FYP Proposal

Emotion Detection from Voice

By

Name: Dahal Prakash,

Student No: 1828421

Email: p.dahal2@wlv.ac.uk

Supervisor,

Mr. Krishna Aryal

Reader,

Mr. Rupak Koirala

Submitted in partial fulfillment

of the requirements for BIT, WLV

in the

Department of Computer Science

The Herald College

Date of submission: 19 December, 2018

Table of Contents

Abstract i
Acknowledgement ii
1. Introduction: 1
1.1. Overview 1
1.2. Aims and Objectives 3
1.3. Project Questions 3
2. Literature Review: 4
2.1. Feature Extraction: 4
2.2. Emotion Recognition using MFCCs and Hidden Markov Models (HMM): 5
2.3. Convolution Neural Network and Recurrent Neural Network: 5
2.4. Speech Recognition in android: 6
2.5. Comparison of MFCCs and LPCCs: 9
2.6. Semi-supervised speech emotion recognition: 9
3. Methodology: 10
3.1. Design: 10
3.2. Implementation 10
3.3. Testing: 11
3.4. Evaluation: 11
3.5. Proposed Solution 11
4. Project Plan: 13
4.1. Work Break Down Structure: 13
4.2. Gantt chart: 15
5. Required Hardware and Software: 16
5.1. Hardware: 16
5.2. Software: 16
References 17

Abstract

Communication with machine is common. These days, human speak and communicate with their phone, tablets and computer. So, emotion detection has been a field for research and development. This proposal gives the overall scenario of the project. The project will focus on the emotion of the speaker based on their voice. It detects the emotion of the speaker extracting important features of voice like MFCC, LPC, and energy and so on. This system will be developed using python and its library. For the part of training, dataset is provided by SAVEE database which will be trained using K-nearest neural network through tensor-flow and keras.

Acknowledgement

I would like to acknowledge for my Supervisor Mr. Krishna Aryal and reader Mr. Rupak Koirala for suggesting me and guiding me in the right track. I am happy and thankful for their support and guidance.

1. Introduction:

1.1. Overview

Interaction of human and computer is common these days. It is still evolving more for better interaction. Computers are able to understand and process Natural Languages. Now emotion detection has been interesting and emerging field for researchers and developers to make better communication. Human gives different emotions according to the situation which are represented from their speech, facial expressions, and gestures and so on. So in order to detect the human emotion through speech, voice pitch, energy, frequency etc. has to be analyzed. Though many research are going on in this field, it is still growing. (Chandni, et al., 2015).

This project is based on the human emotion recognition from voice where feature extraction is the primary work. Different voice parameters has to be extracted and analyzed like Mel-frequency cepstral coefficients (MFCCs), pitch, linear predictive analysis (LPC), etc. in order to detect emotion of a person. If the features are extracted properly then training for machine using neural network will give better result. Another important task is to choose machine learning algorithm. Mainly there are two types of machine learning; supervised and unsupervised. Supervised learning provides input and output both data while training and then it will be able to analyze other inputs according to the training but un-supervised learning is something different here, input are given but output are not provided so it has to be able to determine itself. Supervised learning are of two types; linear regression or prediction and other is classification. This project classifies the emotion and display the result (Akash Shaw, 2016) (Standford University, 2018).

In this project, for feature extraction and training of data python library, Tensor-flow will be used. Since Tensor-flow is open sourced and mainly developed for the machine learning purpose. For the training algorithm, K-nearest neural network has been selected for the classification of the emotion. KNN is simple and easy to implement therefore to complete the project on time, KNN is suitable. Similarly for the database, SAVEE Database will be used which has given its seven different

emotions like anger, fear, happiness, neutral, surprise, disgust and sadness separately with proper format of the data.

1.2. Aims and Objectives

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

Some objectives to fulfill these above aims are:

a. Understand the use of speech parameters.
b. Knowledge for the feature extraction of the given audio.
c. Analysis of KNN drawbacks and its improvement.

1.3. Project Questions

Some project questions to clarify the project and to obtain the better output.

a) How to extract the required audio features?
b) Which machine learning algorithm is feasible and better performing?
c) Is there any way to improve the result?
d) How to integrate trained model in the system?

2. Literature Review:

2.1. Feature Extraction:

Emotion detection a very interesting topic so the author has classify emotion through speech using Artificial Neural Network. Main features to be used for the detection of the emotions are Prosodic features which contains pitch, energy formant frequencies and Spectral features which are like mel frequency cepstral coefficient. These features has been focused by the author to detect emotion.

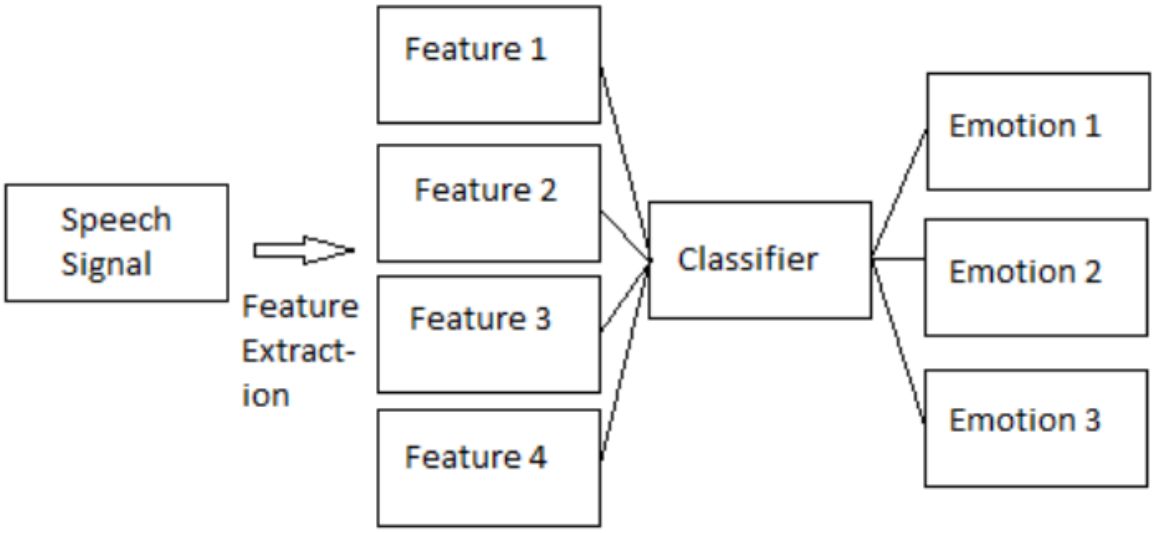

Figure 1: Block Diagram of input flow, process and output

The diagram represents the working mechanism. 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 classified output is obtained (Akash Shaw, 2016).

2.2. Emotion Recognition using MFCCs and Hidden Markov Models (HMM):

The investigation of recognizing emotion from speech signal using Hidden Markov Models. It worked on phonetic features from speech especially on MFCCs to improve accuracy with less feature set. The speech emotional utterances was taken from the SAVEE emotional corpus. Here after training data with the features extracted using HMM was tested and the trial average accuracy was 78% and the highest accuracy was 91.25%.

It gives four different emotions outputs; sadness, fear, disgust and surprise.

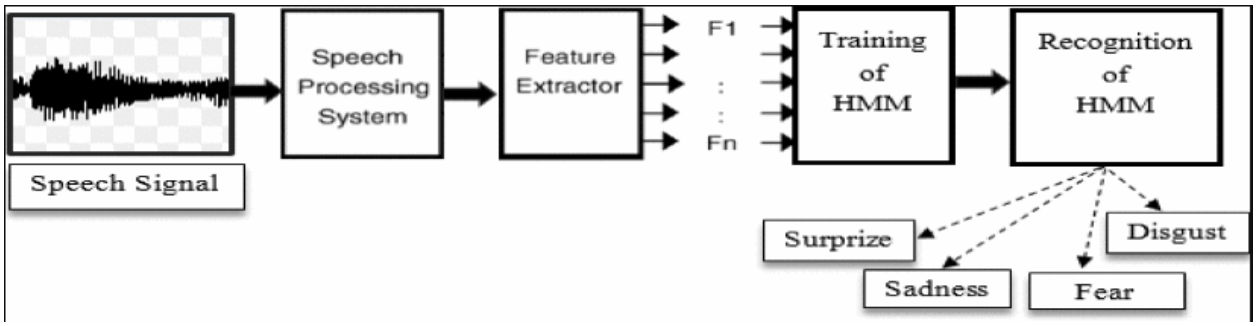

Figure 2: Block Diagram for emotion recognition

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 recognize the class of emotions and the output is given (Chandni, et al., 2015).

2.3. Convolution Neural Network and Recurrent Neural Network:

Challenge of recognizing emotion from speech signal was chosen by the author using corpora speech database. Classification of emotion is done using Convolution neural network and recurrent neural network. Author used 13 MFCCs, 13 velocity and 13 acceleration components as feature.

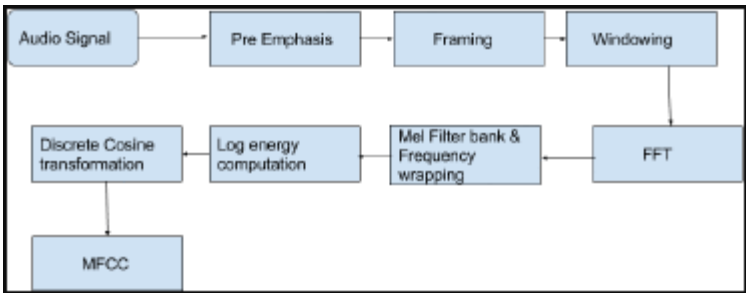

Figure 3: MFCC block diagram

Since MFCCs is taken as a very important speech feature for the emotion recognition. Author has extracted MFCC passing from different Pre Emphasis, Framing, Windowing, Fast Fourier Transform (FFT), and so on to get the proper MFCC. Thus, the output obtained could be more appropriate. Approximately 80% of accuracy was obtained when testing data (Saikat Basu, 2017).

2.4. Speech Recognition in android:

(Dhankar, 13-16 Sept. 2017) Android also supports different external voice recognition APIs. CMU Sphinx, Kaldi, Hidden Markov Model (HTK), etc. are some speech recognition toolkits.

CMU Sphinx is simply called Sphinx. Sphinx is an open source which supports Java, Python and C languages. It also provides offline services. Sphinx uses two different language models; grammar and n-gram. Grammar specification approach is used in command and control where limited vocabulary is enough but N-gram is used for broader range of vocabulary. In both language models, an acoustic model and a dictionary is required. The research paper had used Acoustic-Phonetic Approach, Pattern Recognition Approach and Deep Learning Approach to recognize speech.

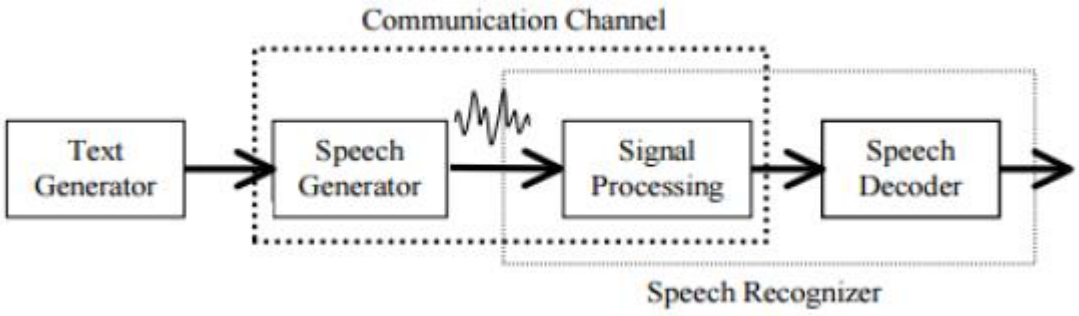

Figure 4: Speech Generation and Recognition Process

Text Generator is brain which generates text and speech generator is the vocal tract. Signal Processing processes the obtained speech and the speech is decoded by speech decoder. Now, the voice pattern is tested form the dictionary. Deep learning makes the voice learning more accurate.

Kaldi is also open source API which is written in C++ programming language to build Language Model (LMs) and Acoustic Models (AMs). It is one of the most popular toolkit which produces high-quality lattices and sufficiently fast for real-time recognition.

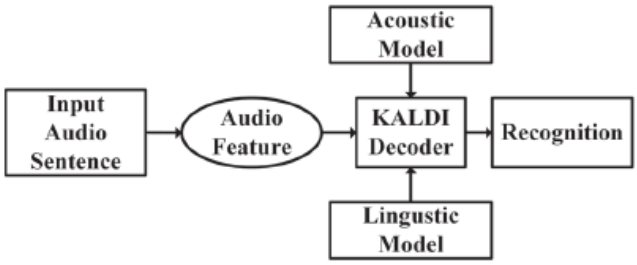

Figure 5: Automatic speech recognition model (Prashant Upadhyaya, 22-24 March 2017).

Similarly, Google has also provided Speech to Text or Text to speech API. Text to speech includes two sub-system; Text to Phoneme Converter Part and Phoneme to Speech Converter Part.

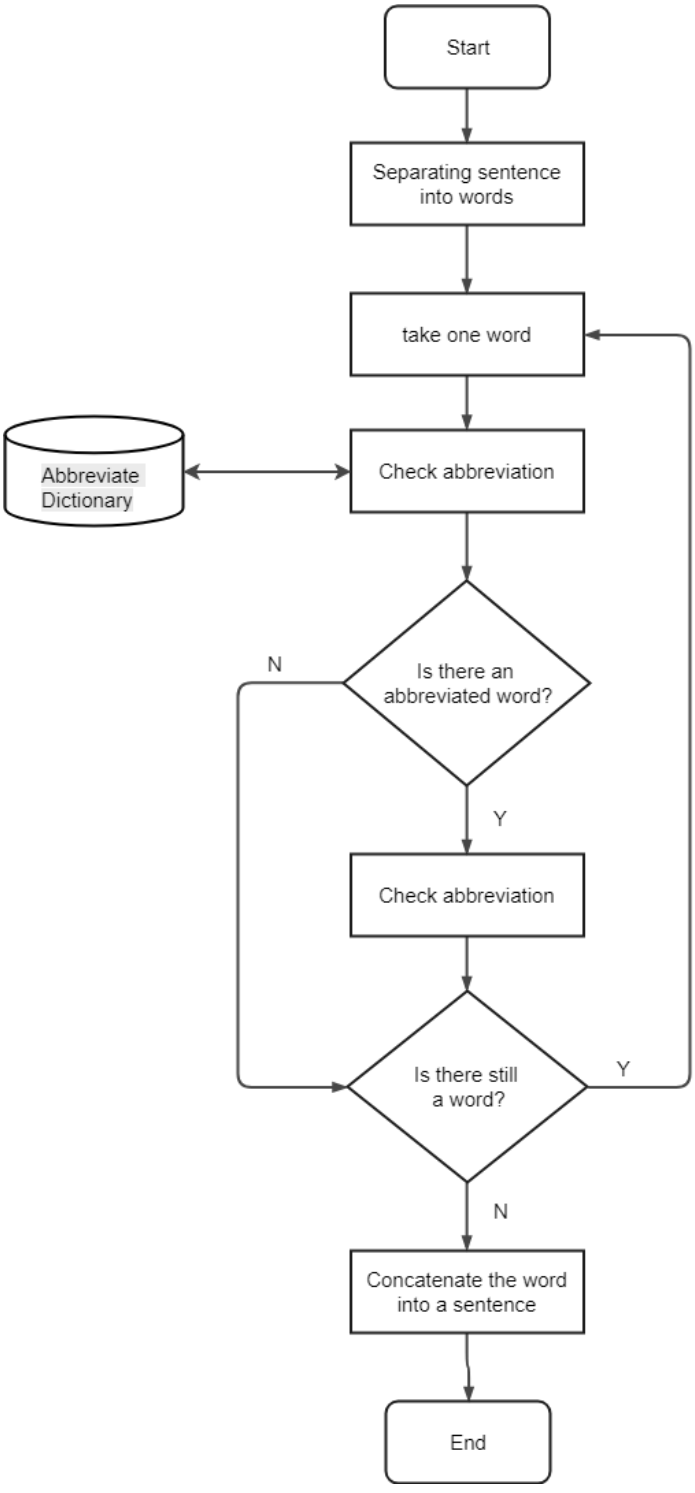

Figure 6: Text Normalization Flowchart ***(Intan Sari Areni, 7-8 Oct. 2017)***

2.5. Comparison of MFCCs and LPCCs:

This paper is comparative study for the evaluation of Linear Predication Cepstral Coefficient (LPCCs) and Mel Frequency Cepstral Coefficients (MFCCs). Data to train the model was used from SUSE database. Only two features MFCCs and LPCCs was used to train the data. After training, when the testing was done, LPCCs gave better result in average so LPCCs is better acoustic feature in comparison to MFCCs for emotion classification.

Emotion	MFCC Approach		LPCC Approach	
	EER	Accuracy	EER	Accuracy
Anger	20.99	79.01	23.67	76.33
Lombard	14.61	85.39	12.61	87.39
Neutral	18.67	81.33	18.67	81.33
Question	27.68	72.32	10.68	89.32
Average	20.48	79.52	16.40	83.60

Figure 7: Performance result using full dimension.

In average, LPCCs approach has obtained better accuracy with minimum error (Sudhakar Kumar, 2014).

2.6. Semi-supervised speech emotion recognition:

Supervised learning method has spread widely for emotion recognition through speech where the developer is in the limitation by the labelled speech data. So the purpose of the author is to improve speech emotion recognition using semi-supervised method. The aim of this journal is to reap the benefits combining labelled and unlabeled data. In future for more the better result or for the improvement of emotion recognition from voice, semi-supervised Recurrent Neural Network can be used (Jun Deng, 2018).

3. Methodology:

3.1. Design:

There are different methodologies for the development of the system. Among these methodologies; Agile, Joint Application Development (JAD), Rapid Application development (RAD), Prototyping, Incremental are some popular methodologies. Out of these methodologies; prototyping and or incremental could be suitable for this type of project, since other are suitable and applicable for group of people working together. For this project, Evolutionary prototyping methodology would be suitable. Prototyping methodology are of four types; Throw-away prototyping, evolutionary prototyping, incremental prototyping and extreme prototyping. Throw-away prototype is thrown after meeting the requirements and again new system is created to fulfill the obtained requirement which is time and effort consuming. In Incremental prototyping, multiple module is created completely and merged together which creates problem in integration. Extreme prototyping is especially suitable for website project. So, finally incremental prototyping is suitable for this project because the performance or result should be analyzed time and again if there room for improvement for better performance considering time. So, improvement is necessary for this project. Therefore, evolutionary prototyping following the software development life cycle is the most suitable and applicable methodology (Tutorials point, 2018).

3.2. Implementation:

a. Planning:

   Planning is an important phase of the project where scope is defined and according to it, schedule, methods, and tasks will be planned properly in order to launch the project successfully. Overall planning for the project is made.

b. Requirement Engineering:

In this phase, the real requirements which are necessary for the project is gathered. In this project, necessary things will be gathered for the completion of the project.

c. Designing:
   Some designing of the system will be represented. UML Diagram for the representation of the overall interaction by the user will be represented.

d. Developing:
   Actual coding takes place in this stage. In this project, actual development is done on python and then in android.

e. Testing:
   Testing includes black box and white box testing. In this phase, actual code testing is done for the bug and error. So white box testing is carried out more than black box.

3.3. Testing:

Here in this stage, overall project is tested. There are two types of testing; black box and white box. So both testing will be carried out on this project. Generally Black box testing will be done at the end for the validation.

3.4. Evaluation:

Evaluation is the final stage where the product is evaluated, whether it has meet the aim and objective of the project or not.

3.5. Proposed Solution:

The project focus on the feature extraction from the speech and the detection of emotion.

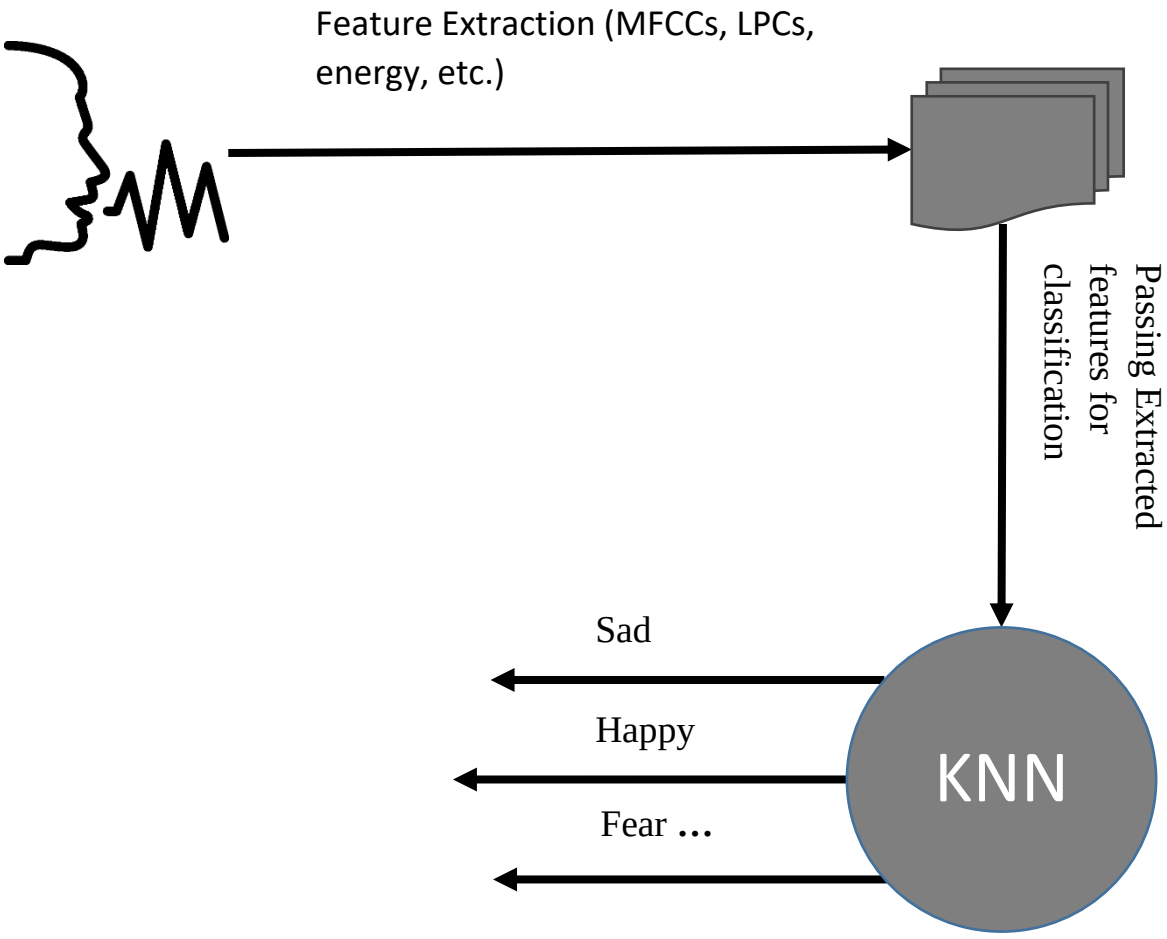

Figure 8: Working mechanism of proposed solution

Above flow diagram represents the overall training of the emotion. From the dataset or speaker, it extracts the different required features like MFCC, LPC, and energy and so on. The extracted features are passed through KNN and it analyze the features and as per the features passed, it directs to different classes of emotion like sad, happy, fear, etc.

4. Project Plan:

4.1. Work Break Down Structure:

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) Prepare Proposal 1.1.4) Resource Planning
	1.2) Requirement Engineering	1.2.1) Requirements Gathering 1.2.2) Requirement 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 1.4.3) Product development the system
	1.5) Testing	1.5.1) Black Box Testing 1.5.2) White Box Testing
	1.6) Deploying	1.6.1) Submission of the project 1.6.2) Evaluation of Final Report

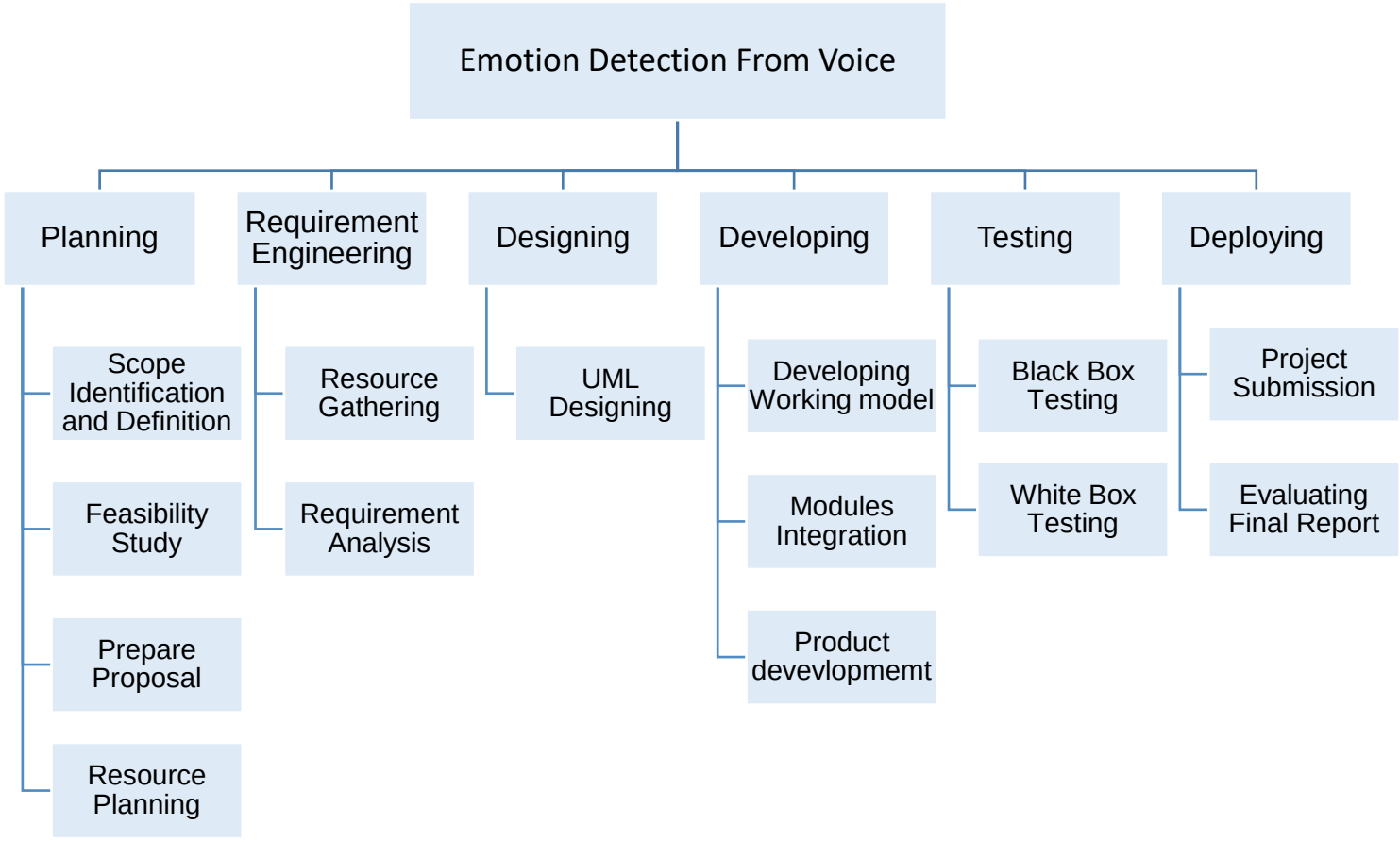

Figure 9: Tree diagram of Work break down structure

4.2. Gantt chart:

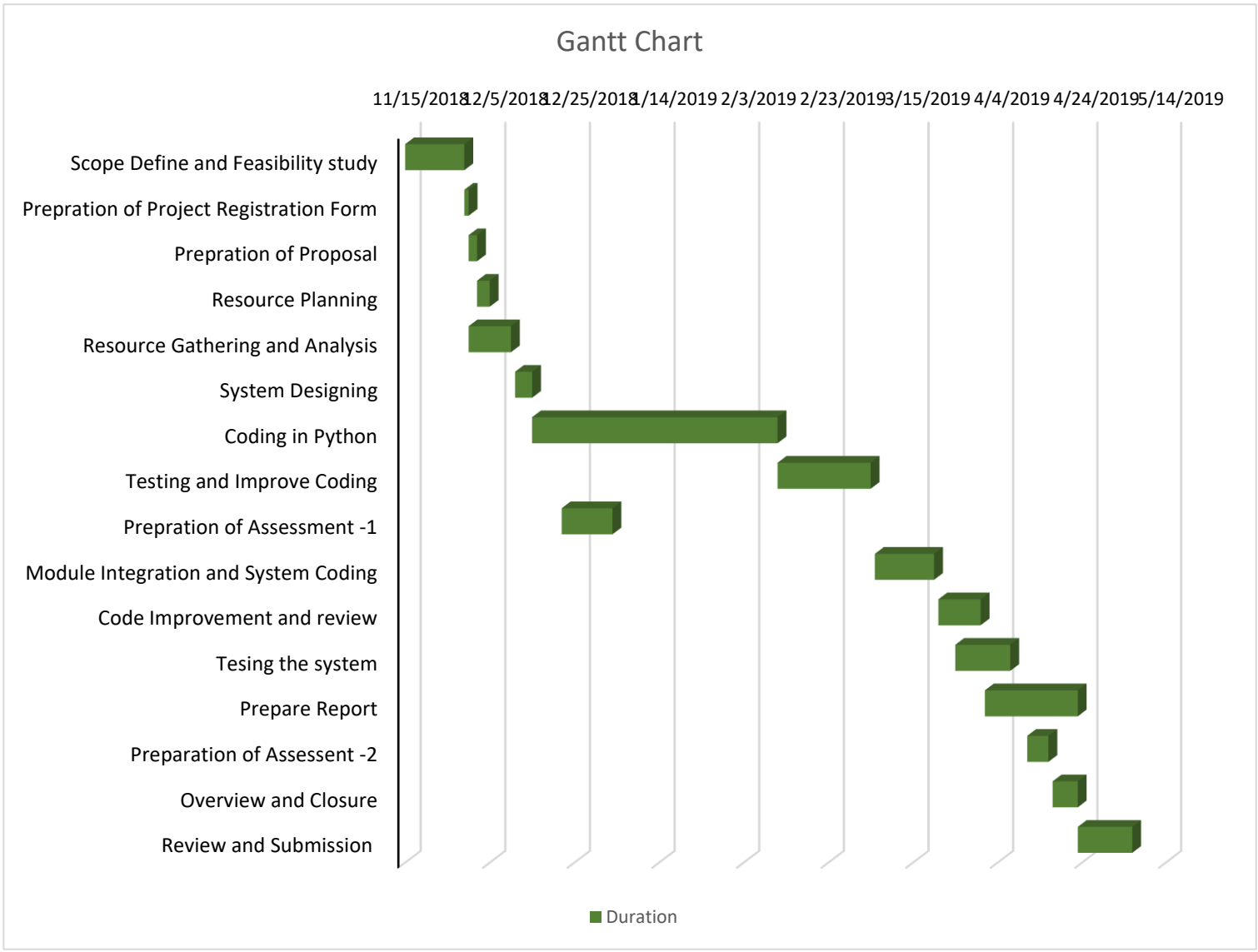

Figure 10: Gantt chart

5. Required Hardware and Software:

5.1. Hardware:

a. Computer:

A computer is required which supports for the training of data in machine learning. A Laptop could be portable and easier.

5.2. Software:

a) Tensor flow

Tensor flow is an open source software library which is used for the machine learning provided by google. So this can be used to use machine learning (Tensorflow, n.d.).

b) Django

Django is a free and open source web framework based on python. Django software is useful for the development of web app.

c) REST software

REST should be used to bring the module build in tensor-flow to Django.

d) Microsoft office

Microsoft is office is required for documentation of the report of the project and to maintain other required file.

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.

Chandni, et al., 2015. *An automatic emotion recognizer using MFCCs and Hidden Markov Models.* Brno, IEEE.

Dhankar, A., 13-16 Sept. 2017. *Study of deep learning and CMU sphinx in automatic speech recognition.* Udupi, India, IEEE.

Intan Sari Areni, S. W. I. ,. A., 7-8 Oct. 2017. *Solution to abbreviated words in text messaging for personal assistant application.* Semarang, Indonesia, IEEE.

Jun Deng, X. X. Z. Z. S. F. B. S., 2018. Semisupervised Autoencoders for Speech Emotion Recognition. *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,* 26(1), pp. 31 - 43.

Prashant Upadhyaya, O. F. M. R. A. Y. V. V., 22-24 March 2017. *Continuous hindi speech recognition model based on Kaldi ASR toolkit.* Chennai, India, IEEE.

Saikat Basu, J. C. M. A., 2017. *Emotion recognition from speech using convolutional neural network with recurrent neural network architecture.* Coimbatore, IEEE.

Standford University, 2018. *Coursera.* [Online]
Available at: www.coursera.org
[Accessed 1 12 2018].

Sudhakar Kumar, T. K. D. R. H. L., 2014. *Significance of acoustic features for designing an emotion classification system.* Dhaka, IEEE.

Tensorflow, n.d. *Tensorflow.* [Online]
Available at: https://www.tensorflow.org/
[Accessed 01 December 2018].

Tutorials point, 2018. *SDLC - Software Prototype Model.* [Online]
Available at: https://www.tutorialspoint.com/sdlc/sdlc_software_prototyping.htm
[Accessed 18 May 2018].