Speech, Age and Gender Recognition using Deep Neural Networks

(Hey Root)

By

GC SARMAD RABBANI

GC M NOUMAN ASLAM

GC AFAQ AHMAD

Supervised by:

Dr Hammad Afzal

Submitted to the Department of Computer Software Engineering,

Military College of Signals, National University of Sciences and Technology, Islamabad,

in partial fulfillment for the requirements of B.E Degree in Software Engineering.

June 2022

In the name of ALLAH, the Most benevolent, the Most Courteous

CERTIFICATE OF CORRECTNESS AND APPROVAL

This is to officially state that the thesis work contained in this report

“Speech, Age and Gender recognition using Deep Neural Networks**”**

is carried out by

GC SARMAD RABBANI

GC M NOUMAN ASLAM

GC AFAQ AHMAD

under my supervision and that in my judgement, it is fully ample, in scope and excellence, for the degree of Bachelor of Software Engineering in Military College of Signals, National University of Sciences and Technology (NUST), Islamabad.

Approved by

Supervisor

Dr Hammad Afzal

Date: ___________

DECLARATION OF ORIGINALITY

We hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification in either this institute or anywhere else.

ACKNOWLEDGEMENTS

Allah Subhan'Wa'Tala is the sole guidance in all domains.

Our parents, colleagues and most of all supervisor, **Dr Hammad Afzal** without your guidance.

The group members, who through all adversities worked steadfastly.

Plagiarism Certificate (Turnitin Report)

This thesis has ____ similarity index. Turnitin report endorsed by Supervisor is attached.

____

GC SARMAD RABBANI

00000278703

____

GC M NOUMAN ASLAM

000000278721

____

GC AFAQ AHMAD

000000278718

____

____

Signature of Supervisor

ABSTRACT

Ample research has been done on language-related applications, especially the use of machine learning for speech recognition. However, recent research has focused on the use of deep learning in voice-based applications. This new machine learning research has become a very interesting field of study, with far superior results compared to other research, depending on the variety of applications. The proposed project uses machine learning and the concept of deep neural networks to score spoken language from a one-second audio file. The proposed project also aims to calculate age and gender using various files as input to the system. A secondary function of the project is to execute voice activation commands from specific speakers. The target audience for this project overview is college students, mostly bachelor's degrees, and may also serve as an open source project on Github. Commercial advertising is more relevant and can target specific age and gender groups, thus increasing sales. Forensic medicine can reduce suspects if there is evidence such as a phone call. Age and gender classification is especially useful in a variety of real-world applications such as security and video monitoring, electronic customer relationship management, biometrics, electronic vending machines, human-computer interaction, entertainment, cosmetics, and forensic arts. The main features of the project and expected features are:

- Speech-to-Text from audio file
- Speech-to-Text from real-time audio
- Gender, Age estimation
- Speaker estimation

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Speech is a multifaceted object, and its creation involves many structural movements that affect the quality and voice characteristics of the speech. Language is an important and easy source of communication. In addition to language knowledge, this also includes speaker-related paralanguage data such as speaker identity, mood, health status, age, and gender. Automatically extracting a system of this information by voice is very useful in many programs, such as: B. Personal identification in the banking system. Customer care applications such as call centers. Voice bot; collaborative and intelligent voice assistant. There are already international and domestic companies in the industry that provide voice processing services such as Google, Amazon and Techmo in the polishing market. Extracting information about a speaker's age and gender can be used in an Interactive Voice Response (IVR) system to redirect or play the speaker to the appropriate coordinator for a particular gender / background music. The Voicebots program allows you to change the behavior of your bot using paralanguage information extraction. For voice assistants, you can use this information to identify relevant ads and select search results that are more appropriate for a particular age / gender group. Overall, the use of paralanguage content can lead to a better user experience, which can bring revenue to companies that decide to use such a system. Speech or speech-to-text recognition, mechanical capabilities or systems that recognize spoken words aloud and convert them into readable text. Rudimentary speech recognition software has limited speech and can only recognize words and phrases when spoken clearly. Complex software can handle native speeches, different voices, and different languages. Speech recognition uses extensive research in computer science, language, and computer engineering. Many modern devices and text-based programs have voice recognition capabilities that make the device easy or hands-free to use**.**Speech recognition and voice recognition are two different technologies and should not be confused with:

- **Speech recognition**: used to identify words in spoken language.
- **Voice recognition**: is a biometric technology used to identify human voices**.**

Speech recognition programs use computer algorithms to process, translate, and translate spoken language into text. The software program converts the voice recording into a computer- and human-understandable microphone**.** following these four steps:

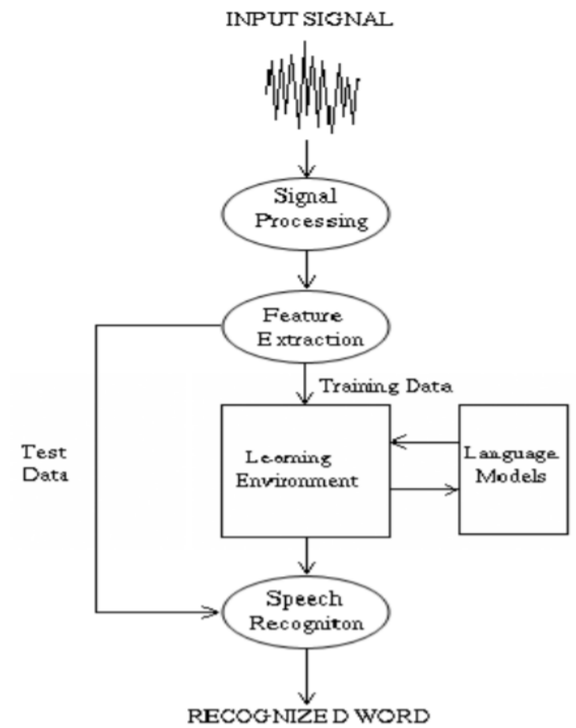
1. Analyze the audio;
2. Divide it into parts;
3. Digitize to a computer-readable format and
4. Use the algorithm to match the most appropriate textual representation.

Speech recognition software needs to adapt to the most flexible environment and the specific context of human speech. Software algorithms process speech into trained text using a variety of speech patterns, speech styles, languages, dialects, pronunciations, and sentences. This software distinguishes between audio and background noise. Background noise is usually accompanied by a signal. The speech recognition system uses two models:

- **Acoustic models.** These represent the relationship between linguistic units and audio signals
- **Language models.** Here, sounds are compared to word sequences to distinguish words with similar pronunciations.

1.1 Overview

An easy-to-use visual interface is the required result of a project, accurate text input, fast output, accurate measurement of gender and age, the project should listen and can work again on Android. To meet our needs, our team will work to create an interesting and easy-to-use interface using machine learning model and deep neural networks to get accurate and fast results. The model will work best on desktop systems and the Android model will be built after the project is completed in the desktop area. The functionality of the ASR system is shown in the figure:



The ML model will use a split algorithm and consideration of language models will be BERT, GPT2, XLNet, RoBERTa. Similarly the project will use deep neural networks. Some of the DNNs are Learning Vector Quantization, WordtoVec, Dynamic Time Wrapping and Artificial Neural Networks. The low noise area is suitable for the project and the latest hardware specifications for Android Studio are well suited to test and implement the project in the Android environment.

1.2 Problem Statement

The research focuses on using in-depth learning of speech-related applications in the last few years. Exploring the age of the platform using real-time speech has also become a hotbed of technological attraction due to increased market demand and increased security and other concerns. The recorded acoustic signal is an analogue signal and the analogue signal cannot transmit directly to ASR systems. These audio signals must be converted to digital signals before processing. These digital signals are sent to the primary filter to down convert the signal. This process improves signal strength at high frequencies. This is the first processing step. The output step detects a set of voice parameters with acoustic communication and speech signals and these parameters are calculated by processing the acoustic waveform.

1.3 Proposed Solution

Acoustic modeling is a fundamental part of the ASR system. In the acoustic model, the relationship between acoustic and phonetic calculations is established. The Acoustic version plays an important role in device performance and is responsible for computer computing. Linguistic translation incorporates the structural boundaries found within a language to create opportunities for that to happen. It creates the possibility of a sentence occurring after a sequence of words. Pattern separation (or thunder) is a process of comparing an unknown test pattern with each sample of sound quality reference and computer to calculate the degree of similarity between them. After completing the machine training during the exit, the pattern is separated to hold the speech.

1.4 Working Principle

1. Development of a smart and intelligent speech, gender and age recognition system
2. To implement Machine Learning techniques and simulate the results
3. To increase productivity by working in a team
4. To design a project that contributes to the welfare of society

1.4.1 Datasets and annotations:

Data is acquired from many datasets available on the internet. Following are a few datasets:

I. The People's Speech
II. LibriSpeech
III. OpenSLR
IV. Google Voice dataset
V. Mozilla Corpus

1.4.2 Dataset training and processing:

The prepared dataset is used as input to train object detection models using machine learning.

After the data is collected the following steps are performed on the data:

I. Pre-processing
II. Feature extraction
III. Classification
IV. Language modeling.

The preprocessing step aims to improve the audio signal by reducing the signal-to-noise ratio, reducing noise, and filtering the signal.

Generally, the features used for ASR are extracted using a specific number of values or coefficients generated by applying different methods to the input. This step needs to be robust against various quality factors such as noise and echo effects.

The majority of the ASR methods adopt the following feature extraction techniques:

I. Mel-frequency cepstral coefficients (MFCCs)
II. Discrete Wavelet Transform (DWT).

The classification model aims to find the voice text contained in the input signal. Gets the features extracted from the preprocessing step and produces the output text. The Language Model (LM) is an important module for capturing language grammatical rules or semantic information. The language model is important for recognizing the output tokens from the classification model and modifying the output text.

1.4.3 Data Dictionary

Following are the data dictionaries:

I. Audio data from datasets in wav format
II. Transcripts of audio in txt format
III. Metadata about dataset in json format

1.4.4 Integration:

The different modules is then integrated in to one stand-alone entity. This stand-alone entity is essential for a compact solution.

1.4.5 GUI presentation:

The visual demonstration of the project is done through the aid of GUI (graphical user interface).

1.5 Objectives

1.5.1 General Objectives:

1. Development of a smart and intelligent speech, gender and age recognition system
2. To implement Machine Learning techniques and simulate the results
3. To increase productivity by working in a team
4. To design a project that contributes to the welfare of society

1.5.2 Academic Objectives:

To explore the utilization of machine learning in Speech recognition and Image processing, learn the innovation in machine learning concepts and a practical implementation of the project using programming knowledge

1.6 Scope

Speech/Speaker and age recognition have many practical applications, among them are:

1. Commercial advertising. Your ads will be more relevant and you will be able to target specific age and gender groups, which will increase your sales.
2. Another application is forensics. If you have evidence such as a phone call, you can reduce the number of suspects.
3. Valuable biometric technology and this procedure apply to multiple areas such as secure access to secure areas, machines such as voice dialing, banks, databases and computers.

1.7 Deliverables

Sr.	Tasks	Deliverables
1	Literature Review	Literature Survey
2	Requirements Specification	Software Requirements Specification document (SRS)
3	Detailed Design	Software Design Specification document (SDS)
4	Implementation	Project demonstration
5	Testing	Evaluation plan and test document
6	Training	Deployment plan
7	Deployment	Complete application with necessary documentation

1.8 Relevant Sustainable Development Goals

The Locally relevant socio-economic issue that the project addresses is SDG#9 INDUSTRY, INNOVATION AND INFRASTRUCTURE, which is focused on developing resilient infrastucture, promoting inclusive and sustainable industralization, and foster innovation.

1.9 Structure of Thesis

- Chapter 2 contains the literature review and the background and analysis study this thesis is based upon.
- Chapter 3 contains the design and development of the project.
- Chapter 4 introduces detailed evaluation and analysis of the code.
- Chapter 5 contains the conclusion of the project.
- Chapter 6 highlights the future work needed to be done for the commercialization of this project.

CHAPTER 2: LITERATURE REVIEW

A new product is launched by modifying and enhancing the features of previously launched similar products. Literature review is an important step for development of an idea to a new product. Likewise, for the development of a product, and for its replacement, related to speech to text, gender and age recognition, a detailed study regarding all similar projects is compulsory. Our research is divided into the following points.

- Existing applications and their drawbacks
- Delivering an application which can be run in Desktop and possibly in Android systems.
- Research Papers

2.1. Deep learning approaches

In-depth learning uses multiple layers of neural systems to capture different exposures from obscure information. The most popular learning models include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), especially Long Term Memory (LSTM). Generally, CNNs are useful for learning local examples over a database while RNNs or LSTMs read sequential examples. Badjatiyaetal has tried different things with the use of three in-depth neural programs, CNN, LSTM and fast content so each included unique and GloVE inserts. CNN has done better than LSTM which has done better than Fast Content. However, combining Gradient Boosted Decision Trees (GBDTs) with LSTM installed with unconventional embedding has produced the best results. Gambck and Sikdar used the convolutional neural system to organize a disgusting Twitter discourse on the data generated by Waseem and Hovy. They explored different approaches to the four CNN models; the first is fixed at random word vectors, the second is prepared with word2vec, the third is prepared with letters and grams and the fourth is prepared with a combination of letters and grams and word vectors. The modified version of word2vec provided the best execution. The use of derogatory language is very small in all facts. Methods of determining information in most research texts are biased towards certain categories of abominable discourses, for example, racism, nationalism, religious gatherings and so on when data collection is collected using many special derogatory terms. Such information does not constitute an actual transfer of offensive language. In addition, fully integrated learning strategies rely on hand-crafted commentary which is a costly process and is therefore insufficient to provide a comprehensive combination of humorous proportions in different proportions. Gao et al. to understand the shortcomings of the controlled reading method and the manual manipulation of information, proposed a two-dimensional approach to bootstrapping is two-component, a missing term reader and an LSTM separator for reading both offensive and understandable speech. The model is also compared to LSTM who is just a term student and was found to have passed the last two frames. Pitsilis et al. has proposed a collection of LSTM dividers to view the content of passionate and disrespectful women on Twitter. They investigated key points associated with the client's tendency to abusive speech by violating customer behavior based on tweet history. Customer-based highlighting fuse has improved the anticipation system and the display of dividers. Representing both local and sequential data in the literature, Zhang et al. introduced another in-depth neural system that used CNN integrated with the crushed RNN. In addition to the fact that they are testing their model only on a religious twitter corpus and refugees hate it, yet they have done a similar wide-ranging survey about freely accessible Twitter companies and won shows created by placing another benchmark on 6 out of 7 databases. In addition to the English language, in-depth learning models are also used in important tests of different dialects, for example, German and Arabic. Mitrovic and Handschuh have developed a framework for identifying hostile language in German tweets. They made three different models; n-gram model, word-component integration model and CNN and RNN affiliate model. They examined the models in three different databases. The method of collecting Word vectors performs better than others in two of the three databases while the n-gram method overrides the other in one of the three data sets. Albadi et al. make the first collection of information for the recognition of religious contempt speech. Think of dictionary-based, n-gram-based strategies and in-depth reading-based strategies in which RNN-based duplicate units come out in different ways.

CHAPTER 3. INTEFACING AND DETACTION

3.1. GUI of Modules

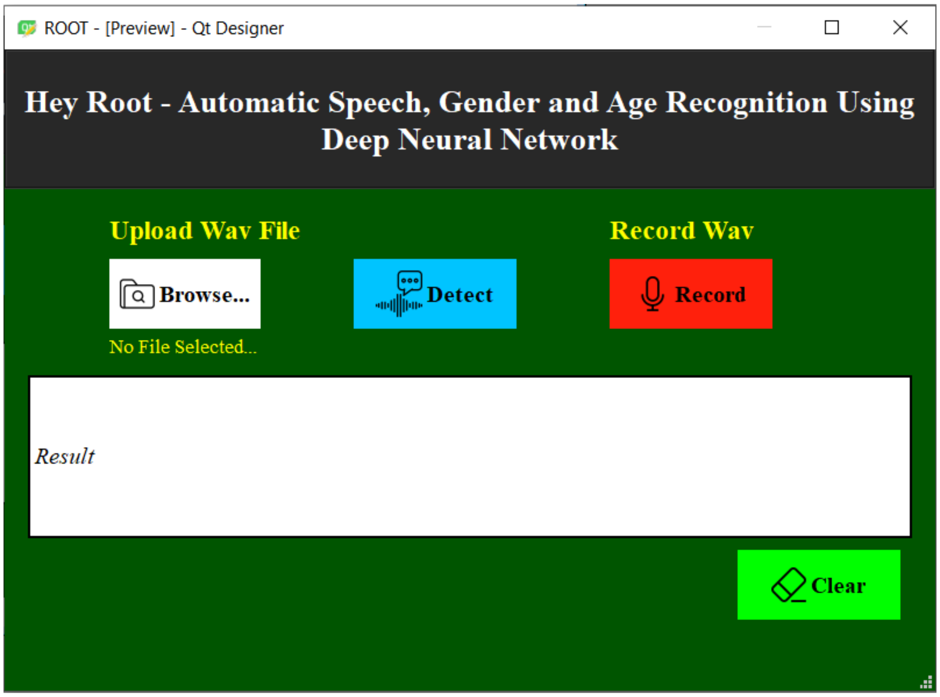


Figure 3.1.1 – User Interface (Screenshot 1)

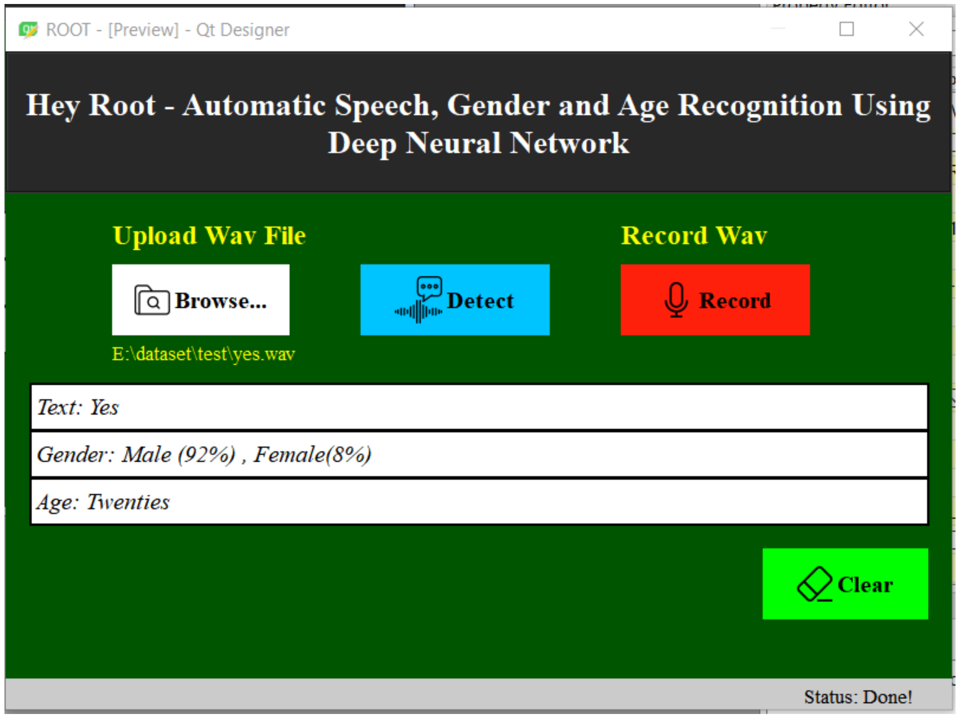


Figure 3.1.2 – User Interface (Screenshot 2)

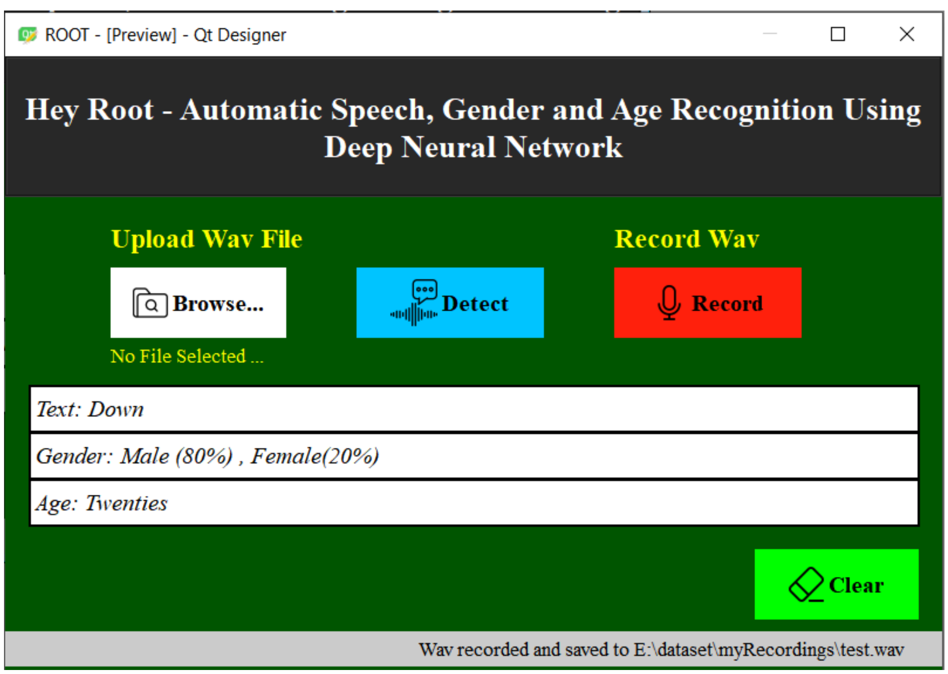


Figure 3.1.3 – User Interface (Screenshot 3)

.

3.2. Process Diagram

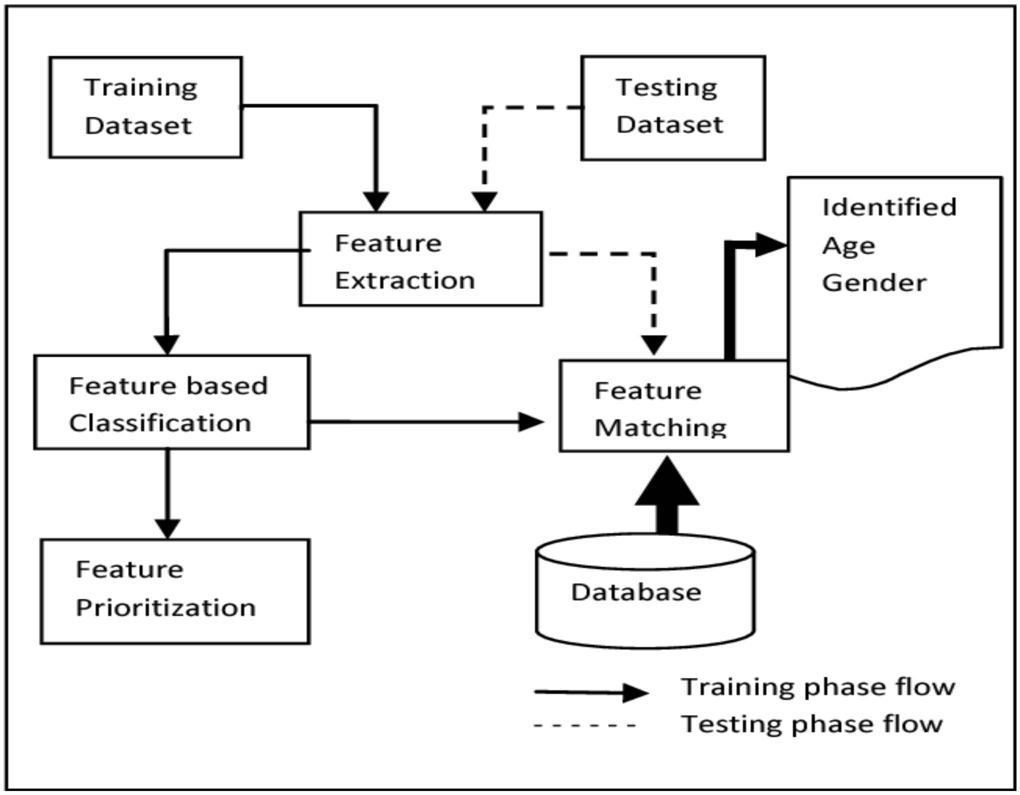


Figure 3.2. – Process Diagram

3.3. Block Diagrams

3.3.1 System Block Diagram

This area includes special expressions of expression, age and gender identity. Demonstrates the use of application in the context of various visual purposes and further demonstrates the interaction between the various components.

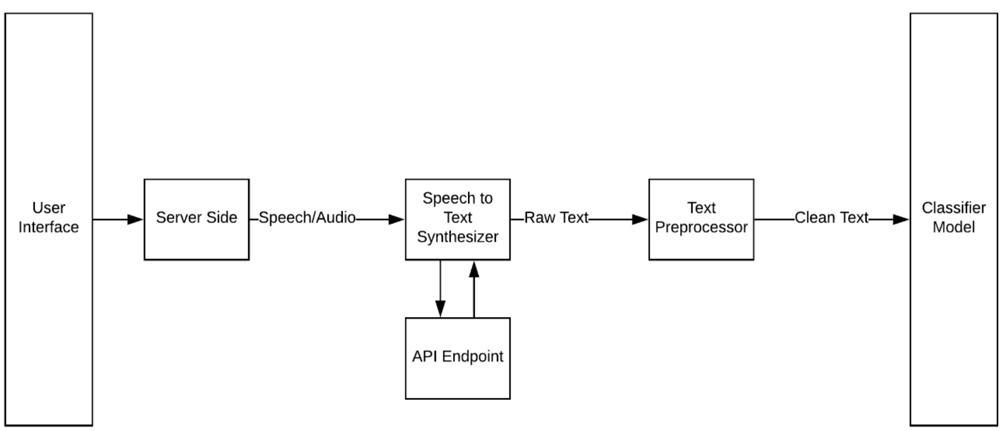


Figure 3.1.1. - System Block Diagram

3.3.2 Activity Diagram

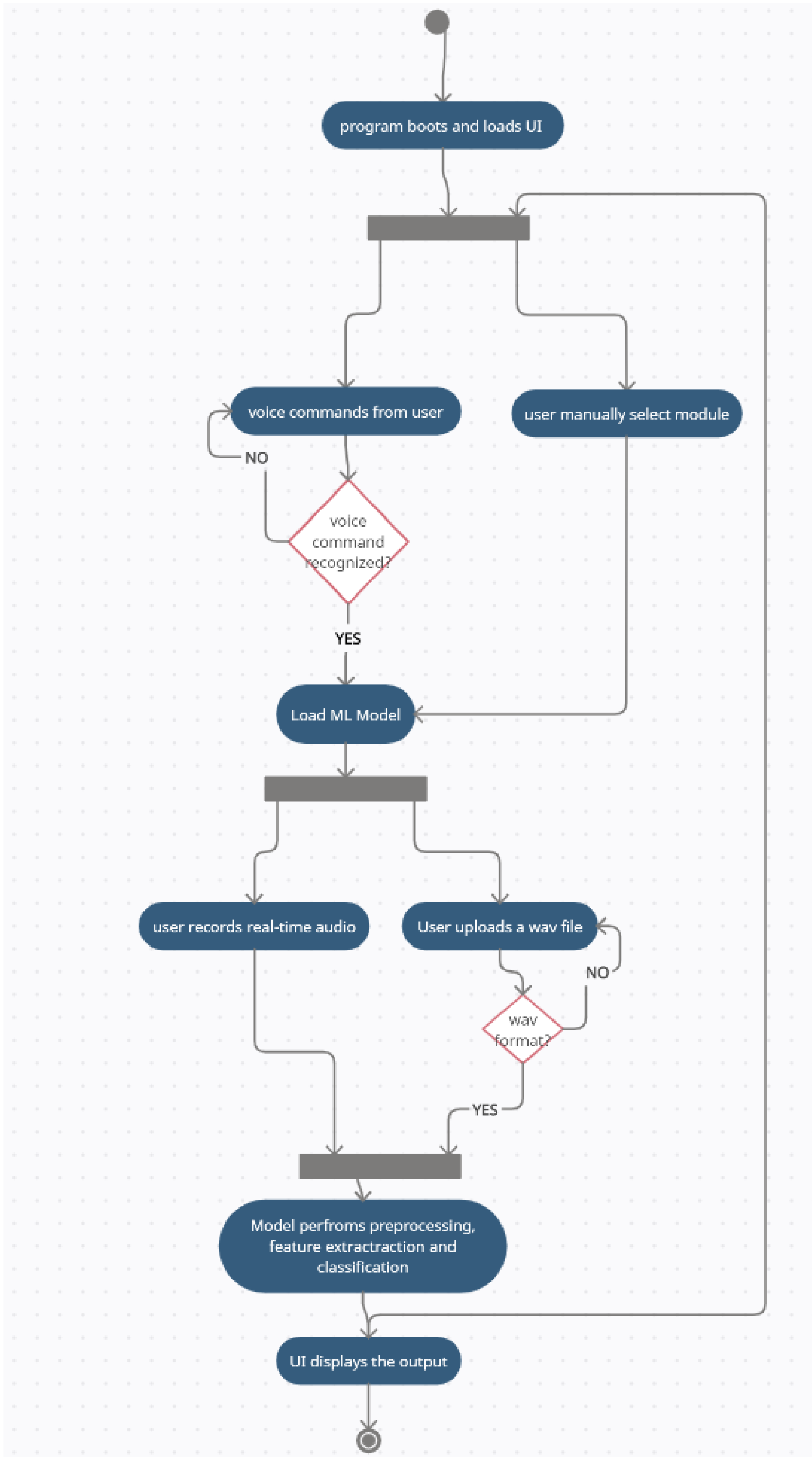


Figure 3.3.2 – Activity Diagram

3.3.3 Class Diagram

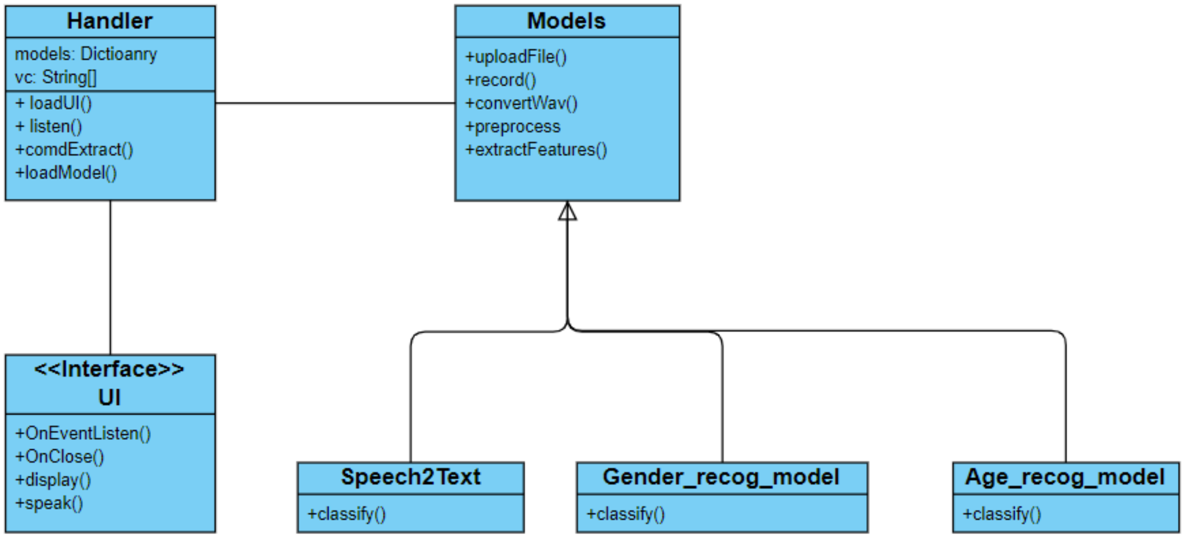


Figure 3.3.3 – Class Diagram

3.3.4. Sequence Diagram

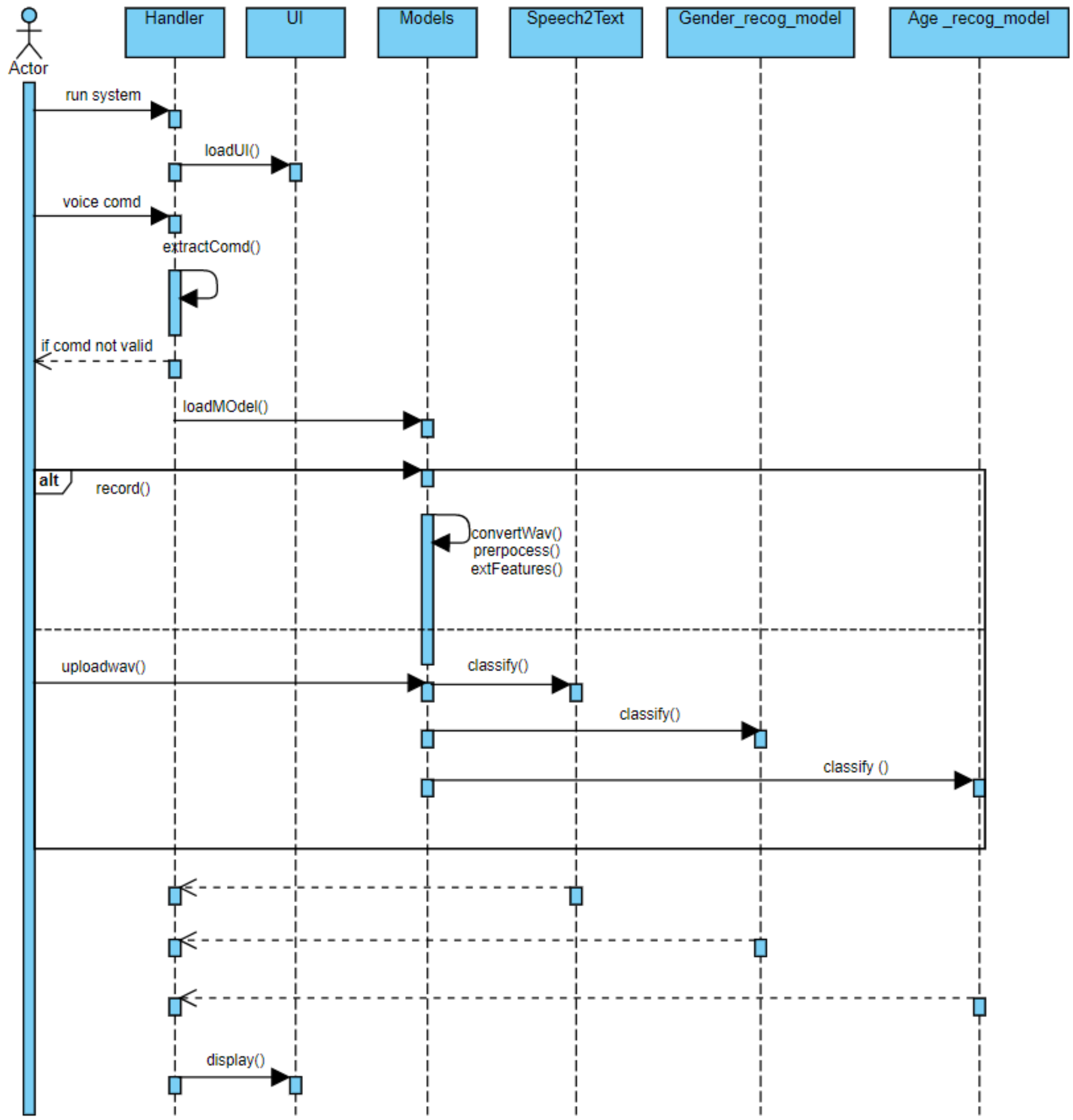


Figure 3.3.4 – Sequence Diagram

3.3.5. User View (Use case diagram)

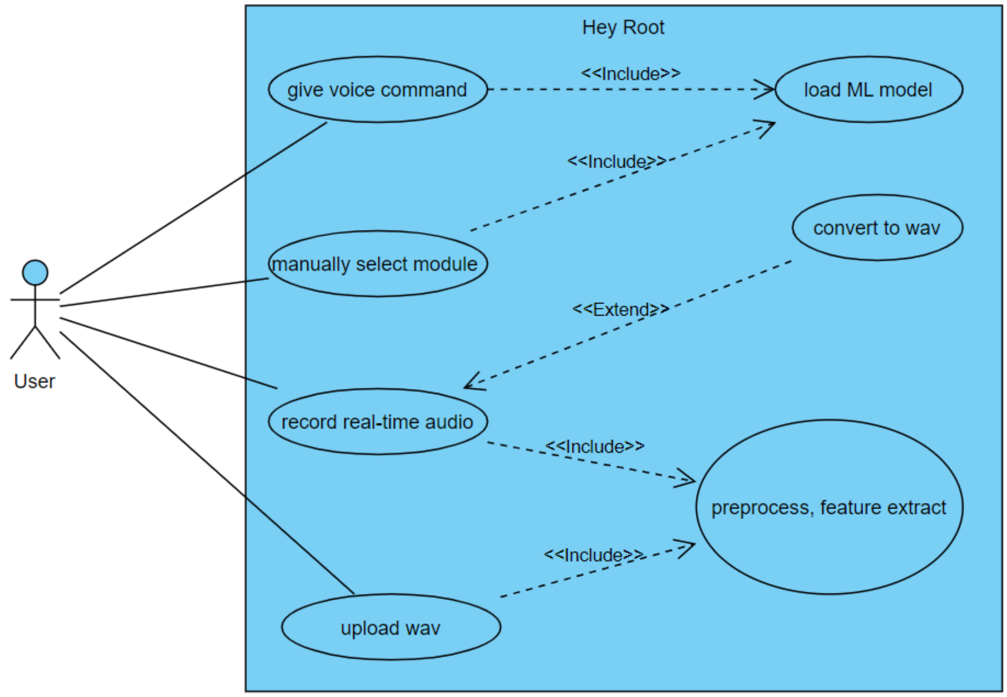
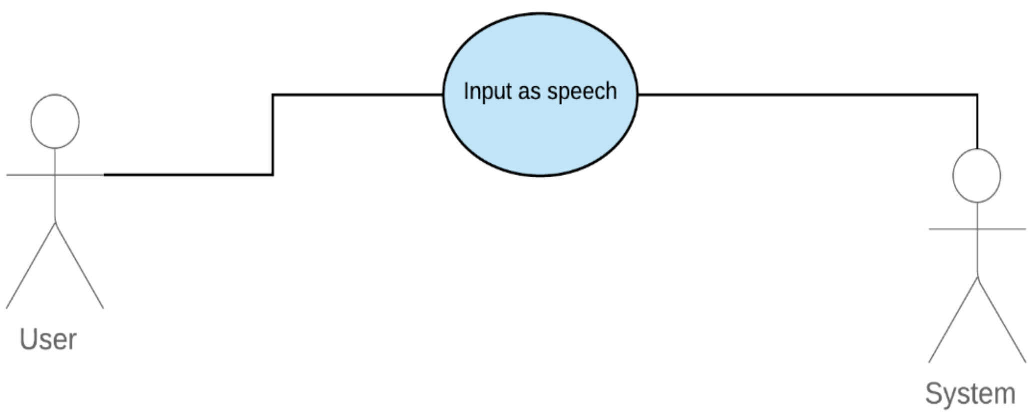


Figure 3.3.5.1 - Use case diagram

Use case 1

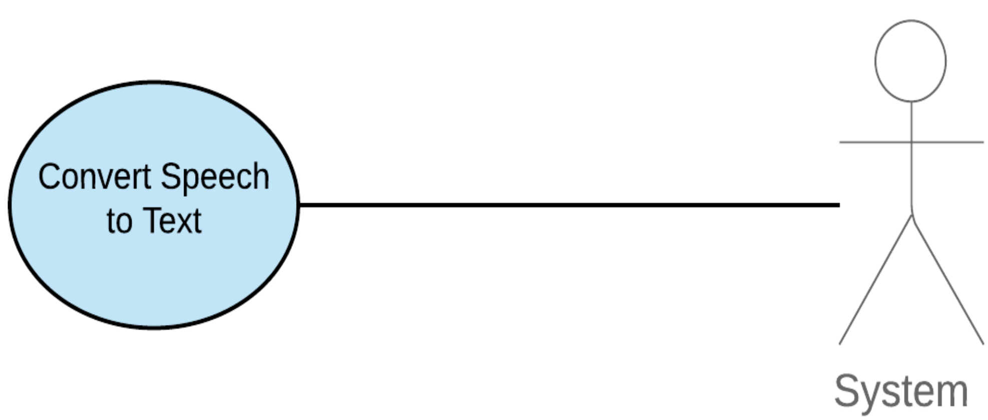


Use case name	Input as speech
Primary actor	user
Secondary actor	System

Normal course	- Initiate process - Give input
Alternate course	- Give input - initiate process - input failed - no process initiate
Pre-condition	System Currently running
Post-condition	Speech to text converter module receive input to convert
Extend	N/A
Include	N/A
Assumptions	Application is running

Table 3-3-5-1 (Use Case 1)

Use case 2

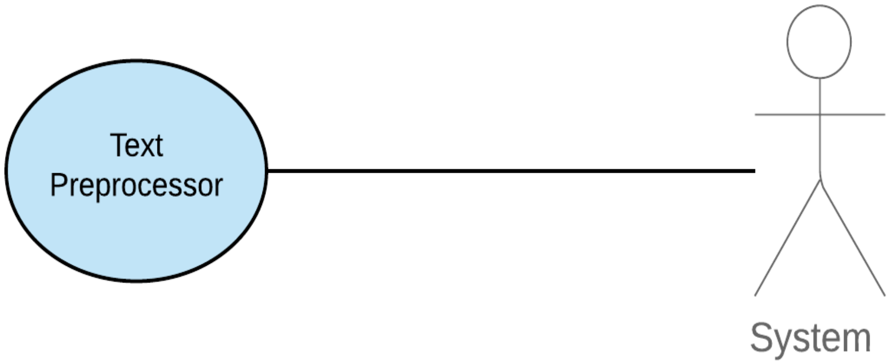


Use case name	**Convert speech to text**

Primary actor	System
Secondary actor	N/A
Normal course	- speech to text converter module receives input to convert - passes converted output to text pre-processor module
Alternate course	- speech to text converter module receives input to convert - error occurs - fails to convert - send error msg
Pre-condition	System has received input speech to convert it into text
Post-condition	Passes converted text to text pre-processor module
Extend	N/A
Include	N/A
Assumptions	App is running

Table 3-3-5-2 (Use Case 2)

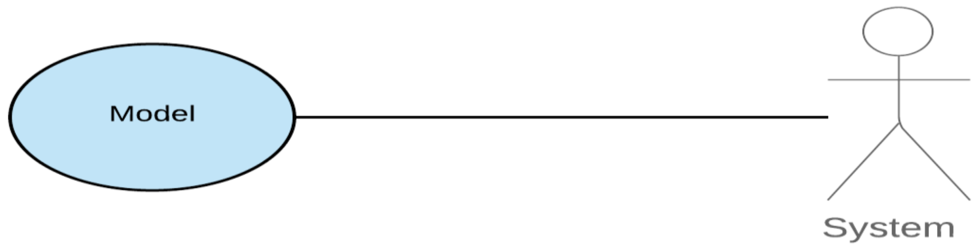
Use case 3



Use case name	**Text pre-processor**
Primary actor	System Control
Secondary actor	N/A
Normal course	- Takes raw text - pre-processes text - produces Word2Vector form
Alternate course	- Takes raw text - Error occurs - Send error msg
Pre-condition	Speech is converted into text successfully
Post-condition	Passes pre-process text to model
Extend	N/A
Include	N/A
Assumptions	App is running

Table 3-3-5-3 (Use Case 3)

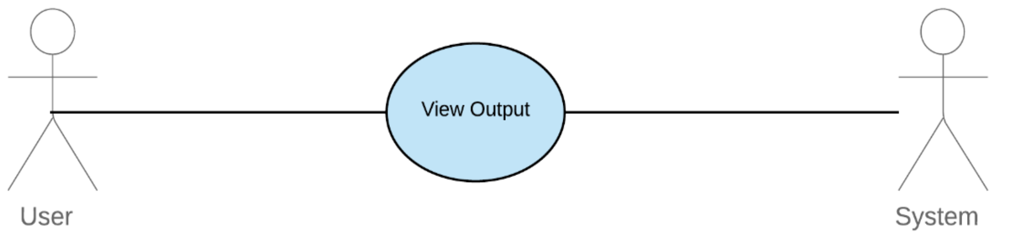
Use case 4



Use case name	**Model**
Primary actor	System Control
Secondary actor	N/A
Normal course	- Take pre-processed text - Pass it through model - Model makes inference
Alternate course	- Take pre-processed text - Pass it through model - Error occurs - Send error msg
Pre-condition	Text pre-processor successfully convert raw text into Word2Vector form
Post-condition	Passes the inference to user interface module through server
Extend	N/A
Include	N/A
Assumptions	App is running

Table 3-3-5-4 (Use Case 4)

Use case 5



Use case name	**View output**
Primary actor	System Control
Secondary actor	User
Normal course	- Show output to user through interface
Alternate course	- Take output/inference - Error occurs - Send error msg
Pre-condition	Inference has been generated and now needs to be shown
Post-condition	User views output
Extend	N/A
Include	N/A
Assumptions	App is running

Table 3-3-5-5 (Use Case 5)

CHAPTER 4. CODE ANALYSIS AND EVALUATION

4.1. Introduction

This record provides test documents to expand Automatic Speech, Age and Gender Recognition, Adaptation 1.0 that will promote specialized test-taking commitment that includes positive assessment of acquisition testing. Each test shows the ownership of the asset playing the test, the conditions required for each experiment, the object to be tested, the information, the expected yield or results, and the progress of the process where necessary.

4.2. Code Analysis

4.2.1 Main GUI

```
# -*- coding: utf-8 -*-

# Form implementation generated from reading ui file 'root.ui'
#
# Created by: PyQt5 UI code generator 5.15.4
#
# WARNING: Any manual changes made to this file will be lost when pyuic5 is
# run again.  Do not edit this file unless you know what you are doing.


from PyQt5 import QtCore, QtGui, QtWidgets


class Ui_MainWindow(object):
    def setupUi(self, MainWindow):
        MainWindow.setObjectName("MainWindow")
        MainWindow.setWindowModality(QtCore.Qt.NonModal)
        MainWindow.resize(800, 546)
        palette = QtGui.QPalette()
        brush = QtGui.QBrush(QtGui.QColor(255, 255, 0))
        brush.setStyle(QtCore.Qt.SolidPattern)
        palette.setBrush(QtGui.QPalette.Active, QtGui.QPalette.WindowText, brush)
        brush = QtGui.QBrush(QtGui.QColor(0, 85, 0))
        brush.setStyle(QtCore.Qt.SolidPattern)
        palette.setBrush(QtGui.QPalette.Active, QtGui.QPalette.Button, brush)
        brush = QtGui.QBrush(QtGui.QColor(0, 127, 0))
        brush.setStyle(QtCore.Qt.SolidPattern)
        palette.setBrush(QtGui.QPalette.Active, QtGui.QPalette.Light, brush)
        brush = QtGui.QBrush(QtGui.QColor(0, 106, 0))
```

```
        self.retranslateUi(MainWindow)
        QtCore.QMetaObject.connectSlotsByName(MainWindow)

    def retranslateUi(self, MainWindow):
        _translate = QtCore.QCoreApplication.translate
        MainWindow.setWindowTitle(_translate("MainWindow", "ROOT"))
        self.label.setText(_translate("MainWindow", "Hey Root - Automatic Speech, Gender and Age Recognit
        self.pushButton.setText(_translate("MainWindow", "Browse..."))
        self.pushButton_2.setText(_translate("MainWindow", "Detect"))
        self.pushButton_3.setText(_translate("MainWindow", "Record"))
        self.label_2.setText(_translate("MainWindow", "Upload Wav File "))
        self.label_5.setText(_translate("MainWindow", "Record Wav"))
        self.label_4.setText(_translate("MainWindow", "No File Selected ... "))
        self.label_3.setText(_translate("MainWindow", "Text: Down"))
        self.pushButton_4.setText(_translate("MainWindow", "Clear"))
        self.label_6.setText(_translate("MainWindow", "Gender: Male (80%) , Female(20%)"))
        self.label_7.setText(_translate("MainWindow", "Age: Twenties"))
        self.label_8.setText(_translate("MainWindow", "Wav recorded and saved to E:\\dataset\\myRecording


if __name__ == "__main__":
    import sys
    app = QtWidgets.QApplication(sys.argv)
    MainWindow = QtWidgets.QMainWindow()
    ui = Ui_MainWindow()
    ui.setupUi(MainWindow)
    MainWindow.show()
    sys.exit(app.exec_())
```

4.2.2 Modules

4.2.2.1 Speech to text

In [36]:

```
from keras.models import load_model
model.save("SpeechRecogModel.h5")
#model=load_model('/kaggle/working/best_model.hdf5')
```

Define the function that predicts text for the given audio:

In [37]:

```
def predict(audio):
    prob=model.predict(audio.reshape(1,8000,1))
    index=np.argmax(prob[0])
    return classes[index]
```

Prediction time! Make predictions on the validation data:

In [38]:

```
import random
index=random.randint(0,len(x_val)-1)
samples=x_val[index].ravel()
print("Audio:",classes[np.argmax(y_val[index])])
ipd.Audio(samples, rate=8000)
```

```
Audio: stop
```

Out[38]:

0:00 / 0:01

In [39]:

```
print("Text:",predict(samples))
```

```
Text: stop
```

4.2.2.2 Gender Recognition

Neural Network

Using neural_network.MLPClassifier to build the model.

In [41]:
```
def nn_error(n,x_train,y_train,x_test,y_test):
    error_rate = []
    hidden_layer=range(1,n)
    for i in hidden_layer:
        model = neural_network.MLPClassifier(solver='adam', alpha=1e-5,
                                       hidden_layer_sizes=i,
                                       activation='logistic',random_state=17,
                                       max_iter=2000)
        model.fit(x_train, y_train)
        y_pred = model.predict(x_test)
        error_rate.append(np.mean(y_pred != y_test))
    kloc = error_rate.index(min(error_rate))
    print("Lowest error is %s occurs at C=%s." % (error_rate[kloc], hidden_layer[kloc]))

    plt.plot(hidden_layer, error_rate, color='blue', linestyle='dashed', marker='o',
             markerfacecolor='red', markersize=10)
    plt.title('Error Rate vs. Hidden Layer Size')
    plt.xlabel('Size')
    plt.ylabel('Error Rate')
    plt.show()
    return hidden_layer[kloc]
```

In [42]:
```
h=nn_error(20,x_train,y_train,x_test,y_test)
```

```
Lowest error is 0.023133543638275498 occurs at C=3.
```

In [43]:
```
model = neural_network.MLPClassifier(solver='adam', alpha=1e-5,
                                     hidden_layer_sizes=h,
                                     activation='logistic',random_state=17,
                                     max_iter=2000)
classify(model,x_train,y_train,x_test,y_test)
```

	precision	recall	f1-score	support
female	0.9771	0.9771	0.9771	480
male	0.9766	0.9766	0.9766	471
micro avg	0.9769	0.9769	0.9769	951
macro avg	0.9769	0.9769	0.9769	951
weighted avg	0.9769	0.9769	0.9769	951

```
In [45]:
model = neural_network.MLPClassifier(solver='adam', alpha=1e-5,
                                     hidden_layer_sizes=h,
                                     activation='logistic',random_state=17,
                                     max_iter=2000)
classify(model,x_train3,y_train3,x_test3,y_test3)
```

```
              precision    recall  f1-score   support

      female     0.9730    0.9771    0.9751       480
        male     0.9765    0.9724    0.9745       471

   micro avg     0.9748    0.9748    0.9748       951
   macro avg     0.9748    0.9747    0.9748       951
weighted avg     0.9748    0.9748    0.9748       951
```

We can see that the highest accurracy is 98.74% which is made by XgBoost. XgBoost is a powerful algorithm, and very popular in Data Science competition. Next time I will try to oppotimize the parameters of XgBoost.

4.2.2.3 Age Classification

```
In [31]: def test_prediction(df, model, scaler, path):
             features = feature_extraction(path)
             gender = features[0]
             features = scaler.transform(features.reshape(1, -1))  # reshape because we have a single sample
             features = features[0]   # beacause the shape is (1, 24), but we want (24, ) as shape
             features[0] = gender     # in this way the gender will be always +1, 0 or -1

             print("true age:      ", get_age(df, path))
             prediction = model.predict(np.expand_dims(features, axis=0))
             print("predicted age: ", labels[np.argmax(prediction)])


         path = join("cv_corpus_v1", "wav-files", "cv-valid-test", "sample-000001.wav")
         play_sound(path)
         test_prediction(df, model, scaler, path)

         true age:       twenties
         predicted age:  twenties

In [36]: path = join("cv_corpus_v1", "wav-files", "cv-valid-test", "sample-000009.wav")
         play_sound(path)
         test_prediction(df, model, scaler, path)

         true age:       fifties
         predicted age:  fifties
```

4.3. Test Items

Based on the requirements of the automated language, age, and gender detection project, the main modules / features that need to be considered during the testing process are:

- Recording audio and converting to Wav
- Uploading Wav File
- Conversion of Wav to English text
- Showing of Age and Gender classification on user interface

4.4. Features to Be Tested

Following features are being tested:

- Ability to allow the user to upload a file (wav).
- Ability to allow the user to record speech in English
- Ability to allow the user to convert the English audio file and recorded audio file (Real-time English audio) into English Text.
- Ability to detect the content from English text.
- Ability to show output on user interface.
- Ability to clear the text box.

4.5. Item Pass/Fail Criteria

Details of the test cases are specified in the section Test Deliverables. Following the principles outlined below, a test item would be judged as pass or fail.

• Prerequisites met
• Input runs as specified
• Results work as specified in output => Pass
• System does not work or matches output specifications => Failure

4.6. Suspension Criteria and Resumption Requirements

Testing will be suspended when a deformity is presented/discovered that can't permit any further testing. Testing will be continued after imperfection expulsion.

4.7. Test Deliverables

Following are the test cases:

Test case name	**File Upload**
Test Case Number	1
Description	This feature let the user to upload audio in English language.
Testing Technique used	Black Box Testing
Preconditions	file must be in wav format.
Input	File must be in wav format.
Steps	• Select audio file (wav) in English. • Upload it by using submit button on interface
Expected output	File will be uploaded and its respected text will be written in Text Area.
Alternative Path	• N/A
Actual output	Confirmed

Table 4-8-1 (Test Case 1)

Test case name	**File Upload (Invalid format)**
Test Case Number	2
Description	This feature let the user to upload audio (wav) in English language.
Testing Technique used	Black Box Testing
Preconditions	File must not be in wav format.
Input	Invalid File format.
Steps	• Select file (any). • Upload it by using submit button on interface
Expected output	“Wrong format” will be written on interface.

Alternative Path	• N/A
Actual output	Confirmed

Table 4-8-2 (Test Case 2)

Test case name	**File Upload (Invalid language)**
Test Case Number	3
Description	This feature let the user to upload audio in English language.
Testing Technique used	Black Box Testing
Preconditions	File must be in wav format.
Input	File must be in wav format other than English.
Steps	• Select file (audio) in other than English language. For Example Urdu • Upload it by using submit button on interface
Expected output	Unknown result
Alternative Path	• N/A
Actual output	Confirmed

Table 4-8-3 (Test Case 3)

Test case name	**Recording**
Test Case Number	4
Description	The user can record a real time audio (wav) in English.
Testing Technique used	Black Box Testing
Preconditions	Recording button must be clicked before recording.
Input	Real-time recorded Urdu audio.

Steps	• Open User Interface page. • Click record button. • Then click the stop button to stop recording.
Expected output	Audio file will be recorded in wav format.
Alternative Path	• N/A
Actual output	Confirmed

Table 4-8-4 (Test Case 4)

Test case name	**Audio to Text conversion**
Test Case Number	5
Description	This feature allow the user to convert user audio file into English text.
Testing Technique used	Black Box Testing
Preconditions	• For recorded audio, recording must be started. OR • Person must select a file to upload
Input	Audio file
Steps	• For recorded audio conversion, press stop button to stop it. OR • Click Submit button to upload file.
Expected output	English text will be displayed.
Alternative path	**N/A**
Actual output	As per audio speech

Table 4-8-5 (Test Case 5)

Test case name	**Gender Recognition**
Test Case Number	6
Description	This feature let users to identify Gender from Audio.

Testing Technique used	Black Box Testing
Preconditions	User upload a file or record the real time audio file in wav.
Input	English audio
Steps	• Click detect button
Expected output	English text
Alternative path	**N/A**
Actual output	Confirmed

Table 4-8-6 (Test Case 6)

Test case name	**Age Recognition**
Test Case Number	7
Description	This feature let users to detect Age from Audio (wav).
Testing Technique used	Black Box Testing
Preconditions	User upload a file or record the real time audio file in wav.
Input	English audio
Steps	• Click detect button
Expected output	English text
Alternative path	**N/A**
Actual output	Confirmed

Table 4-8-7 (Test Case 7)

Test case name	**Speech to text**
Test Case Number	8
Description	This feature let users to convert speech to text from Audio (wav).
Testing Technique used	Black Box Testing
Preconditions	User upload a file or record the real time audio file in wav.
Input	English audio

Steps	• Click detect button
Expected output	English text
Alternative path	**N/A**
Actual output	Confirmed

Table 4-8-8 (Test Case 8)

CHAPTER 5: CONCLUSION

Estimating the age, gender of speaker using real-time speech has also gained focus in technology due to the market demands and increasing security and other concerns. The requirement of the user of this project are a user-friendly interface, accurate speech to text, fast output, accurate estimation of gender and age, probably model should listen and also that the project is available for Android. To cope up with the requirements our team will work on Automatic Speech, Gender and Age Recognition using Deep Neural Networks, creating an attractive and simple interface, machine learning model and deep neural networks for accurate and fast outputs. The proposed project has applications in commercial advertisements, forensic science, biometric recognition technology and suspicious call detection. The project has a simple and interactive GUI which make it easy of understanding for the user. Future work of the project includes recognizing speech, gender and age using Urdu language. Main module of the project also allows the user to load pre-trained model and the GUI displays results, visualize datasets and plots MFCC spectogram. The proposed project has been implemented in Python using Tensorflow and Keras and accuracy greater than 90% is achieved in each model.

CHAPTER 6: FUTURE WORK

The Project can be extended further by adding following modules:

1. Addition of Urdu language model.
2. Addition of extracting text (English) from video.
3. Making an Android & IOS Applications of it.

Further improvements and accuracy of detecting model can be achieved by retraining of model again and again with respect to time.

REFERENCES AND WORK CITED

1. Dhanashri, D., Dhonde, S.B.: Speech recognition using neural networks: a review. Int. J. Multidiscip. Res. Dev. 2(6), 226–229 (2015)
2. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural network. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (2013)
3. Halageri, A., Bidappa, A., Arjun, C., Sarathy, M., Sultana, S.: Speech recognition using deep learning. Int. J. Comput. Sci. Inf. Technol. 6(3), 3206–3209 (2015)
4. Lekshmi, K., Dr. Sherly, E.: Automatic speech recognition using different neural network architectures a survey. Int. J. Comput. Sci. Inf. Technol. 7(6), 2422–2427 (2016)
5. S. B. Kalluri, D. Vijayasenan and S. Ganapathy, "A Deep Neural Network Based End to End Model for Joint Height and Age Estimation from Short Duration Speech," ICASSP 2019 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6580-6584, doi: 10.1109/ICASSP.2019.8683397.
6. Rita Singh, Bhiksha Raj, and James Baker, "Short-term analysis for estimating physical parameters of speakers," in Proc. of IWBF. IEEE, 2016, pp. 1–6
7. Y. Xie, L. Le, Y. Zhou and V. V. Raghavan, "Deep learning for natural language processing" in Handbook of Statistics, Amsterdam, The Netherlands:Elsevier, 2018.
8. J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey", *IETE*
9. A.A. Assim, V. Davydov, V. Yushkova, A. Glinushkin, V. Rud, Age and gender recognition from speech signals. J. Phys. Conf. Ser. 1410, 012073 (2019).

PLAGIARISM REPORT

Speech, Age and Gender Recognition using Deep Neural Networks

ORIGINALITY REPORT

12%	10%	2%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	2
2	www.techtarget.com Internet Source	1%
3	senior.ceng.metu.edu.tr Internet Source	1%
4	theaisummer.com Internet Source	1%
5	fics.nust.edu.pk Internet Source	1%
6	Submitted to Kingston University Student Paper	1%
7	Assim Ara Abdulsatar, V V Davydov, V V Yushkova, A P Glinushkin, V Yu Rud. "Age and gender recognition from speech signals", Journal of Physics: Conference Series, 2019 Publication	1%

8	nanopdf.com Internet Source	1 %
9	Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, Tetsuya Ogata. "Audio-visual speech recognition using deep learning", Applied Intelligence, 2014 Publication	1 %
10	Submitted to University of Malaya Student Paper	1 %
11	keepass.info Internet Source	1 %
12	www.slideshare.net Internet Source	1 %
13	Submitted to UNITEC Institute of Technology Student Paper	<1 %
14	Submitted to Staffordshire University Student Paper	<1 %
15	www.scribd.com Internet Source	<1 %
16	hdl.handle.net Internet Source	<1 %
17	Submitted to University of Technology, Sydney Student Paper	<1 %
	Submitted to University College London	

18	Student Paper	<1 %
19	fdocuments.in Internet Source	<1 %
20	Submitted to University of Lancaster Student Paper	<1 %
21	repository.tudelft.nl Internet Source	<1 %
22	Submitted to Curtin University of Technology Student Paper	<1 %
23	Susan K. Land, Douglas B. Smith, John W. Walz. "Practical Support for Lean Six Sigma Software Process Definition", Wiley, 2008 Publication	<1 %
24	Submitted to University of Wolverhampton Student Paper	<1 %
25	etd.aau.edu.et Internet Source	<1 %
26	shashank-srikant.github.io Internet Source	<1 %
27	ijrjournal.com Internet Source	<1 %
28	www.coursehero.com Internet Source	<1 %

29	Jayaprada S. Hiremath, Shantakumar B. Patil, Premjyoti S. Patil. "Human Age and Gender Prediction using Machine Learning Algorithm", 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC), 2021 Publication	<1 %
30	www.restfulwhois.org Internet Source	<1 %
31	suraj.lums.edu.pk Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		