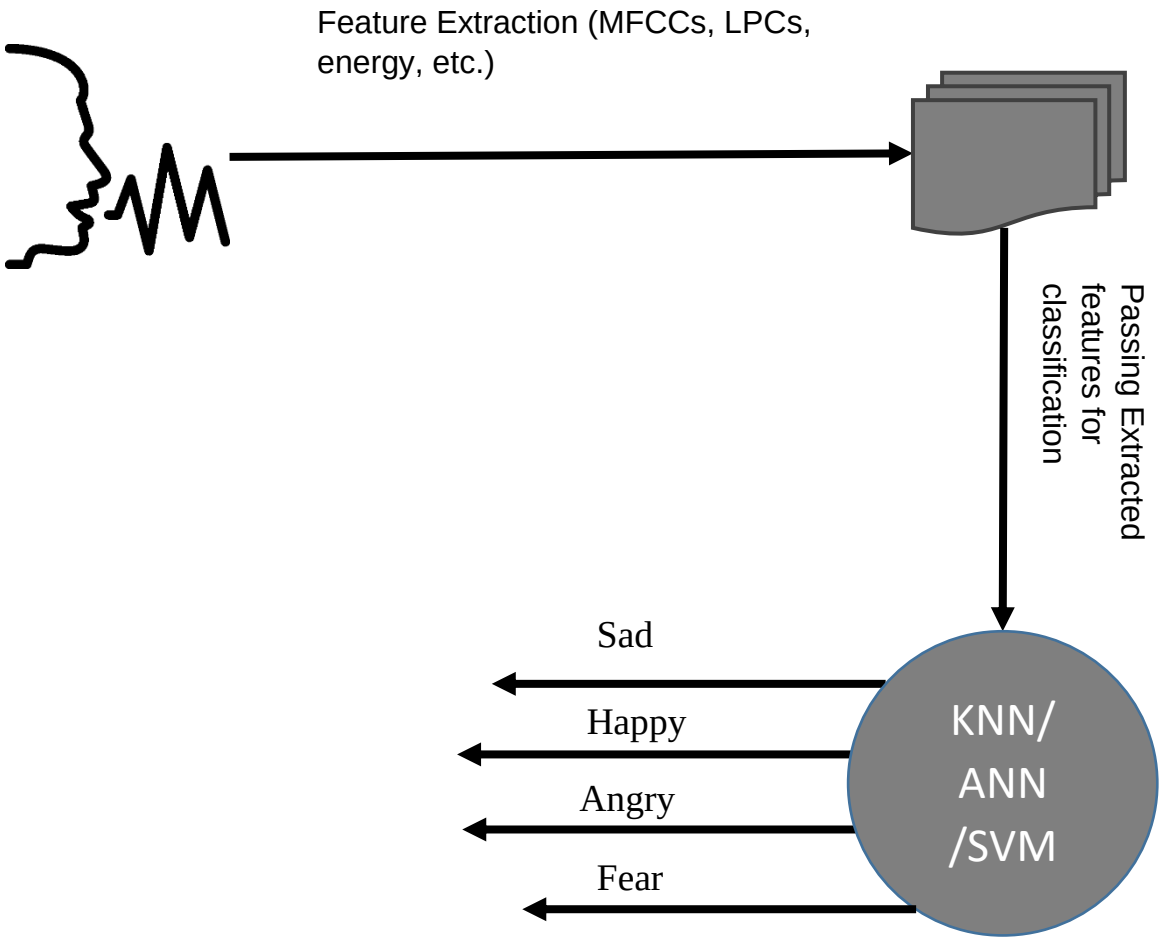


Figure 3: System Workflow

1.7. Report overview

The report is the documentation of the product developed. To build the product research was carried out. There are different research methods like application prospective, objective and enquiry mode. In this report qualitative research and pure and applied research are used which are under enquiry mode and application respectively because this report involves the study of different books, journals, conference and the application of the knowledge gathered (Kumar, 2011) (MASON, 2002).

The project is limited up to emotion extraction. This project is aimed to take user voice save it in a file (say output.wav) and display the emotion of the user. Model is developed through Anaconda software and saved model is used in Django Frame-work to take user voice and display emotion.

The report structure is represented below:

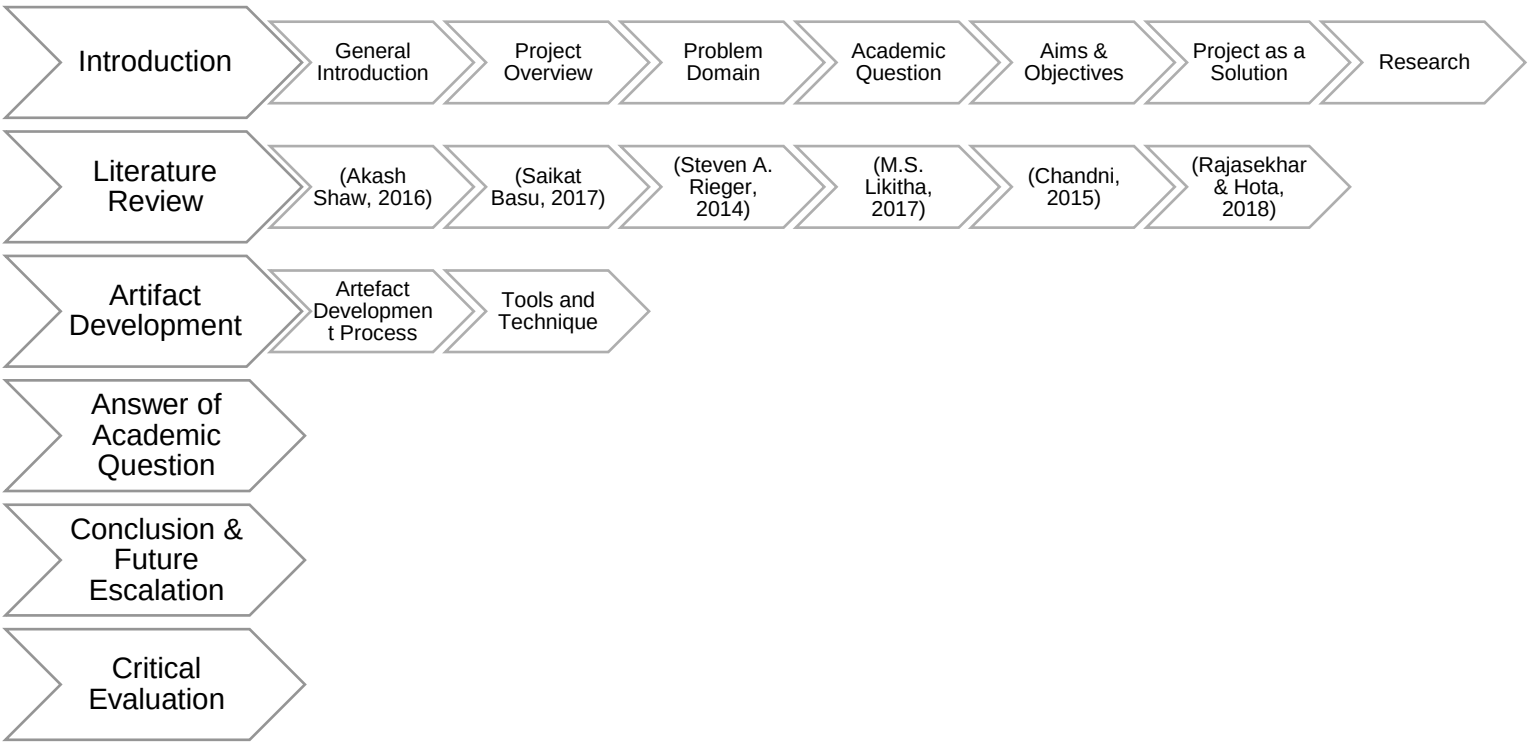


Figure 4: Report Structure

The above chart shows the structure of the report. In introduction, general introduction with research, problem and solution of the project are discussed. Literature review reflects similar projects and their working mechanism and different algorithm used. In Artefact development total development processes and the reason of using such tools and technique is presented. Answer of Academic questions are presented. In conclusion and future escalation report and project is concluded and further improvement are disclosed, and finally critical evaluation is presented.

2. Literature Review

2.1. Emotion Recognition and Classification in Speech using ANN

The interaction of human and computers are common, and it is a popular area of research. People these days give command to their car, computer and many other electric devices. Therefore, machine understanding the emotion of the speaker is impactful for the better interaction, but it is a very interesting as well as challenging task. Better interaction is possible only if the recognition is better. For the effective recognition of the emotion, proper training has to be done for which proper features are required. Author has used spectral features like MFCC and LPC along with formant, pitch, energy and so on. Author wants to classify four different emotions; neutral, happy, angry and sad using Artificial Neural Network.

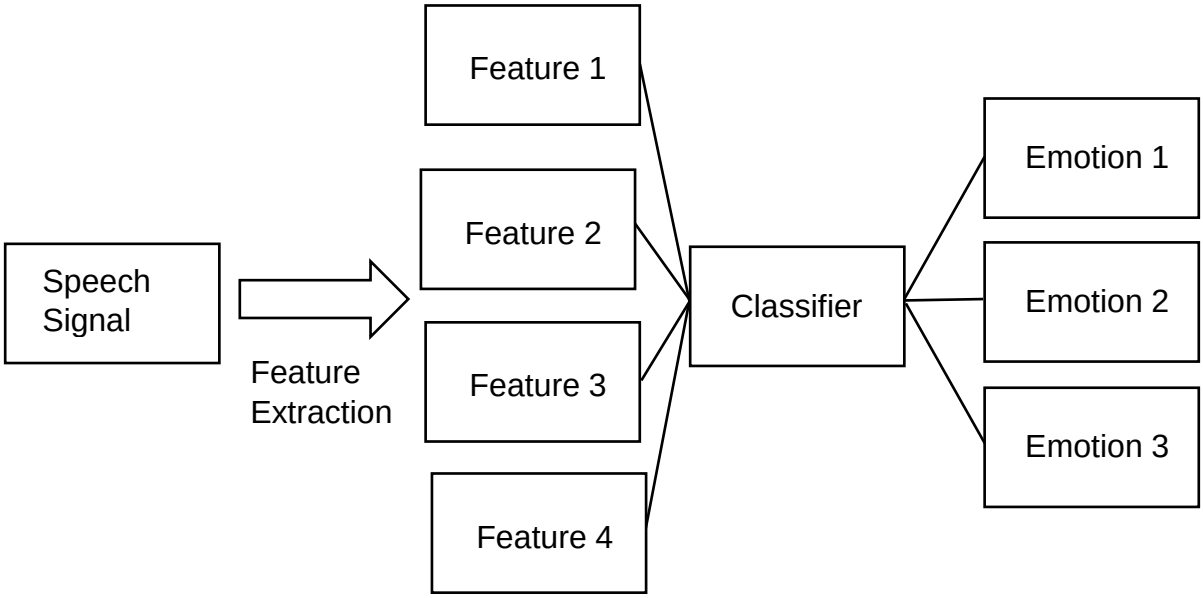


Figure 5: Block Diagram of input flow, process and output

The diagram represents the overall working mechanism for the classification of emotions. Here speech is received from the user and its features are extracted then it is passed through classifier layer which classifies the emotion from the given speech features and the actual output is obtained. But the audio file has to be pre-processed so that the required features are extracted.

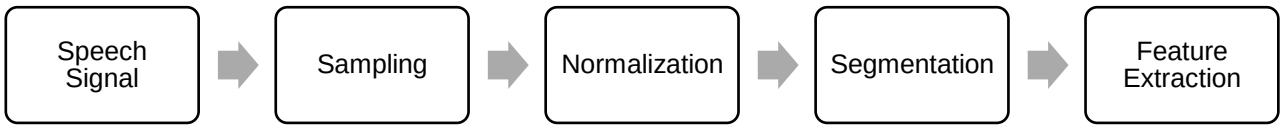


Figure 6: Pre-processing for emotion recognition

Speech signals is in Analog format which needs to be converted into digital format. Sampling converts Analog signal into discrete time signal. Normalization ensure for the volume level of each sentence is comparable because volume is an important factor to calculate speech energy as well as other features. Then segmentation is done since whole signal has to be divided into different frame. Overlapping is used to avoid loss.

Important features used by Author are:

2.1.1. Energy

Energy is important for emotion recognition because most of the emotions are based on energy. It is regarded as the basic feature in speech signal processing. Happiness and anger are represented by the energy since both have higher energy while comparing with sadness or neutral.

2.1.2. Pitch

Tone of the speech rises and falls which is regarded as pitch. Vibration frequency of vocal fold is represented by the pitch while the user is speaking. There are many methods to work with the pitch. Here Author uses Auto-correlation method which is easier to make short term analysis.

2.1.3. Formant Frequencies

Formant represents the timber of vowel from the speech. This feature is used in different papers because of its importance. Formants are the peaks of the frequency response from a linear prediction filter.

2.1.4. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is a spectrum feature. It is an accurate representation of the short time power spectrum. MFCC is capable

to imitate the reaction of human ear to sounds. It does not use linear spaced frequency, but Mel scale is used. The extraction of MFCC has to be passed from different stages.

After extracting all these features, Author used Artificial Neural Network to train these features.

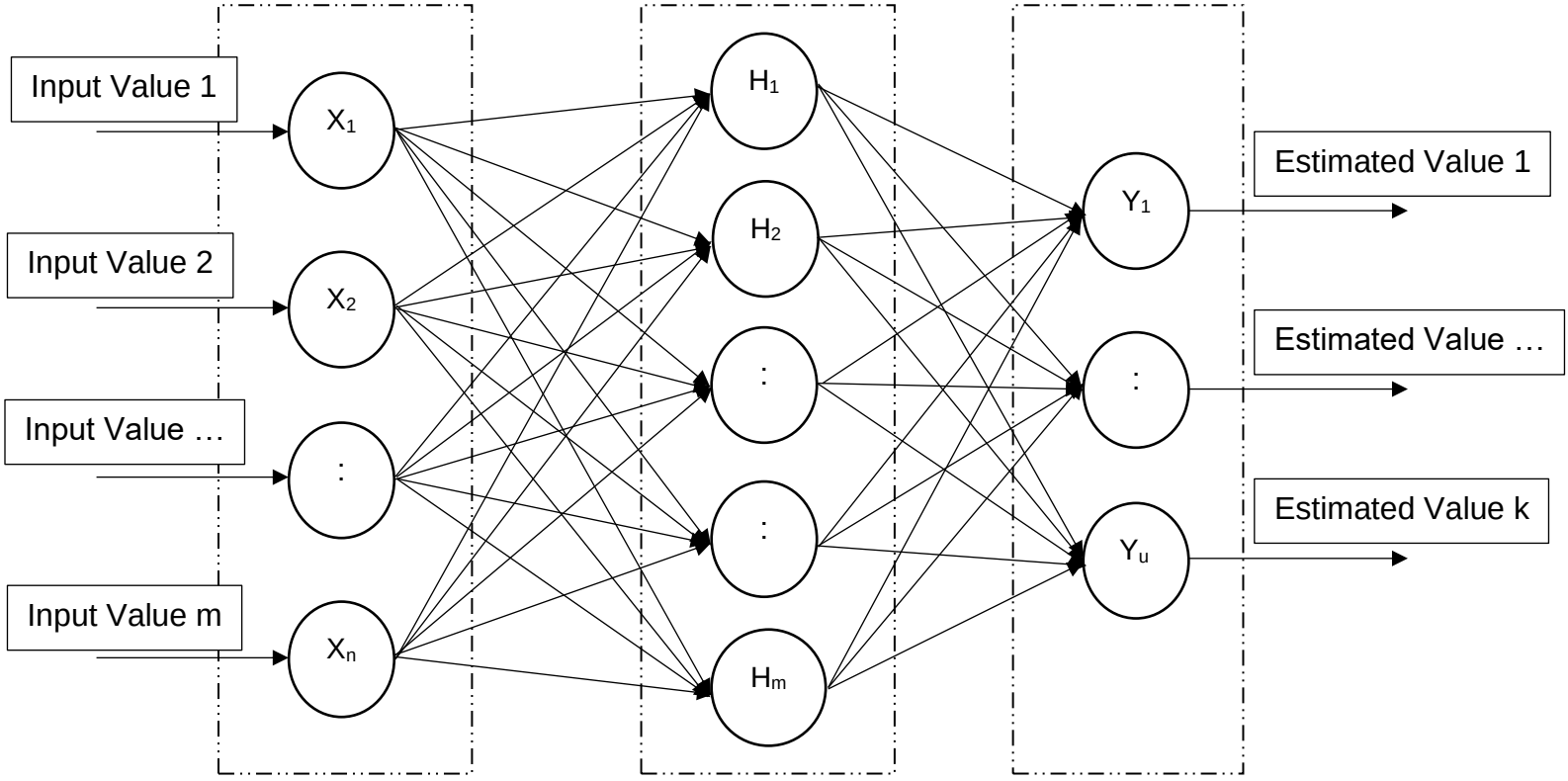


Figure 7: Artificial Neural Network

The general workflow of Artificial Neural Network (ANN) is represented by Figure 5. In this paper, Author has used ANN to classify the emotion. Matrix of input dataset has categorised into three parts; training, validation and testing. Author trained dataset several times to minimize the error. The obtained mean square error rate shows how good the model has learned (Akash Shaw, 2016).

Analysis:

In this paper author has used different features which are important for emotion recognition and four different emotions are classified using artificial neural network. Different other features like quality-based features can be included which can increase the output of the result as Author explains in the future work section that more emotions can be classified improving the features.

2.2. Emotion Recognition from Speech using CNN with RNN Architecture

Interaction best way is communication. Human speak with machine in different way these days. It is challenging to detect emotion of the speaker by the machine. Robotics engineering, medical science, call centre are some major area where emotion of speaker is essential. Therefore, it is needed to build human like system which can detect emotions effectively and efficiently. The universal emotions include happiness, sadness, surprise, neutral, disgust, fearful, stressed, etc. Author has selected; fear, disgust, happiness, boredom, neutral, sadness and anger emotions to classify using Convolution Neural Network (CNN) and then feeding the obtained result in Neural Network where Author has selected Recurrent Neural Network (RNN). Different features are required for this classification where Author highly focused on spectral features like; MFCC, LPCC. Xianglin Cheng et al. has performed emotion classification based on Gausian Mixture Model (GMM) using MFCC and pitch and the recognition rate was 81%. But the Author is using Convolutional Neural Network over Recurrent Neural Network. Berlin Database of Emotional Speech (EmoDB) has been used which contains 535 utterances spoken by 10 different actors. This database has seven classes of emotions; happy, angry, anxious, fearful, bored, disgusted and neutral.

For feature extraction, pre-processing is done before extracting MFCC feature. Below illustrated diagram explains the steps for MFCC extraction.

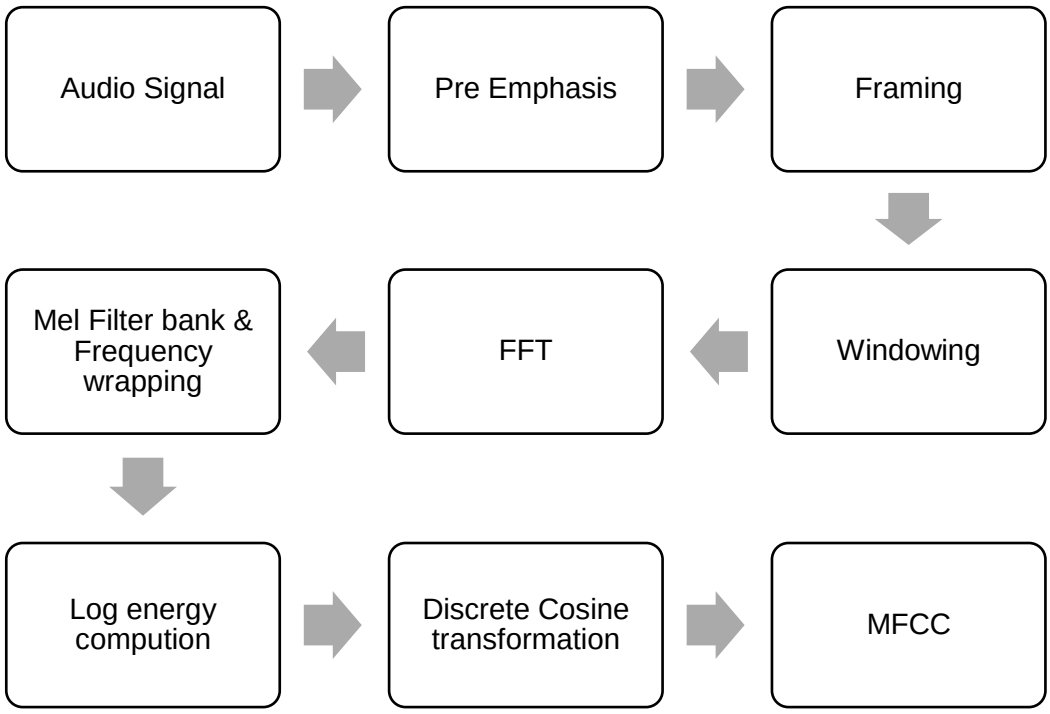


Figure 8: Block Diagram of MFCC

Signal energy is increased using pre-emphasis. Framing segments, the speech sample into 20-40 ms frames. Signal discontinuity has to be reduced which is done by windowing. Fast Fourier Transform (FFT) generates frequency spectrum for each frame. Mel Scale filter gives the energy existing in each frame. Log function is applied on each frames of filter bank energy. Then finally, Discrete Cosine transformation extracts the 13 MFCC features, 13 acceleration and 13 velocity as required.

After extraction of features, classifiers are used. At first CNN has been used and obtained result is passes through LSTM. Convolutional Neural Network (CNN) has several convolution layers. Input features are passed through convolution layers and the input is convoluted by the different filters passed through rectified linear unit (ReLU) layer and feature map is generated. Single convolution layer gives a feature map. Then pooling layer is used which reduces the size of input. Max pooling is used by author which takes the maximum values and stores in a feature map. When input features are ready from CNN then it can be passed through

Multilayer Perceptron (MLP) or Recurrent Neural Network. The output from CNN is an input feature for RNN or MLP. RNN alone cannot give good result since it has a drawback of vanishing gradient which can be solved using Long Short-Term Memory (LSTM). LSTM is a memory cell which can remember or forget the information. Using LSTM and RNN together can work efficiently.

Author has distributed dataset into training and testing into 80% and 20% respectively. The model was trained with 500 epochs. The accuracy obtained in training dataset is 96% whereas accuracy in testing dataset is 80% (Saikat Basu, 2017).

Analysis

The result obtained from the CNN and RNN with LSTM. But Author convey that the result can be more fruitful if the dataset is increased and Bidirectional LSTM is used. The algorithm used is effective since it works on time frame which can give more accurate result in automatic speech recognition but the drawback of RNN has to be reduced as much as possible. In CNN, only max pooling is used instead average pooling can be used to compare the result if the result is effective. On the other hand, input data set are not equally distributed which causes class unbalance issue and can affect on the recognition which can be maintained making all the dataset equal on its respective classes.

2.3. Speech Emotion Recognition Using Ensemble of KNN classifiers

Security is a major thing in this developing technology. Coming generation can face threats on security issue in both way; internal and external. Areas like healthcare, transportation and logistics, smart environment, etc. demands for the emotion detection of user. Security can be enhanced by the emotion recognition system which reveals the state of mind of a person and human biometric system can reveal the identity. Secure website can be controlled by the emotional state of the person. If anger or panic is presented, then system can refuse to serve that person with the identification through biometrics. Using

password and user id, personal account can be hacked. Cyber security needs support beyond what it is now. Author has used k nearest neighbour (kNN) classifier for the emotion classification. This report is more based on the pattern recognition with the spectral features extracted from the speech. Acoustic features like energy, pitch, zero cross rate, etc. are also required features but in this paper only Linear Predictive Cepstral (CEP) and MFCC features haven been used for the emotion recognition. Here single kNN and an ensemble system with multiple kNNs are applied.

Author used kNN with the squared Euclidean distance measure because it is simple to use, and data trained on different subset using multiple kNN are also presented. LDC emotional prosody speech database is used by for his all experiment. There are three male and four female speakers in the LDC database. Author has selected six emotions; disgust, happy, anger, neutral, panic and sad. The classes are unequally distributed; 113 disgust, 143 happy, 72 neutral and so on.

Non-recursive filter has been used in the speech pre-emphasis stage. The frame duration used is 30 ms and the overlap is 20 ms. LP analysis is done with the autocorrelation which gave the result as LP polynomial A(z) of order p = 12. Also 12 dimensional MFCC feature vector is computed in each frame. Alpha is used as 1 where Beta is used as 0.9. On each frame kNN classification is used to identify emotion present there.

Ensemble of kNN was used by taking 2,3,4 and 5 kNNs. Five trials were taken with full training and performance evaluations steps. Datasets are divided into train and test set in 80% and 20%. Synthetic Minority Over-sampling Technique (SMOTE) is used to make balanced class. SMOTE is the overall result of the features for each class. Training is done using bagging algorithm and then performance is evaluated on the result calculated on each frame. Since each frame is evaluated therefore different emotions could have been processed on each frame therefore the maximum emotion is taken (Steven A. Rieger, 2014).

Analysis:

K nearest neighbour is used for the emotion classification with the Euclidean distance technique. Ensemble kNN technique used is better technique than just using a single kNN algorithm. Here in this paper, Euclidean distance has been used but there are other distances like Minkowsky which result can be better. On the other side, only limited features are used which can be enhanced by increasing the number of features because Author is aiming to provide better security than that we are having now. So, good result is necessary.

2.4. Human Speech Emotion Recognition Using MFCC

Emotion recognition is an attractive field. Processing speech signal and extracting emotion from it is an interesting task. Emotion is essential on different fields like fraud detection, surprise/amusement, healthcare, criminal investigation, intelligent assistance and so on. Speech emotion recognition is the state of identifying the emotional state of the speaker. To identify the emotional state through speech, different speech features are required. Different features like speech, pitch, formant is used by researchers and for better prediction Artificial Neural Network (ANN), Linear prediction cepstral coefficients (LPCC), Mel Frequency cepstral coefficient (MFCC) or combination of LPC and MCC as LPCMCC, Support Vector Machine (SVM) or combination of SVM and HMM can be used. Here Author has proposed this paper just using MFCC and decision making using standard deviation.

Feature is extracted using MFCC. Here MFCC is the key feature for the emotion detection. Extracted features have to be used for training and testing before that, pre-processing of speech is necessary:

Speech has to be framed and divided because it has to be processed in short time intervals called as frames. Generally, frame is sized between 20-40 ms. After this overlapping is done for the smooth transition between frames. The first frame value starts with N = 256 (typical value) and second value starts with M = 100. The overlapping is N-M and it goes through the same process till the speech is finished. Then Windowing function is used to minimize the

discontinuity and spectral distortion. The formula of windowing signal x(n) is given below:

$Y(n) = x(n)w(n), \quad 0 \leq n \leq N - 1$

Hamming windowing is used by an Author. Then Fast Fourier Transform (FFT) is used in windowing signal to convert it into frequency domain. Result after FFT is generally spectrum features. Mel Frequency is extracted from it and then MFCC is extracted.

Mean of resulted MFCC is extracted to reduce huge set of values from MFCC. The formula is given below:

$$x = \frac{x1+x2+x3+\cdots+xn}{n}$$

Then standard deviation (σ) is extracted. From this standard deviation, Author wants to make a decision boundary by evaluating the result. The result obtained are as follows:

Sad: 0.1700 – 0.2199, Happy: 0.2200 – 0.2899, Angry: 0.2900 – 0.3999.

For new speech, it's MFCC features is extracted passing through the above-mentioned steps and its mean and standard deviation is calculated. Then it is passed through if-else statement to get the result. Then if the if-else statement is satisfied then, it gives the emotion result. The accuracy obtained is 80% (M.S. Likitha, 2017).

Analysis

MFCC can be regarded as the powerful feature for the emotion recognition because here Author just using MFCC features, emotion has been classified. Mean and standard deviation is extracted for the Here just three different classes have been used for classification; sad, happy and angry. For three classes of emotion standard deviation has been given better result but when the classes increase for examples surprise, cold anger, disgust etc are added then the feature standard deviation of MFCC only would not be suitable. Only

if else statement has been used by Author, if neural network would have been used then the result could be different because the learning would have taken place while fitting data into neural network. It would not have been limited only with the condition defined inside if statement, but it would have been based on the learning from data. This can be further enhanced by training dataset in neural network.

2.5. An Automatic Emotion Recognizer using MFCCs and HMM

Beside facial expression, speech signal can be used for the emotion detection. In speech, only words are not converted but it also flows the wealth of the speech. Many researchers have used common speech features like pitch, intensity, spectral density, formants, etc. for the emotion classification but still the classification efficiency is low. In this paper, (Chandni, 2015) is providing a method to attach an emotional label using Hidden Markov Model Toolkit and Mel Frequency Cepstral for a continuous speech. Silence part in the speech was neglected and the accuracy was enhanced. The main reason of this better result is the optimization of the acoustic parameters, states of the HMM and the transitional probability between the states. Author reveals the importance of MFCC features rather than conventional prosodic features. The algorithm is neither complex but gives better efficiency. Here dataset is used from SAVEE emotional corpus and the obtained trial average accuracy is 78% and the highest accuracy is 91.25%. Even in future the use of this algorithm can give better result that the existing algorithm.

Automatic emotion recognition from the given speech is a very challenging task. Acoustic features are not clear for the emotion recognition system. Author suggests using MFCC for acoustic analysis with HMM as a classifier can provide efficient result. It worked on phonetic features from speech especially on MFCCs to improve accuracy with less feature set. The step wise methodology for emotion extraction using MFCC and HMM technique is; labelling emotional database splitting train and test data in the ratio of 2:1 using

Wavesurfer for four emotions, Acoustic features are transformed into a sequence of coefficient vectors, HMM model is defined, Task grammar and Task Dictionary is defined, Data testing is done with test dataset and finally performance is evaluated.

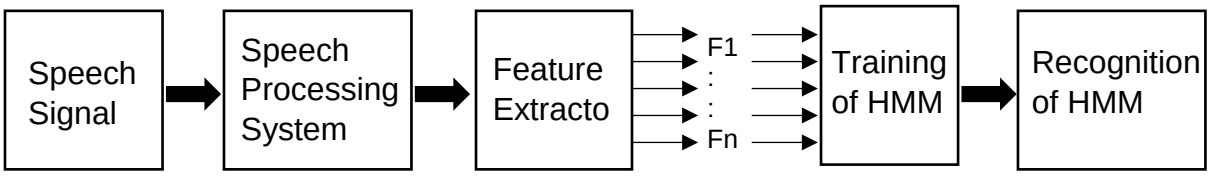


Figure 9: Block Diagram for emotion recognition

Here speech signal is received from the user and it is processed to extract features. Different features are extracted and passed through different functions to train the machine through features using HMM. After training, it recognizes the class of emotions and the output is given.

Form very long time, HMM has been used for the speech recognition but it is rarely used for the emotion classification. Speech is in large frame so HMM model can be flexible since it supports sequence of states. Hidden Markov Model Toolkit (HTK) prototype is built using HMM. For HTK passes through different stages. In Dara preparation phase; HSLAB, HCOPY HLIST and so on can be used for Speech and for transactions; HLED, HLSTATS are used. In this paper, HCOPY is used for data preparation on speech. In training phase; HINIT, HREST, etc. can be used. In testing phase; HVITE is used and for analysis phase; HRESULTS is used. After training recognition is done. For recognition, it assigns an emotional label to the query used for data provided. It gives result in the log probability and which class has the highest log probability that is the class we need (Chandni, 2015).

Analysis:

Using HMM Author has classified four different emotions; sad, fear, surprise and disgust. Only 50 audios have been used for the training which can be enhanced using more datasets. Here silence in the voice has been neglected.

This cleaning process can be beneficial in speech processing. Due to this cleaning the processing time for executing can be relatively fast while working on large speech database. MFCC has been used as a major feature which shows the importance of MFCC feature in emotion recognition.

2.6. A Study of Speech, Speaker and Emotion Recognition using MFCC and SVM

Humans have vocal tracts and articulators to produce speech and cochlea to detect the speech. These days computer can process the speech signal from the user and can give different result which has been possible from the digital signal processing. Computers can understand verbal commands or action from human and can work accordingly. For speech processing in-depth analysis is necessary in speech samples. Here (Rajasekhar & Hota, 2018) has used the concept of MFCC, and the Euclidean Distance. The emotion recognition process is more extensive than speech and speaker recognition. Pitch and amplitude can represent for the emotional state of the speaker. Happiness, sadness, fear, disgust, boredom or neutrality can be presented by these features. Stress detection and detection of other emotional states are essentials for different purposed. Here in the paper, Author has discussed more on Mel Frequency Cepstral Coefficient (MFCC) and Support Vector Machine (SVM).

MFCC is taken as the key feature for most of the speech related work like speaker or speech recognition and so on including emotion recognition. Spectral features are highly sensitive so MFCC is also affected by the environment very soon. It takes background noise and the performance of MFCC is not high. MFCC ignores the impact of pitch frequency which eventually decreases its performance. According to (Rajasekhar & Hota, 2018) speech refining process can be categorised into three types; filter-based compensation, noise model-based compensation and empirical compensation. Since new evolvement is going on, modern MFCC extraction can be done.

Stepwise MFCC extraction process is given in Figure:6. Speaking frequency envelope (also called as SMFCC) and weighted window can improve the speech recognition rate in comparison to the traditional MFCC. SMFCC is called as Smooth Mel Frequency Cepstral Coefficient which extracts the envelope of the signal post DFT/FFT and then only it is passed it to Mel Frequency filter banks while extracting the MFCC features. The influence is also accounted by the windows. Raised cosine window upgrades the Cepstrum which is suitable for speech recognition whereas half raised sine window is suitable for a speaker recognition. Traditional plain MFCC has given 84% of accuracy whereas SMFCC with a weighted has given 91% of accuracy.

Research of the (Rajasekhar & Hota, 2018) reveals that windowing is necessary where hamming has given the better result. Authors have used Vector Quantization (VQ) classification method which converts the audio signal into code vectors form. Experiment of the authors has shown the result form VQ is better than Nearest Neighbour.

SVM has been used for the emotion recognition problem since it can identify the patterns and analyse the information. It is one of the most popular algorithms used for emotion detection. SVM has evolved into four types; general separating hyperplane, maximum margin hyperplane, soft margin and Kernel method. Authors has used Kernel method since it works on non-linear. Kernel function defines the inner products to emulate polar coordinates of the given data in the Cartesian frame (Rajasekhar & Hota, 2018).

Analysis

Authors has used MFCC features with VQ and SVM algorithm. The technique of cleaning MFCC (modern SMFCC) has improved the accuracy result. Windowing is necessary, but the research of Authors has given hamming windowing as a better windowing process. Not only this but the VQ is better than Nearest Neighbour. MFCC cleaning and windowing techniques can be better if it is used for the other emotional recognition problem since spectral features are highly sensitive by the noise and background environment.

In this paper only MFCC has been used but other important features have been missed. Other features like LPCC can be used by cleaning the noise factor so that the result obtained is more efficient.

The over analysis of the literature review is given below:

Literature Review Comparison Table:

Literature Review Paper	(Akash Shaw, 2016)	(Saikat Basu, 2017)	(Steven A. Rieger, 2014)	(M.S. Likitha, 2017)	(Chandni, 2015)	(Rajasekhar & Hota, 2018)	This Project
Datasets	-	Berlin Emotion Speech dataset (EmoDB)	LDC	-	SAVEE	-	SAVEE
Features	Energy, Pitch, Formant Frequency, MFCC	13 MFCC with 13 Velocity and 13 acceleration	CEP, MFCC, LSF, ACW, PFL	MFCC	MFCC	MFCC	13 MFCC
Algorithm	ANN	CNN, LSTM, RNN	KNN	If-else condition	HMM. HTK	SVM, Vector Quantization (VQ)	KNN, SVM, ANN
Accuracy	86.87% (average)	80%	95%	80%	78% (average)	91%(SMFCC), 84%(MFCC)	76%, 68.75%, 66.66%
Emotions	Neutral, Angry, Happy, Sad	Disgust, Fear, Happy, Boredom, Neutral, Sad, Angry	Disgust, Happy, Angry, Neutral, Panic, Sad,	Happy, Angry, Sad	Surprise, Sad, Fear, Disgust	Happy, Fear, Sad, Angry	Sad, Happy, Angry, Fear
Merits	Most of the spectral features are used	Overcome of RNN limitation	Ensemble KNN has been used which is better than KNN	Standard Deviation has been used.	Silence in audio has been removed.	Noise Cleaning (using SMFCC), Hamming windowing, NN and VQ comparison.	Mean and standard deviation, both has been used.
Lacking	Noise is not cleaned.	Bidirectional LSTM could give better result.	Minkowsky distance can be used and compared.	Increase in emotion class, may not be suitable.	Only 50 audios have been used for training.	Only MFCC features has been used.	No cleaning Limited dataset

Table 1: Literature Review Comparison

In this project, performance of different algorithm has been compared. It takes Mean and Standard deviation of 13 MFCC to train the model.

2.7. Other required research

Speech is very complex which is represented in a signal format. Speech is in Analog form which need to be converted into digital form which includes various mathematical steps. Human voice is Analog wave which is captured by microphone and converts to Analog signal and the stored data are represented in digital format (binary 0 and 1). Digital Signal Processing (DSP) has some merits to process like it increases the accuracy, reliability in the digital communication field also reduces noise and distortion. It has some drawback that if signal is weak, signal cannot be amplified if it is already digitized, loss of data or sometimes it is very hard or impossible to convert to digital. But DSP has great application on speech recognition, processing, signal analysis and so on which has been used in telephone, military, space and so on (Nilu Singh, 2015). The features extracted from speech signal rely on two features; prosody and spectral. Prosodic features are generated from vocal cord's fold and Spectral features are generated from spectral content of the original speech signal. Speech rate, pitch, intensity are samples of prosodic whereas Mel Frequency Cepstral Coefficients (MFCC) and Log Frequency Power Coefficients (LFPC), Linear Prediction Cepstral Coefficients (LPCC) are samples of spectral features (Anjum, 2019). These features can be used for emotion classification of other speech processing work. While using these features on machine learning, features need to be fit and trained using different machine learning algorithms. K Nearest Neighbours (kNN) algorithm called as lazy learning method. In kNN, Similar samples belong to the same class having high probability. kNN is sensitive to the selected values of k which can be resolved with different techniques where spare learning based kNN method has been used in (Debo Cheng, 2014) paper. kNN is unsupervised learning. There are other supervised learning algorithms like ANN, SVM, CNN, RNN and so on. Artificial Neural Network (ANN) imitate the working of human brain.

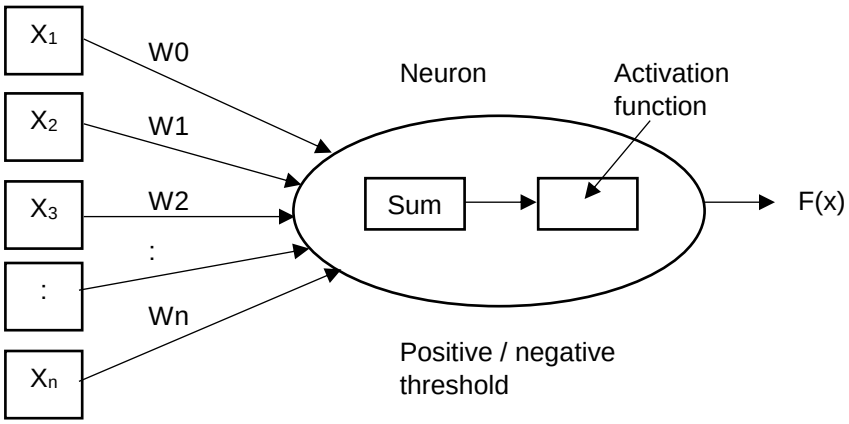


Figure 10: ANN one neuron model work

Each neuron carries its values multiplied with weight and bias is added if any then all values are added passed through activation function like sigmoid, softmax, relu and so on and one output is generated which can be used in another layer as input as shown in Fig:7 (Harsh Kukreja, 2016).

3. Artefact Development

3.1. Artefact Development Process

The development of artefact is not easy. Different research is done and slowly the development is carried out. Evolutionary prototyping has taken place for the development of this project because the field of artificial intelligence is new and so the requirements were not clear. Therefore, it has to be researched and implement in the system which could be changing as per the research. Different stages like planning, requirement analysis, designing, developing and testing were carried out in an iterative way. Different Algorithms like KNN, ANN, SVM is used while developing the product and finally trained model is integrated in Django. Django takes the voice of the user and gives the output from the saved model.

All stages are explained in detail

Level 1	**Level 2**	**Level 3**
1) Emotion Detection from Voice	1.1) Planning	1.1.1) Scope Identification and Definition 1.1.2) Feasibility Study 1.1.3) Resource Planning
	1.2) Requirement Engineering	1.2.1) Requirements Gathering and Analysis
	1.3) Designing	1.3.1) UML Designing
	1.4) Developing	1.4.1) Developing of working Model in Python Framework 1.4.2) Modules Integration in Django 1.4.3) System Development
	1.5) Testing	1.5.1) Black Box Testing 1.5.2) White Box Testing
	1.6) Deploying	1.6.1) Evaluation of Final Report

Table 2: Tabular representation of work-break down structure

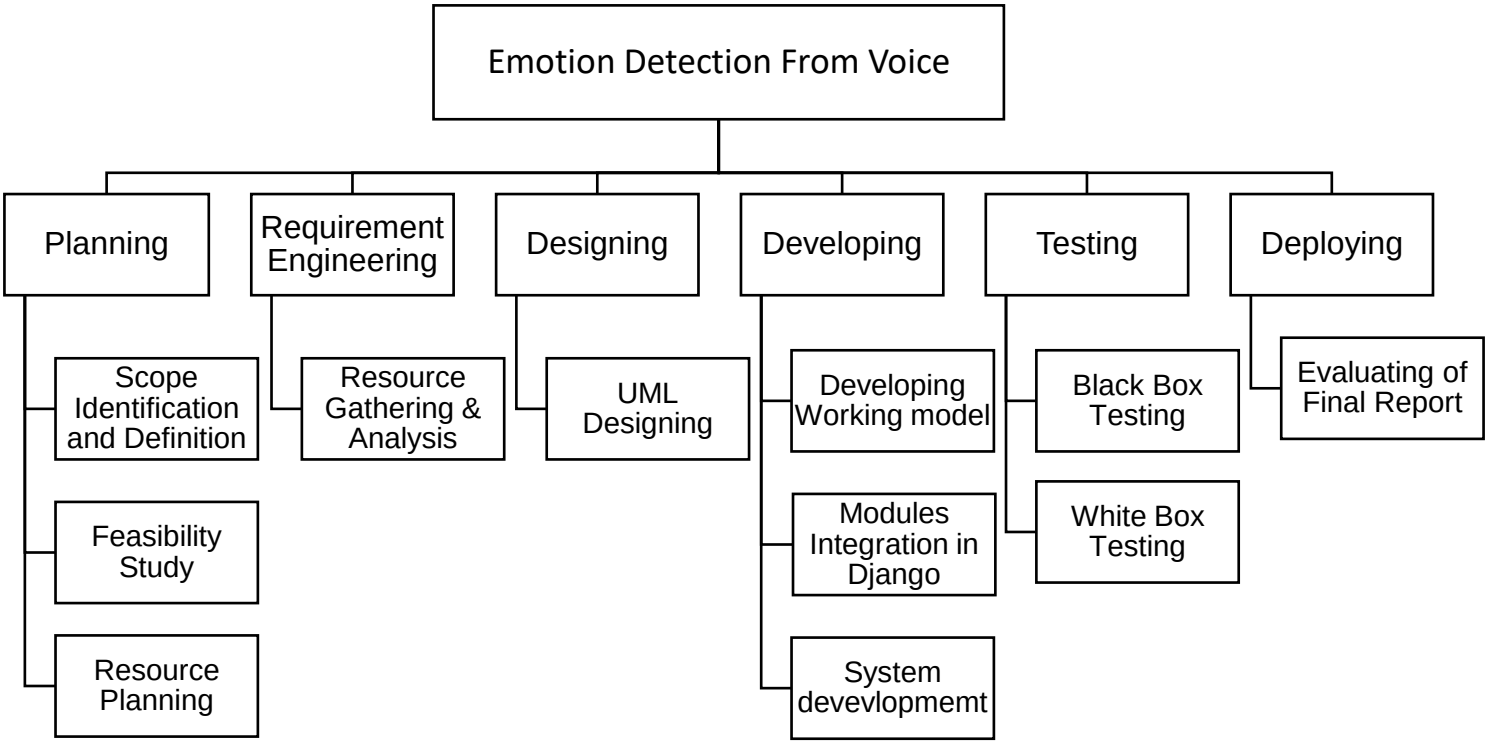


Figure 11: Tree diagram of Work-breakdown structure

3.1.1. Planning and Analysis

Brainstorming is done, what are the research required, possible steps to carry out the project successfully. Scope of the project is defined and feasibility study for technical, scheduled, economic and operational is done. Resources required are also planned along with the risk planning for backup.

Project scope is defined so that the project can be completed on given time. Research planning is done on the speech signals and its parameters and to know how speech signals flows and how speech signals are converted

in digital signals. Then extraction of features from voice is done and different planned to use features like energy, pitch, MFCC, LPCC to train the model. Using K-Nearest Neighbour (KNN), Artificial Neural Network (ANN) and Support Vector Machine (SVM) learning algorithms and comparing the result obtained. But it is difficult to carry out all planned things mainly because of time constraints. So, the planning is changed time and again. Finally, feature extraction planning is limited up to MFCC and energy as chroma.

Apart for the development planning, time management and feasibility planning are another difficult task. Technical and schedule feasibility are more applicable here.

3.1.2. Requirement Engineering

Most time-consuming phase is requirement gathering and analysis phase. Gathering of database took lots of time. Even going through lots of research paper, it took lots of time to find required dataset. Finally, SAVEE databased is found. Similar work done research paper is gathered to enhance knowledge. Feature extraction process, different algorithm is studied for number of days. Different online course is taken in order to get the knowledge to complete the project. Features from voice signal is taken and saved in the csv format. But the problem is started there. The extracted feature is saved but it is saved in an uncleaned format. 13 MFCC features is taken and 12 chroma energy was extracted and saved but each MFCC from 1 to 12 is in large array form with exponential symbol and 12 chroma is in complex number format like 2 + 3j.

MFCC_4	MFCC_5	MFCC_6
[1.70321168e+01 2.62643775e+01 2.94168616e+...	[21.34045448 15.88653582 14.85254689 16.33...	[4.29654062e+00 6.42556511e+00 7.11390531e+...
[1.74628399e+01 2.39660202e+01 2.87282268e+...	[22.77101234 16.20472163 12.37627163 12.20...	[1.22895162e+00 4.78902291e-01 5.13272630e+...

Figure 12: MFCC feature exponential problem

chroma_6	chroma_7	chroma_8
[0.00000000e+00+0.00000000e+00j 0.00000000e+00...	[0. +0.j 0. +0.j ...	[0. +0.j 0. +0.j ...
[0.00000000e+00+0.j 0.00000000e+00+0.j...	[0.00000000e+00+0.j 0.00000000e+00+0.j...	[0. +0.j 0. +0.j ...
[0. +0.j 0. +0.j ...	[0. +0.j 0. +0.j ...	[0. +0.j 0. +0.j ...

Figure 13: Chroma feature complex problem

3.1.3. Designing

System designing is necessary to understand the working of the system. Use-case diagram, System working mechanism diagram, Machine learning workflow diagram are presented in this project.

Gantt Chart Designing:

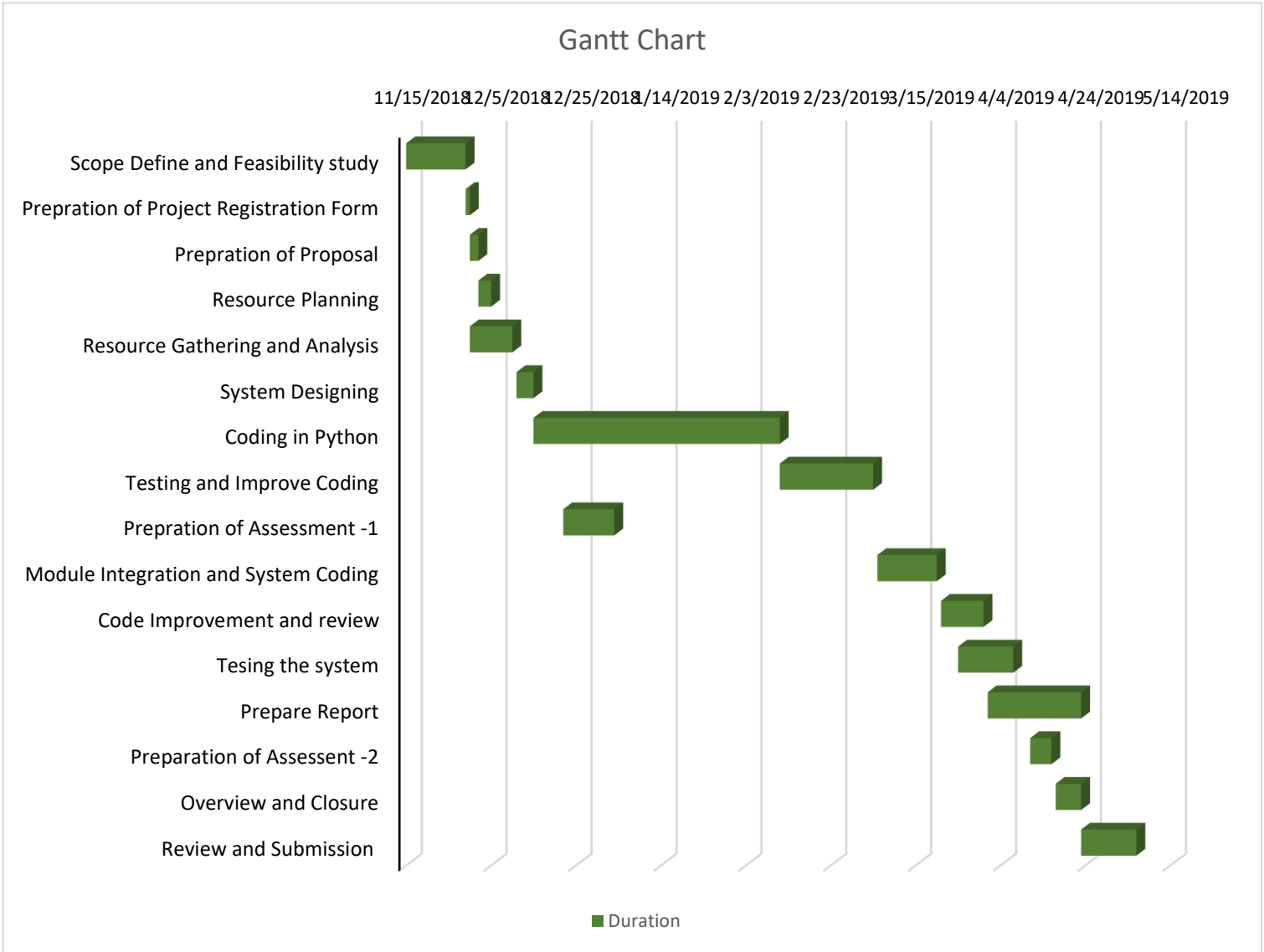


Figure 14: Gantt Chart representation

The above-mentioned Gantt chart is from proposal. Most of the duration are as per it but some stages took longer than expected. Resources gathering, and analysis took longer period than it was expected. In coding section, more time was consumed to clean extracted features format. Since submission time increased so code review and report managing time increased.

Use Case Designing:

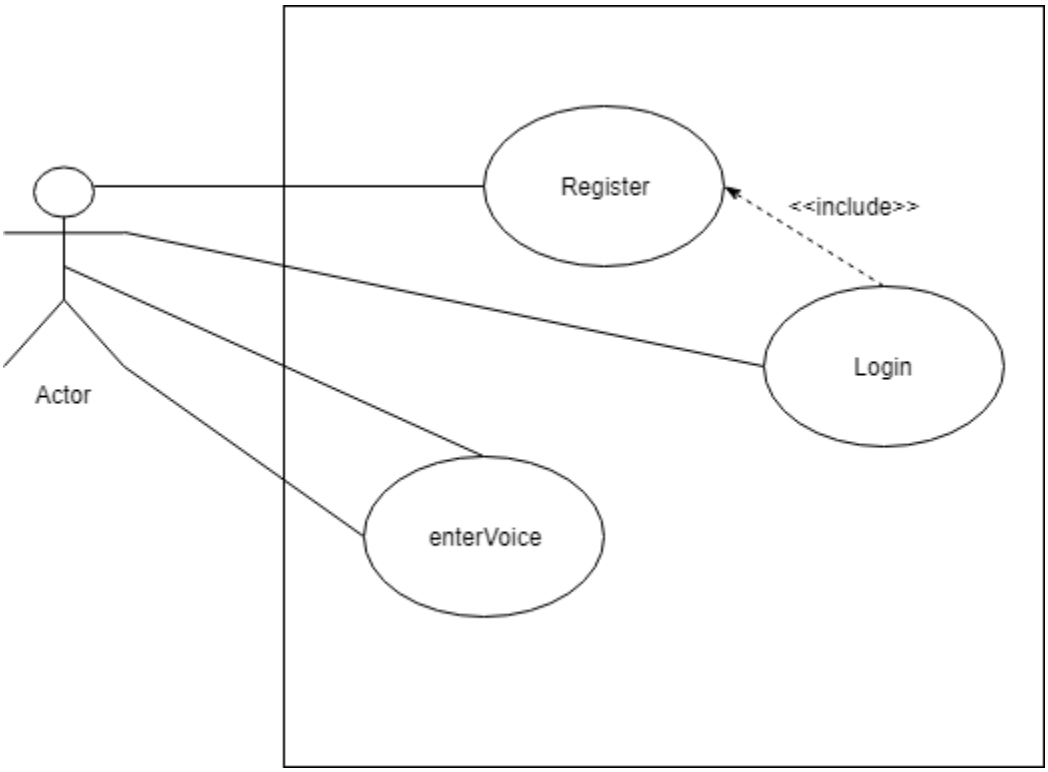


Figure 15: Use-Case Diagram

Use-Case diagram, figure 14 explains the user interaction with the system. User can register into the system. To login into the system, registration is compulsory. User cannot login without registration. In this system login is not compulsory to check user's emotion state. So, user can directly check emotional state without registration or login.

System takes the speech from the user and extracts its features mainly MFCC and it is passed to the trained model and the result is given by the system to the user.

Overall System Representation:

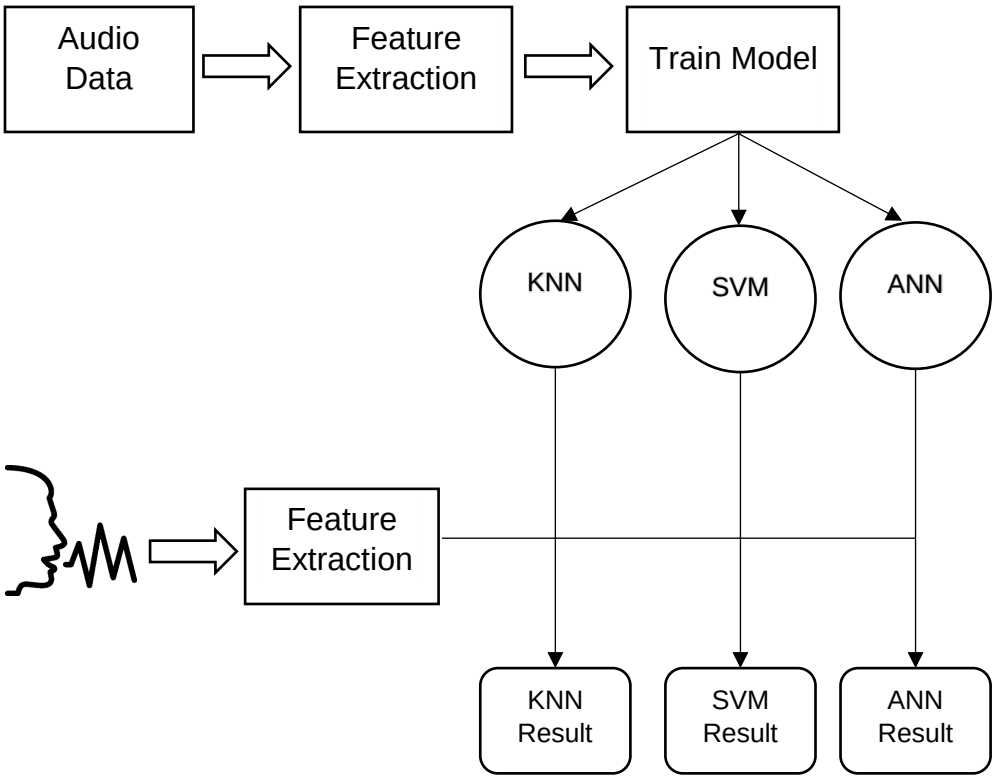


Figure 16: Full Flow of the system

The overall work of the system is represented in the above graph. Features are extracted from the audio data set (SAVEE database) and it is trained using different algorithms (KNN, SVM, ANN). Now, when user gives speech, features are extracted, and result can be obtained from the trained model. The result can be used as per the requirement.

3.1.4. Developing

Development is the actual coding part. In this part all the gather knowledge and requirements are used to develop the actual product. Here project has two sections; training and building model part, another is taking user voice and extracting required features from the given voice. The continuous development was done and it leaded for the different versions. Finally saved model is saved and integrated in Django Web framework. The whole steps and progress have built three prototypes.

3.1.4.1. First Prototype:

Features are the primary requirement so collected data from SAVEE database are used to extract features from it. Python Librosa library is used for the feature extraction. First 13 MFCC and first 12 chroma features are extracted and saved in csv file. The result obtained was complex number and uncleaned data as shown in figure 11 and 12. This problem stuck for longer period of time. To solve this problem, before storing the features obtained from Librosa library directly, the obtained 13 MFCC features are converted into np array list and chroma (array of complex number) converted tuple of rare and imaginary number using numpy library and then the feature is saved in csv. Now the result obtained was 13 MFCC with number of array values in each MFCC and 12 chroma features with number of array values in each chroma in tuple format.

	MFCC_1	MFCC_2
1	4739, -512.6311741116173, -509.48367762254827, -509.97173334905335, -515.2681681893441]	872, 57.33939073992123]
2	301114, -530.8778295069496, -532.143563967187, -531.0155217620235, -447.76096910384064]	56, 38.081233308878765]
3	1282, -499.7806181035588, -499.5472604450893, -500.44608487436415, -503.68778486936577]	944, 57.37908966335927]
4	4445119, -744.918585372184, -742.3682030031817, -743.9335592900121, -747.3244568951083]	899, 94.60599599549603]
5	611474, -505.1222718888428, -503.2824233521147, -504.0122014557964, -444.2949370399144]	65, 34.020934955968826]
6	590183, -558.3759508386614, -556.8333878545842, -557.4049071239866, -558.4223747787152]	107, 97.22931131385354]
7	6649, -477.2143885525139, -476.6102855057175, -478.7820703048773, -454.23566608020826]	099, 32.78974035795579]
8	023609, -744.7079802154356, -746.2421951821757, -745.5135003631394, -744.0836159033869]	506, 93.77827370351498]
9	7433681, -519.9659526671431, -519.806652232275, -519.5709756461108, -519.6105820024742]	85, 62.017506898173664]
10	098138, -557.6251078585395, -558.649065409907, -496.9444186518622, -423.55182200751034]	134, 20.50482871238883]
11	268374, -434.4149155619186, -435.3187426934045, -432.7660327447847, -431.4888922822054]	39, 27.705764413618112]
12	766351, -763.5477064911267, -756.7565391153597, -739.5616854188991, -709.8398748991218]	189, 51.21512336404892]
13	826089, -510.1347745538544, -507.7429857115262, -505.6968137718226, -440.3705603640072]	674, 37.37998963135459]
14	71949, -546.5779823254785, -546.5681020807808, -546.0885644645915, -441.40362097004055]	53, 29.135367227198863]
15	54, -470.60875464930643, -469.46288759019774, -468.29261754988545, -471.49034379687066]	854, 35.94926686964885]
16	4052315, -722.0490033058455, -722.6696036144547, -724.145778573956, -725.4707973199471]	789, 77.48007727205604]
17	390573, -536.9762685404161, -534.2212571063749, -531.5626257918144, -530.4012603410231]	649, 58.75274029386755]

Figure 17: MFCC feature cleaned list format

	CHROMA_11	CHROMA_12	OUTPUT
	0.39114135416156964), (0.0, 0.0)]	0.44536468503711313, -0.04540876861741398), (0.0, 0.0)]	0
	2083351635), (0.0, 0.0), (0.0, 0.0)]	2816824), (0.06306863230057776, 0.19863428045335219)]	0
	25819666, 0.6599487481448154)]	), (0.0, 0.0), (0.7046194412900437, 0.5644654759266051)]	0
	0.16494044619637943), (0.0, 0.0)]	(0.40993022080979424, -0.9121168861866518), (0.0, 0.0)]	0
	0284666906), (0.0, 0.0), (0.0, 0.0)]	-), (0.2740095716579263, 0.6789796420465285), (0.0, 0.0)]	0
	, 0.6347306807377245), (0.0, 0.0)]	(0.7176370540012764, -0.27587825356610807), (0.0, 0.0)]	0
	, 0.2999190060780725), (0.0, 0.0)]	), (0.9352639651335868, -0.3539510072349005), (0.0, 0.0)]	0
	1205637013), (0.0, 0.0), (0.0, 0.0)]	, (0.41396148650939407, 0.7096566957194621), (0.0, 0.0)]	0
	54), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)]	13, -0.29756623556800654), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)]	0
	74466419, 0.7733786107646355)]	32905067948613), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)]	0
	2005160987), (0.0, 0.0), (0.0, 0.0)]	81458529043, -0.9478261070850151), (0.0, 0.0), (0.0, 0.0)]	0
	46), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)]	82479331283161), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)]	0

Figure 18: Chroma feature cleaned tuple format

Training was not possible in these data so MFCC are taken and mean value is calculated from its respective array values and saved in respective places.

MFCC_1	MFCC_2	MFCC_3	MFCC_4
-409.23355144427916	95.17242115521368	26.971322602888247	43.20156876957321
-426.10604548987396	108.43788688251144	38.655204742643676	48.4766453664175
-344.71095943864566	109.11870270946865	16.926360467099762	52.960039913942886
-653.9905249452675	129.99384599843657	22.854716227164563	47.45151898420735
-385.79161117087136	109.20238676252178	18.21352812045713	27.5064224587625
-438.0175647549851	135.81318908922995	24.52831992903859	29.08858522425039
-327.4817906415072	114.75448124203673	7.4478549426200145	30.032823146485867
-654.3248619735847	154.29536909578965	31.38437419616405	18.990318386870012
-434.86992222129027	98.56596435162825	25.44497143261313	31.652937185474016
-499.08625722855396	121.567376903786	37.6165722464132	23.9656555536768
-339.82524835569615	77.53273988320164	22.272818134778824	29.371330841492114

Figure 19: Mean value of MFCC

Since chroma features are very high and low as well as it is in rare and imaginary format, mean value was not possible so only MFCC values are taken to train the model. Sk learn and Tensorflow has been used for Machine learning. KNN and SVM is used from sk learn and ANN is used from Keras (Tensorflow).

At first KNN algorithm is used to get the result. In KNN, 4 neighbours are used and minkowski metric is used. The result obtained is 70%.

```
#show first 5 model predictions on the test data
from sklearn.metrics import confusion_matrix
out = knn.predict(X_test)
# print(out)
# print(y_test)
print(confusion_matrix(out, y_test))

#check accuracy of our model on the test data
print(knn.score(X_test, y_test)*100)
```

```
[[14  0  0  2]
 [ 0 11  6  3]
 [ 1  2  7  0]
 [ 0  2  2 10]]
70.0
```

Figure 20: KNN accuracy using mean (MFCC)

When using SVM model 64.583% accuracy is obtained using poly kernel.

```
y_pred = svclassifier.predict(X_test)
# print(y_pred)
# print(y_test)
print(confusion_matrix(y_pred, y_test))
#check accuracy of our model on the test data
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)*100
```

```
[[11  1  1  1]
 [ 0  7  4  3]
 [ 0  2  6  1]
 [ 1  2  1  7]]
64.58333333333334
```

Figure 21: SVM accuracy using mean (MFCC)

Finally, ANN is used with gradient descent optimizer as 0.0001 and hidden layer 1024,512,128,4. 500 epochs is done.

```
test_loss, test_acc = model.evaluate(X_test, y_test)
print('Test accuracy:', test_acc*100)
```

```
72/72 [==============================] - 0s 125us/step
Test accuracy: 62.5
```

Figure 22: ANN accuracy using mean (MFCC)

Best accuracy obtained is from KNN algorithm.

3.1.4.2. Second Prototype:

To increase the accuracy, some changes are made on features. Only mean has been used so additionally, standard deviation of all 13 MFCC are calculated and stored as mean values. Therefore, there were 13 MFCC mean values and 13 MFCC standard deviation values. These features are trained.

In KNN, with the same hyper-parameters, accuracy increased to 76.67%.

```
#show first 5 model predictions on the test data
from sklearn.metrics import confusion_matrix
out = knn.predict(X_test)
# print(out)
# print(y_test)
print(confusion_matrix(out, y_test))

#check accuracy of our model on the test data
print(knn.score(X_test, y_test)*100)
```

```
[[15  0  0  2]
 [ 0 10  5  2]
 [ 0  2 10  0]
 [ 0  3  0 11]]
76.66666666666667
```

Figure 23: KNN accuracy using mean and std (MFCC)

When using SVM model with the same hyper-parameter obtained result is given below:

```
y_pred = svclassifier.predict(X_test)
# print(y_pred)
# print(y_test)
print(confusion_matrix(y_pred, y_test))
#check accuracy of our model on the test data
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)*100
```

```
[[9 1 1 1]
 [0 9 4 2]
 [0 0 7 1]
 [3 2 0 8]]
68.75
```

Figure 24: SVM accuracy using mean and std (MFCC)

In ANN with the same hyperparameters, obtained result is 66.67%

```
test_loss, test_acc = model.evaluate(X_test, y_test)

print('Test accuracy:', test_acc*100)
```

```
72/72 [==============================] - 0s 139us/step
Test accuracy: 66.66666666666666
```

Figure 25: ANN accuracy using mean and std (MFCC)

3.1.4.3. Third Prototype:

Third prototype is the integration of model in Django Framework. This is also another challenging task to integrate the trained model in Django. Trained model from sk learn are saved in .sav format and tensorflow model are saved in .h5 format. Since the accuracy from KNN is better so far in comparison to other learning algorithms. Therefore, KNN saved model has been integrated in Django. In Django, user clicks the button and the system records user voice for 5 seconds and the audio file is saved in Django as output.wav. The same audio is loaded, and features are extracted (i.e. 13 MFCC mean and standard deviation values). Then the result is predicted through the knn saved model.

The result output is given below:

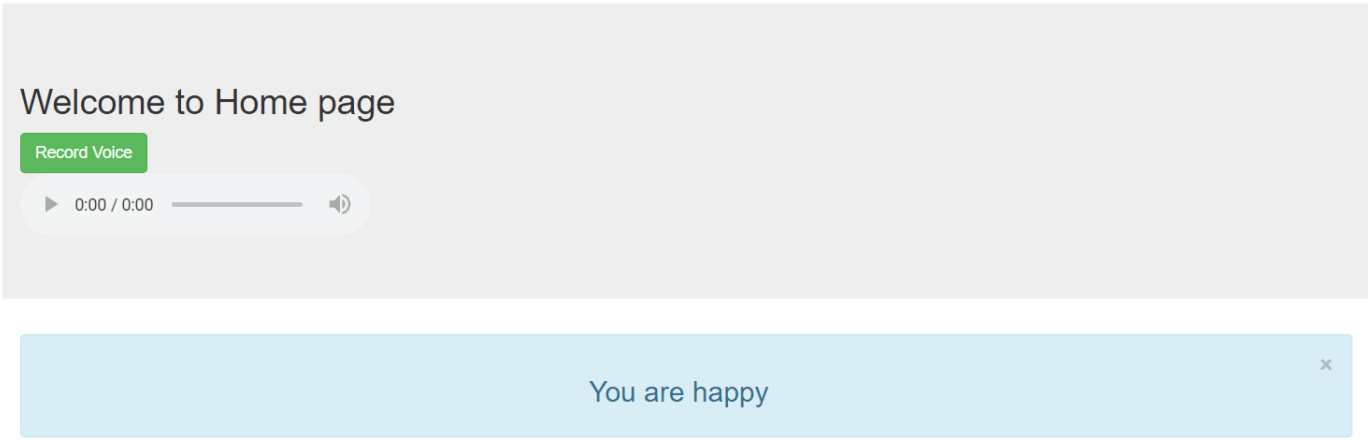


Figure 26: Output in Django

3.1.5. Testing

Testing is essential to know whether module is working fine or not and overall system is solving the problem or not. There are different types of testing like white box, black box, unit, and so on. Here white box testing and black box testing has been used.

Testing:

<table>
<tr><td>Test case no.</td><td colspan="2">1</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether audio is being recording or not</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>User voice</td><td>Should show the audio file</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
# Recording voice

p = pyaudio.PyAudio()
stream = p.open(format=FORMAT,
                channels=CHANNELS,
                rate=RATE,
                input=True,
                frames_per_buffer=CHUNK)
print("* recording started.  You have 5 seconds to record")
frames = []
for i in range(0, int(RATE / CHUNK * RECORD_SECONDS)):
    data = stream.read(CHUNK)
    frames.append(data)

print("* Successfully recording is completed")


stream.stop_stream()
stream.close()
p.terminate()
```

```
* recording started.  You have 5 seconds to record
* Successfully recording is completed
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Audio is recording</td></tr>
</table>

Table 3: Audio recording (Test 1)

<table>
<tr><td>Test case no.</td><td colspan="2">2</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check audio is being saved or not</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>User voice</td><td>Audio should be saved</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
# storing file

wf = wave.open(WAVE_OUTPUT_FILENAME, 'wb')
wf.setnchannels(CHANNELS)
wf.setsampwidth(p.get_sample_size(FORMAT))
wf.setframerate(RATE)
wf.writeframes(b''.join(frames))
wf.close()
print("Your audio has been saved")
```

```
Your audio has been saved
```

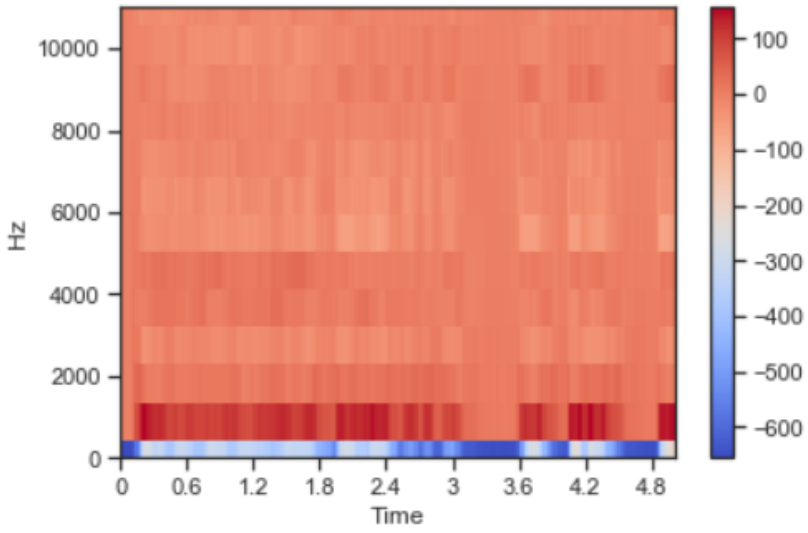
</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Audio is saved</td></tr>
</table>

Table 4: Audio saving (Test 2)

<table>
<tr><td>Test case no.</td><td colspan="2">3</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether MFCC is being extracted or not</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Audio voice</td><td>Should give 13 MFCC features</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
# mfcc
MFCC = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)
librosa.display.specshow(MFCC, x_axis='time', y_axis='hz')
plt.colorbar()
```

```
<matplotlib.colorbar.Colorbar at 0x1cb06f79588>
```



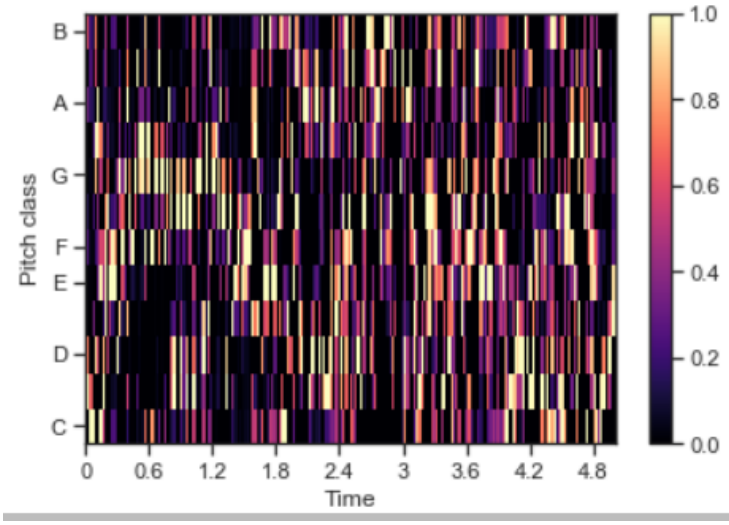
</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. MFCC data are displayed</td></tr>
</table>

Table 5: MFCC Extraction (Test 3)

<table>
<tr><td>Test case no.</td><td colspan="2">4</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether Chroma is being extracted or not</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Audio voice</td><td>Should give 12 chroma features</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
#Energy
chroma = librosa.feature.chroma_cqt(C=c, sr=sr)
print(len(chroma))
librosa.display.specshow(chroma, x_axis='time', y_axis='chroma')
plt.colorbar();
```

```
12
C:\Users\dahal\Anaconda3\lib\site-packages\librosa\display.py:696:
d.
  warnings.warn('Trying to display complex-valued input. '
```



</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. chroma data are displayed</td></tr>
</table>

Table 6: Chroma Extraction (Test 4)

<table>
<tr><td>Test case no.</td><td colspan="2">5</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether all audios files are reading of each class and extracting features (MFCC and chroma)</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Folder and file names</td><td>Display result for all files MFCC and chroma features</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
# Checking to load audio files
for i in range(15):
    if i>8:
        y, sr = librosa.load(str("audio_files/DC/sa")+str(i+1)+str(".wav"))
        y1, sr1 = librosa.load(str("audio_files/JE/sa")+str(i+1)+str(".wav"))
        y2, sr2 = librosa.load(str("audio_files/JK/sa")+str(i+1)+str(".wav"))
        y3, sr3 = librosa.load(str("audio_files/KL/sa")+str(i+1)+str(".wav"))
    else:
        y, sr = librosa.load(str("audio_files/DC/sa0")+str(i+1)+str(".wav"))
        y1, sr1 = librosa.load(str("audio_files/JE/sa0")+str(i+1)+str(".wav"))
        y2, sr2 = librosa.load(str("audio_files/JK/sa0")+str(i+1)+str(".wav"))
        y3, sr3 = librosa.load(str("audio_files/KL/sa0")+str(i+1)+str(".wav"))

    MFCC0 = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)
    c = librosa.cqt(y,sr)
    chroma0 = list(librosa.feature.chroma_cqt(C=c, sr=sr))
print(MFCC0, chroma0)
```

```
[[-522.39939461 -534.08877838 -535.9074478  ... -539.28459557
  -534.1616445  -438.24241851]
 [  58.29642087   74.95319702   77.19618653 ...   71.71761373
    72.62053898   20.63676789]
 [  57.30608149   54.7137718    52.08255048 ...   47.09893859
    44.23447143   26.65686576]
 ...
 [  11.24257619    7.41442916    2.53248732 ...    2.03776235
     2.51549318    1.83210347]
 [   4.76729511    2.86881818    3.40449714 ...    1.40858979
    -0.57029943   -4.4853406 ]
 [   6.30797697    6.6717287     4.56876163 ...    3.06229963
     0.86035432   -2.13660482]] [array([0.        +0.j        , 0.        +0.j
       0.        +0.j        , 0.        +0.j        ,
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. All files are read, and features are extracted.</td></tr>
</table>

Table 7: Audio files reading and Feature Extraction (Test 5)

Test case no.	**6**	
Test case objective	To check whether all features are saved in csv	
Test Data	Data	Expected Result
	Features	All features should be saved in csv
Actual Result		

```
# ------------------------------------------------MFCC-----------------------------------------

# creating mfcc csv file
def create_file(fileName):
    with open('extracted_feature/'+'/'+str(fileName)+'.csv', mode='w', newline='') as csv_file:
        fieldnames = ['MFCC_1', 'MFCC_2', 'MFCC_3','MFCC_4','MFCC_5','MFCC_6','MFCC_7','MFCC_8','MFCC_9','MFCC_10','MFCC_11','MFCC_12
        writer = csv.DictWriter(csv_file, fieldnames=fieldnames)
        writer.writeheader()

# appending mfcc row in created csv file
def append_row(fileName,MFCC,chroma,output):
    with open('extracted_feature/' + '/'+str(fileName)+'.csv', mode='a', newline='') as csv_file:
        fieldnames = ['MFCC_1', 'MFCC_2', 'MFCC_3','MFCC_4','MFCC_5','MFCC_6','MFCC_7','MFCC_8','MFCC_9','MFCC_10','MFCC_11','MFCC_12
        writer = csv.DictWriter(csv_file, fieldnames=fieldnames)
        writer.writerow({'MFCC_1': MFCC[0], 'MFCC_2': MFCC[1], 'MFCC_3': MFCC[2],'MFCC_4': MFCC[3], 'MFCC_5': MFCC[4], 'MFCC_6': MFCC[
```

MFCC_1	MFCC_2	MFCC_3	MFCC_4	MFCC_5	MFCC_6
97173335 -515.26816819]	95e+01 5.73393907e+01]	4.96867314 39.3809872]	31e+01 2.36638205e+01]	1.88357392 15.47085642]	83e+00 1.18842984e+01]
.01552176 -447.7609691]	9.45820604 38.08123331]	3e+01 3.44933949e+01]	99e+01 1.69704751e+01]	5.83412857 11.91688166]	28e-01 2.55755624e-01]
44608487 -503.68778487]	52e+01 5.73790897e+01]	.06062008 44.61252197]	9.61935313 28.88266649]	4.80238915 15.15433427]	5.81266756 6.17606401]
.93355929 -747.3244569]	95.4704103 94.605996]	5.86615346 37.71295263]	1.25590763 26.92228532]	7.9914096 11.67444289]	43e+00 8.92751399e+00]
01220146 -444.29493704]	1.48221587 34.02093496]	3.57996516 33.69753883]	29.74139852 18.82066]	7.09389354 16.26871211]	8.21860874 7.04039871]
40490712 -558.42237478]	9.34090286 97.22931131]	3.14813434 51.01544958]	3.38438454 16.79353889]	6.77074279 6.86504139]	72e+00 5.95211913e+00]
.7820703 -454.23566608]	3.32175926 32.78974036]	.54873034 30.56037257]	5.40855287 17.16158051]	5.78789572 17.69267725]	8.80934217 9.82543825]
.51350036 -744.0836159]	2.38833285 93.7782737]	0.7298746 41.31059466]	4.82261521 25.8001093]	0.93874381 11.30732897]	28e+01 9.45332255e+00]
9.57097565 -519.610582]	52.2016159 62.0175069]	8.36427435 47.8576956]	1.8586557 31.05280311]	3.09134976 17.38185042]	79e+00 9.54500087e+00]
94441865 -423.55182201]	9.59315637 20.50482871]	2.85863753 26.63297053]	3.18109604 16.03575955]	3.25859171 14.75637095]	4.23747364 4.25822743]
76603274 -431.48889228]	3.13778558 27.70576441]	3.54241346 24.46964425]	9.74227954 19.82082304]	5.39095679 14.65871718]	1.17242783 9.87842804]
.56168542 -709.8398749]	9.53927464 51.21512336]	3.52248878 37.88963482]	3.09368673 15.47332604]	5.51727525 16.49532753]	7.73329902 12.51565792]
69681377 -440.37056036]	2.32296018 37.37998963]	11.71754944 31.496009]	17e+01 1.46752117e+01]	9.45865673 14.10102593]	32e+01 6.74420428e+00]
98856446 -441.40362097]	59e+01 2.91353672e+01]	2.1213995 31.2360354417.70456679 17.96672836]		5.90176691 16.77135215]	5.77661453 8.61539116]

Test Result	Testing is successful. All features are saved in csv.

Table 8: Saving Features in CSV (Test 6)

<table>
<tr><td>Test case no.</td><td colspan="2">7</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether stored csv features are extractable</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>CSV file</td><td>Output: number format</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
df = pd.read_csv('extracted_feature/first_exponencial_features.csv')
df.MFCC_1[0]
```

```
'[-505.42593124 -512.02233338 -512.71170979 -513.12405621 -514.16350428\n -514.3605
-514.83942141 -515.2403744  -514.76012238 -513.74038047 -516.91388287\n -516.881405
-514.60769793 -502.69758197 -449.15205634 -396.20545422 -367.27208747\n -360.714816
-263.19739243 -285.91174263 -301.04590495 -340.43080283 -346.82016922\n -325.426726
-446.01412356 -327.0573924  -244.08051118 -223.71153275 -243.59170161\n -289.579920
-365.37017272 -435.26076036 -425.9321187  -344.14112086 -316.14211169\n -329.157316
-297.9584854  -311.9064128  -348.378037   -362.00568004 -374.38187065\n -399.425178
-355.61641979 -382.81368172 -425.22540531 -480.23676844 -518.309361\n -525.79242458
19.11475619 -517.61196676 -520.2701755  -521.0071707  -502.11825367\n -431.36416548
23.64272399 -378.839603   -333.08517882 -318.15321725 -325.48481628\n -336.99812187
50.50781735 -360.47535222 -371.46025303 -371.06968447 -381.38437896\n -412.16552061
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing failed. Features are not readable in numeric, but it is in string format</td></tr>
</table>

Table 9: Stored features format -String (Test 7)

<table>
<tr><td>Test case no.</td><td colspan="2">8</td></tr>
<tr><td>Test case objective</td><td colspan="2">Converting string features to number format</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Feature data</td><td>Should display numeric array for MFCC</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
df = pd.read_csv('extracted_feature/features.csv')
ast.literal_eval(df.MFCC_1[0])
```

```
[-505.42593124260526,
 -512.0223333762358,
 -512.7117097930369,
 -513.1240562091313,
 -514.163504276468,
 -514.3605546563824,
 -514.1423053870188,
 -514.3018471843733,
 -513.6646158722322,
 -514.4280175997261,
 -514.839421414764,
 -515.2403743973856,
 -514.7601223826058,
 -513.7403804742362,
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Features are converted into numeric format.</td></tr>
</table>

Table 10: Converting Features into numeric format (Test 8)

<table>
<tr><td>Test case no.</td><td colspan="2">9</td></tr>
<tr><td>Test case objective</td><td colspan="2">Taking mean value of all 13 MFCC of each feature.</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>MFCC features</td><td>Should give single mean value.</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
import ast
import numpy as np

for index, row in df.iterrows():
    meanV = []
    output = row[25]
    for i in range(13):
        m1 = np.asarray(ast.literal_eval(row[i]))
        meanV.append(np.mean(m1))
    append_row(fileName3, meanV, output)
#     print(meanV)
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Accurate mean is given</td></tr>
</table>

Table 11: Taking Mean from MFCC (Test 9)

<table>
<tr><td>Test case no.</td><td colspan="2">10</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether mean values are being saved or not</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Calculated mean</td><td>Mean value should be saved in csv file</td></tr>
<tr><td>Actual Result</td><td colspan="2">
<table>
<tr><th>MFCC_1</th><th>MFCC_2</th><th>MFCC_3</th></tr>
<tr><td>-409.23355144427916</td><td>95.17242115521368</td><td>26.971322602888247</td></tr>
<tr><td>-426.10604548987396</td><td>108.43788688251144</td><td>38.655204742643676</td></tr>
<tr><td>-344.71095943864566</td><td>109.11870270946865</td><td>16.926360467099762</td></tr>
<tr><td>-653.9905249452675</td><td>129.99384599843657</td><td>22.854716227164563</td></tr>
<tr><td>-385.79161117087136</td><td>109.20238676252178</td><td>18.21352812045713</td></tr>
<tr><td>-438.0175647549851</td><td>135.81318908922995</td><td>24.52831992903859</td></tr>
<tr><td>-327.4817906415072</td><td>114.75448124203673</td><td>7.4478549426200145</td></tr>
<tr><td>-654.3248619735847</td><td>154.29536909578965</td><td>31.38437419616405</td></tr>
<tr><td>-434.86992222129027</td><td>98.56596435162825</td><td>25.44497143261313</td></tr>
</table>
</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Mean value is saved in csv file.</td></tr>
</table>

Table 12: Saving Mean to CSV (Test 10)

<table>
<tr><td>Test case no.</td><td colspan="2">11</td></tr>
<tr><td>Test case objective</td><td colspan="2">Calculation of standard deviation</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Dataset</td><td>Standard deviation values should be returned.</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
import numpy as np
a = [1,2,3,4,5]
print(np.mean(a))
print(np.std(a))
```

```
3.0
1.4142135623730951
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Sample standard deviation values are correct.</td></tr>
</table>

Table 13: Calculating Standard Deviation (Test 11)

<table>
<tr><td>Test case no.</td><td colspan="2">12</td></tr>
<tr><td>Test case objective</td><td colspan="2">To add standard deviation and mean in a new csv file</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Dataset</td><td>Should save mean and standard deviation in csv.</td></tr>
<tr><td>Actual Result</td><td colspan="2">
<table>
<tr><th>MFCC_13_m</th><th>MFCC_1_v</th></tr>
<tr><td>0.04634035004494635</td><td>84.6291463120363</td></tr>
<tr><td>0.2745775937730099</td><td>84.04001815919807</td></tr>
<tr><td>-2.307078801506743</td><td>113.40958189237892</td></tr>
<tr><td>-0.5757428987194816</td><td>74.95317666928831</td></tr>
<tr><td>-4.740435691754554</td><td>92.46932306749862</td></tr>
<tr><td>0.8410681076127802</td><td>103.0067698624432</td></tr>
<tr><td>0.12355116058133732</td><td>127.51437130254098</td></tr>
</table>
13th mean MFCC and 1st standard deviation MFCC
</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Mean and standard values are saved in csv file.</td></tr>
</table>

Table 14: Saving Mean and Standard Deviation in CSV (Test 12)

<table>
<tr><td>Test case no.</td><td colspan="2">13</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether KNN is working and giving result</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Mean and standard deviation csv</td><td>Features should be trained and give result.</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
#show first 5 model predictions on the test data
from sklearn.metrics import confusion_matrix
out = knn.predict(X_test)
# print(out)
# print(y_test)
print(confusion_matrix(out, y_test))

#check accuracy of our model on the test data
print(knn.score(X_test, y_test)*100)
```

```
[[15  0  0  2]
 [ 0 10  5  2]
 [ 0  2 10  0]
 [ 0  3  0 11]]
76.66666666666667
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Model is trained, and accuracy is obtained</td></tr>
</table>

Table 15: KNN Working (Test 13)

<table>
<tr><td>Test case no.</td><td colspan="2">14</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether ANN is working and giving result</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Mean and standard deviation csv</td><td>Features should be trained and give result.</td></tr>
<tr><td>Actual Result</td><td colspan="2">

```
test_loss, test_acc = model.evaluate(X_test, y_test)

print('Test accuracy:', test_acc*100)
```

```
72/72 [==============================] - 0s 139us/step
Test accuracy: 66.66666666666666
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Model is trained, and accuracy is obtained</td></tr>
</table>

Table 16: ANN Working (Test 14)

<table>
<tr><td>Test case no.</td><td colspan="2">15</td></tr>
<tr><td>Test case objective</td><td colspan="2">To check whether SVM is working and giving result</td></tr>
<tr><td rowspan="2">Test Data</td><td>Data</td><td>Expected Result</td></tr>
<tr><td>Mean and standard deviation csv</td><td>Features should be trained and give result.</td></tr>
<tr><td></td><td colspan="2">

```
y_pred = svclassifier.predict(X_test)
# print(y_pred)
# print(y_test)
print(confusion_matrix(y_pred, y_test))
#check accuracy of our model on the test data
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)*100
```

```
[[9 1 1 1]
 [0 9 4 2]
 [0 0 7 1]
 [3 2 0 8]]
68.75
```

</td></tr>
<tr><td>Test Result</td><td colspan="2">Testing is successful. Model is trained, and accuracy is obtained</td></tr>
</table>

Table 17: Working of SVM (Test 15)

System whole testing:

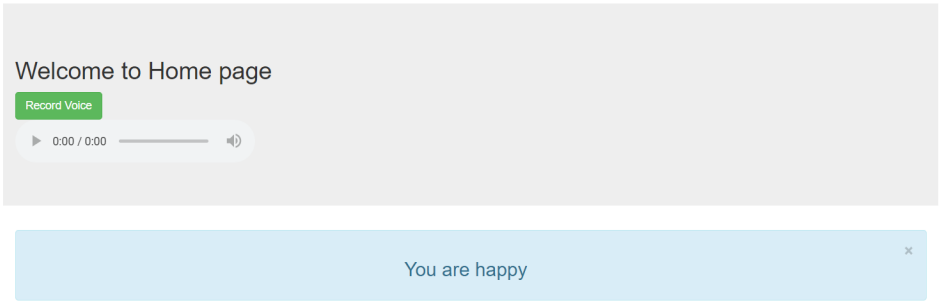
Test case no.	**16**	
Test case objective	To check Django gives the result of audio or not	
Test Data	Data	Expected Result
	User voice	Should give emotions (sad, happy, angry, fear)
Actual Result		
Test Result	Testing is successful. Django gives notification message with his / her emotions.	

Table 18: Overall system outcome (Test 16)

3.2. Tools and techniques:

Different tools and techniques are used to accomplish the aim of the project. Used tools are explained below:

a. Ms. Office and .txt

From Ms Office, Excel and word are used. Ms. Office is not an open software, but it is provided by the University. Ms. Office is easier and user friendly and it is runs offline. So Ms office is selected for this project. Excel is used to make some graphs and words are used to save documents research planning and so on. Sometimes .txt file is used to save data like word count planning for report is stored in count.txt file.

b. Anaconda

Anaconda is one of the good open source python software. Anaconda has been used here to extract features and train the model using python language. Python is a high-level language. Large dataset can be used easily, and machine learning is easy to perform in python because of its different available library.

c. Django

Django is a web framework on python. Django can understand python language easily. Since this project is done using python, it would be easy to work on python framework, therefore Django has been selected for the web view.

d. Draw.io

Draw.io is a free online designing software to draw different charts, figures and graphs. Use case is an example designed using draw.io.

e. Librosa Library

Librosa library is a python library. It allows to extract speech features easily and suggested by many other researchers. It focuses more on

spectrum features like MFCC. It is the most essential feature for project like speech processing. Therefore, this library has been selected.

f. PyAudio

PyAudio is also a python library which is easy to take user voice specifying the time to record the voice. Here in this project, user has to give his/her voice and being based on that, emotion is provided back to user.

g. Sk-learn

Sk-learn is a machine learning library of python. Here KNN and SVM are done using sklearn library. Sk learn is easy to use.

h. Tensor-flow

Tensor-flow is also machine learning library of python but more focused on deep learning. ANN learning algorithm has been used through keras library of tensorflow. Tensor-flow provides more features than sk learn for deep learning like epoch, optimizer, graphical representation. Hence for ANN, tensor-flow has been used.

Used Techniques are given below:

a) Learning Algorithm (KNN/ANN/SVM)

Different learning algorithm has been used in this project to compare the results. KNN is an unsupervised learning algorithm where ANN and SVM are supervised learning algorithm. Here KNN has been proposed to use as learning algorithm but research reflected to use and compare ANN and SVM. Therefore, ANN and SVM has also been used along with KNN.

b) Feature Extraction Technique

Feature extraction is one of the difficult tasks. But librosa and pyaudio are good libraries which allows to work on speech. The technique to record audio and extract features like MFCC and chroma are possible.

c) Model integration

Trained model can be used in many ways. REST-Api was supposed to be used to use it in Django framework but direct use of trained file storing it in Django was easy. So, integration was done in this way.

4. Answer of Academic Question:

How to detection emotion of a person is a challenging problem to solve. But to search for the better algorithm which can perform well is a part which requires lots of research and the result obtained after then are always amazing. Here in this report, 13 MFCC features are extracted where in most of the research papers (Rajasekhar & Hota, 2018) (Saikat Basu, 2017) has used MFCC as a major feature. (Saikat Basu, 2017) project has used standard deviation of MFCC and got 80% of accuracy for three different emotional classes. Comparison of different learning algorithm with the result they are gives the answer of better algorithm. And successfully the model has been saved in Django. The system is working exactly how it was aimed with 76% of accuracy from KNN algorithm.

5. Conclusion and Future Escalation:

Finally, Emotion detection is a difficult task because same situation can give different feelings and emotions to people. It is more difficult if it is only based on speech factor. Different speech parameters are required for the emotion recognition process. Here in this project, MFCC and chroma has been extracted and cleaned. Complex number has been changed in tuple format. To train the model only MFCC is selected. Its mean and standard deviation is used. It is trained through different algorithms like KNN, SVM and ANN. Just training the model with mean values gave highest accuracy of 70% and with mean and standard deviation gave 76.67% in KNN algorithm. KNN has given better accuracy than SVM and ANN. But the accuracy of ANN can be increased more by increasing number of epochs and other hyper-parameters. At last the trained and saved model is integrated in Django for the user interface. In Django, can click the button record audio and receive his/her emotion.

This project can be built more efficiently in the future. Only MFCC has been selected for the emotion recognition which can be enhanced by adding other additional spectrum features like LPCC, energy, pitch and quality-based features and so on. Since spectrum feature, MFCC is affected by the noise and environmental sound, so noise can be removed or neglected with different windowing techniques or algorithms like cocktail party problems. The dataset used are limited up to 60 on each class which can be increased using different available databases like EmoDB, google research, etc. Only four classes have been selected for classification, but it can be increased more. In future by enhancing this project, this can be used in different modules.

6. Critical Evaluation:

Most of the time we dream for big but cannot accomplish it. While taking this project, “Emotion Detection from voice” the doubt was hovering whether it would be complete or not but finally it has been completed successfully which is a great achievement. This project has not been achieved as much as it was planned for. Different other features like LPCC, energy, pitch was also supposed to be use for the better result, but all these were not applicable in the project within the limited given time. Feature handling can be taken as proud step in this project because the problem of exponential and complex number took number of days to be solve. Even after having all these troubles, planning made was implemented successfully one by one. Time management was challenging due to which project was completed late. But throughout the completion of the project different knowledges were gathered from different sources like Journals, Conferences, Books and Online courses. The project was completed successfully because of the different software and libraries used. Finding of all software, techniques, database are from the references given in the research papers. In overall, the project has been completed successfully.

References

Akash Shaw, R. K. V. S. S., 2016. Emotion Recognition and Classification in Speech using Artificial Neural Networks. *International Journal of Computer Applications,* Volume 145, pp. 5-9.

Anjum, M., 2019. *Emotion Recognition from Speech for an Interactive Robot Agent.* Paris, IEEE.

Chandni, G. V. M. K. D. K. R. J. P., 2015. *An automatic emotion recognizer using MFCCs and Hidden Markov Models.* Brno, IEEE.

Debo Cheng, S. Z. Z. D. Y. Z. M. Z., 2014. *kNN Algorithm with Data-Driven k Value.* Switzerland, Springer International Publishing.

Dongyang Dai, Z. W. R. L. X. W. J. J. H. M., 2019. *Learning Discriminative Features from Spectrograms Using Center Loss for Speech Emotion Recognition.* Brighton, IEEE.

Harsh Kukreja, B. N. S. C. S. K. S., 2016. AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK. *IJARIIE,* 1(5), pp. 27-30.

Kerkeni, L. S. Y. M. M. R. K. M. M., 2018. *Speech Emotion Recognition: Methods and Cases Study.* s.l., SCITEPRESS.

Kumar, R., 2011. *RESEARCH METHODOLOGY a step-by-step guide for beginners.* 3rd ed. London: SAGE.

Kun-Yi Huang, C.-H. W. Q.-B. H. M.-H. S. Y.-H. C., 2019. *Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds.* Brighton, IEEE.

M.S. Likitha, S. R. R. G. K. H. A. U. R., 2017. *Speech Based Human Emotion Recognition Using MFCC.* Bangalore, IEEE.

MASON, J., 2002. *Qualitative Researching.* 2nd ed. London, Thousand Oaks, New Delhi: SAGE Publications .

Moataz El Ayadi, M. S. K. F. K., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition,* XI(44), pp. 572-587.

Nilu Singh, R. A. K., 2015. *Digital Signal Processing for Speech Signals.* Lucknow, India, Bilingual International Conference on Information Technology.

Rajasekhar, A. & Hota, M. K., 2018. *A Study of Speech, Speaker and Emotion Recognition Using Mel Frequency Cepstrum Coefficients and Support Vector Machines.* Chennai, IEEE.

Saikat Basu, J. C. M. A., 2017. *Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture.* Kolkata, IEEE.

SAMUEL J. LING, J. S. W. M., 2016. *University Physics Volume 1.* s.l.:OpenStax.

SHICHAO ZHANG, X. L. M. Z. X. Z. D. C., 2017. Learningkfor kNN Classification. *ACM Transactions on Intelligent Systems and Technology,* 8(3).

Sreenivasa Rao Krothapalli, S. G. K., 2013. *Emotion Recognition.* New York: Springer.

Steven A. Rieger, R. M. R. P. R., 2014. *Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers.* Singapore, IEEE.

Appendix

Project Title: Emotion Detection from Voice

Student Name: Prakash Dahal

Student Number: 1828421

Supervisor Name: Krishna Aryal

 We confirm that the information given in this form is true, complete and accurate.

Student Signature: Prakash Dahal Date: 13 November 2018

Supervisor Signature: _____ Date: 30 November 2018

Thank you for completing this form. The MCS Ethics Committee will process the information provided and inform you of their decision shortly.

Log files with supervisor:

Frequently meeting was done with the supervisor but only in main 10 meetings log file is maintained. All ten logs are given below:

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

1

PROJECT MANAGEMENT LOG

First Name: Prakash Dahal **Surname:** ~~Krishna Aryal~~

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

Research on the topic : Saver App
All findings.

What do you aim to complete before the next meeting?

Detail project plan

Supervisor comments:

It is difficult to define critical or difficult situation. The idea has to be modified or changed.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2018/11/23

Supervisor Signature: _____ Date: 2018/11/23

Faculty of Science and Engineering
School of Mathematics and Computer Science

UNIVERSITY OF WOLVERHAMPTON

2

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Supervisor: Krishna Aryal

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Selection of New topic : Emotion Detection from Voice

Research & findings

What do you aim to complete before the next meeting?

Project planning

Supervisor comments:

1) - Selected topic is ok.
2) - Research on the topic is not sufficient.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2018/12/02

Supervisor Signature: _____ Date: 2018/12/02

Faculty of Science and Engineering
School of Mathematics and Computer Science

UNIVERSITY OF WOLVERHAMPTON

3

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Supervisor: Krishna Aryal

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Research for the classification problem.
Planning for the project development

What do you aim to complete before the next meeting?

Starting actual development.

Supervisor comments:

1) Try to find specific and related paper and solution. The classification of image and emotion from speech are different.
2) Research on Algorithm suitable for speech classification

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/1/13

Supervisor Signature: _____ Date: 2019/1/13

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

4

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

KNN algorithm is selected for speech classification.

Feature from speech (MFCC) is extracted. Extracted feature is in complex format and in exponential which is creating problem.

What do you aim to complete before the next meeting?

Train model using KNN algorithm.

Supervisor comments:

1) Try to clean data before saving it to csv file.
2) Convert complex number in readable format.
   Eg; (3,5) -> seperate real and imag number.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/2/18

Supervisor Signature: _____ Date: 2019/2/18

Faculty of Science and Engineering
School of Mathematics and Computer Science

UNIVERSITY OF WOLVERHAMPTON

5

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

From the MFCC features, mean value is selected and training is done.

What do you aim to complete before the next meeting?

I will try to use different other features to train.

Supervisor comments:

1) Try to increase your accuracy more than 70%.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/3/25

Supervisor Signature: _____ Date: 2019/3/25

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

6

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

Only Mean value is taken from MFCC.
13 MFCC are taken and each MFCC mean values are used for training.
~~Is this ok~~
Can you guide me further?

What do you aim to complete before the next meeting?

Complete AI part.

Supervisor comments:

1) Mean feature is ok
2) Try to include other features
3) Take some statistical data to increase your performance.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/3/29

Supervisor Signature: _____ Date: 2019/3/29

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

7

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal.

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

Mean and standard deviation is used. It has increased the accuracy to 76 %. ~~ok~~

What do you aim to complete before the next meeting?

Complete training and evaluation.

Supervisor comments:

1) Good progress of increasing accuracy.
2) Try to use other algorithms, apart from kNN so that you can compare.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/10

Supervisor Signature: _____ Date: 2019/4/10

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

8

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

Decision tree and SVM algorithm has been used. The accuracy is comparatively low.

What do you aim to complete before the next meeting?

Evaluation

Supervisor comments:

1) Less accuracy does not matter, try to compare different algorithms for your problem.
2) ANN can be better algorithm, so try to implement it.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/15

Supervisor Signature: _____ Date: 2019/4/15

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

9

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Supervisor:** Krishna Aryal

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

ANN has been used, the accuracy depends on the epoch, learning rate and other hyper-parameters.

Problem in integration of Model in Django.

What do you aim to complete before the next meeting?

Model integration in Django framework.

Supervisor comments:

1) Use RESTful Api to take data from your trained model.
2) Justify why Django is selected.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/18

Supervisor Signature: _____ Date: 2019/4/18

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

10

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Supervisor: Krishna Aryal

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Directly saved .sav model is used in Django. So, RESTful Api is not used.

What do you aim to complete before the next meeting?

Complete Development.

Supervisor comments:

1) If model is working fine, then its ok.
2) Make simple UI for your project.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/23

Supervisor Signature: _____ Date: 2019/4/23

Log files from Reader

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

1

PROJECT MANAGEMENT LOG

First Name: Prakash **Surname:** Dahal

Student Number: 1828421 **Reader:** Rupak Koirala.

Project Title: Emotion Detection from Voice **Month:**

What have you done since the last meeting?

Report started and completed:
- introduction
- Academic Question
- Literature Review

What do you aim to complete before the next meeting?

Artefact Development

Reader comments:

i) Compare other literature review with your's project.

ii) Try to include mathematical explaination (models)

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/12

Reader Signature: _____ Date: 2019/4/12

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

2

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Reader: Rupak Koirala

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Improved Literature Review
Artefact Development

What do you aim to complete before the next meeting?

Artefact development (improving)

Reader comments:

i) Focus on your testing and increase your testing.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/21

Reader Signature: _____ Date: 2019/4/21

Faculty of Science and Engineering
School of Mathematics and Computer Science

UNIVERSITY OF WOLVERHAMPTON

3

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Reader: Rupak Koirala

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Artefact Development

What do you aim to complete before the next meeting?

Tools and Technique

Reader comments:

i) Justify why you choose development method as evolutionary prototyping.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/4/25

Reader Signature: _____ Date: 2019/4/25

Faculty of Science and Engineering
School of Mathematics and Computer Science

UNIVERSITY OF WOLVERHAMPTON

4

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Reader: Rupak Koirala

Project Title: Emotion Dection from Voice Month:

What have you done since the last meeting?

Tools and Technique
Conclusion and Future Work
Answer of Academic Question

What do you aim to complete before the next meeting?

Critical Evaluation

Reader comments:

i) Justify why ~~tools~~ these tools are selected.

ii) Complete your report.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: _____ Date: 2019/5/1

Reader Signature: _____ Date: 2019/5/1

UNIVERSITY OF WOLVERHAMPTON

Faculty of Science and Engineering
School of Mathematics and Computer Science

5

PROJECT MANAGEMENT LOG

First Name: Prakash Surname: Dahal

Student Number: 1828421 Reader: Rupak Koirala

Project Title: Emotion Detection from Voice Month:

What have you done since the last meeting?

Tools and Technique justified
Critical Evaluation

What do you aim to complete before the next meeting?

Reader comments:

i) Overall work is ok

ii) Try to make your report rich.

We confirm that the information given in this form is true, complete and accurate.

Student Signature: prakash Date: 2019/5/2

Reader Signature: Date: 2019/5/2