Software Requirements Specification

for

Recognition of Speech Emotion

Version 1.0 approved

Prepared by Group-13

Gonuguntla Harichandana (BT19GCS038)

Anayna Nidhi Singh (BT19GCS114)

Kalluri Divija (BT19GCS122)

Archa Dave (BT19GCS003)

Ayush Pandey (BT19GCS098)

NIIT University

25-09-2022

Table of Contents

Revision History

Name	Date	Reason For Changes	Version
Recognition of Speech Emotion-SRS Version-1.0	25-09-2022	First edit.	Version-1.0

1. Introduction

1.1 Purpose

This software requirement specification document provides a complete description of all the functionalities and specifications of a system for the recognition of speech emotion. This will explain the different functional and non-functional requirements of the system, the interface of the system, and how the system will behave and interact with the user. This SRS will provide a clear understanding of what is expected by the client. The main goal of the project is to design a scalable and extensible system for the audience. This system aims to provide recognition of emotions from speech. The system will be designed with a user-centric approach to ensure that the user requirements are fulfilled.

1.2 Document Conventions

This document was created based on the IEEE template for System Requirement Specification Documents. To make the document easier to read, important jargon and technical terms have been presented in bold font, for better segmentation bullet points and numbering have been used wherever possible, and for better readability, a slightly larger font is selected for the headings as compared to normal explanation.

1.3 Intended Audience and Reading Suggestions

This document is formulated in a very sequential manner whilst the users are relieved to jump to any section they find relevant below is a brief overview of each part of the document.
Part 1 (Introduction): This section offers a summary of the project, including goals and objectives, project scope, general system details, and some major constraints associated with the intended platform.
Part 2 (Overall Description): This section describes the system class by class, including interface details, class hierarchies, performance/design constraints, process details, and algorithmic models.
Part 3 (External Interface Requirements): Readers interested in how to organize and handle data should consult this section, which covers data and flow patterns utilized by the system.
Part 4 (System Features): This section covers all of the details related to the structure of the user interface, including some preliminary mockups of the web application. Readers can view this section for a tentative glimpse of what the final product will look like.
Part 5 (Other Nonfunctional Requirements): Any other requirements which are not covered in the SRS will be mentioned in this section.

1.4 Product Scope

Speech emotion recognition systems can be used in multiple sectors. It can be integrated with smart voice assistants like Alexa and Siri and be used in a variety of applications.
It can also be used in music listening apps and in music recommendation systems where the app can recommend the users songs to listen to on the basis of their current emotions. This judgment of current emotion can be done via speech recognition. If the user is feeling sad, the app can play happy songs to cheer him/her up.
The speech emotion recognition system can also be used in healthcare applications where the user's emotion can be detected by what they speak. It can also be used by psychologists to treat patients. Apart from that, this system can be used in mental health apps to gauge the emotion of the users by interacting with them and making them speak something. Then this app can provide some valuable counseling or feedback to the user. It can also recommend users try out different activities to pursue based on the type of emotion they are encountering.
In the marketing sector, this system can be used to sell products and gain customers. With the easy-to-use user interface of the system, the user can either record instant audio or upload an existing file to the system and perform emotion recognition. This system allows big corporate companies to measure customer satisfaction and perform the necessary analysis.
Determination of a user's emotional state with facial and voice analysis plays a fundamental part in human-machine interaction (HMI) systems since it employs non-verbal cues to estimate the user's emotional state. This software system will be able to perform emotion recognition from audio.

1.5 References

[1]Apoorva Ganapathy, "Speech Emotion Recognition Using Deep Learning Techniques", BC Journal of Advanced Research, Speech Emotion Recognition Using Deep Learning Techniques | Request PDF

[2]Zhou Qing, Wang Zhong, Wang Peng, "Research on Speech Emotion Recognition Technology Based on Machine Learning", 2020 7th International Conference on Information Science and Control Engineering (ICISCE),20 September 2021, DOI: 10.1109/ICISCE50968.2020.00247,https://ieeexplore.ieee.org/document/9532213

[3] Sung-Woo Byun, Seok-Pil Lee 2, “A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms”, MDPI Journal Special Issue in AI, Machine Learning and Deep Learning in Signal Processing, 21 February 2021, A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms

[4]Qingli Zhang, Ning An, Kunxia Wang, Fuji Ren, Lian Li, “Speech emotion recognition using a combination of features”, 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 25 July 2013, Speech emotion recognition using combination of features | IEEE Conference Publication

2. Overall Description

2.1 Product Perspective

A speech emotion recognition web app can detect the emotional state of a person from his speech information and audio files. In this scope, an audio emotion recognition system requires evaluating the emotion of a person from his speech which evaluates certain features that help determine the emotion of the person of interest. The model will focus on evaluating the pitch, time, and other features and give an output of the emotion of the person.

2.2 Product Functions

The software described in this SRS will be used to detect people's emotions. The main functions of the product include audio data extraction, detection, and analysis. Apart from that, the major function of the network model implemented in the web app is to give accurate emotion as the output using feature extraction and feature analysis.

2.3 User Classes and Characteristics

The user who uses this system should have basic knowledge of computers and the basic idea of navigating through a web app since the system deals with authentication, the user must be able to create a user id and create a password to save his data. When the user is working with the main application of this web app, the user is expected to know what kind of result will come and for what purpose he/she is using it. The user classes and their domain usage depends on the scope of this web application. It extends to starting from the tech world to advance the automation in personal assistants, robots, makes websites or chatbots smarter and ranging to marketing and customer attraction for selling products in some domains of marketing. The user class, like the medical field, can use this application towards psychology and treat patients by knowing the mood or intention of the patient so they can approach it in a clinical way. The user class of applied sciences also uses this product for exploration and for curiosity reasons. Any user from any class is expected to go through the user manual and apply it for smooth functioning.

2.4 Operating Environment

Since docker is going to be used to package and containerize the application, it will be easier to deploy it on any operating system which has docker run time on it. Other than that the operating environment should preferably be updated to the latest version possible. The software and hardware requirements are anything that is compatible with recent versions of our web app.

2.5 Design and Implementation Constraints

The design and implementation constraints include achieving the maximum possible accuracy because human emotions are varied and complex. Designing a model that covers a range of emotions requires a huge training dataset and implementing that will require complex systems.

2.6 User Documentation

The product will be provided with a user manual, where users can find the directions and instructions to navigate through the web application and find the feature's usage to make use of the feature for the associated function. Also, the user manual will be provided with the limitations of the application.

2.7 Assumptions and Dependencies

Here assumptions include that a fixed set of human emotions exist and we are working with a set of emotions. The speech is always available in a proper audio format without any noise. The features always belong to a predefined category of defined functions in the code. The web app heavily depends on the training data.

3. External Interface Requirements

3.1 User Interfaces

The user interface shall provide a simple and aesthetic interface to the user and the administrator. It will be a website and the user can easily log in via his/her email and navigate through it. There will be a simple option to upload the user's speech for 4 seconds. After uploading the voice via a microphone or uploading an audio file from the system, it will display the results on the website.

3.2 Hardware Interfaces

Sound cards- Speech requires a lower bandwidth relatively so a 16-bit sound card will work fine. The proper driver should be installed and sound must be enabled. Some speech emotion recognition systems may require sound cards that are a bit more specific.

Microphone- A microphone is of utmost importance when building a speech emotion recognition system. The best microphone is a headset-style microphone. These microphones reduce ambient noise. Desktop microphones have the propensity to pick up comparatively more ambient noise so they are not suggested, though it can be considered.

Processors- Speech emotion recognition systems can be heavily dependent on processing speed. This is due to the fact that a huge amount of digital filtering and signal processing takes place in ASR.

Monitor- The monitor screen shall display the information to the user.

Keyboard- The keyboard will be used to interact with the system via basic keyboard keystrokes.

3.3 Software Interfaces

The computer to be used must have the libraries attached to python. Some of these libraries are Librosa, Scikit-Learn, NumPy, etc.

Librosa- Librosa is a python package that is used for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It has a flatter package layout, standardized interfaces and names, backward compatibility, modular functions, and readable code.

Scikit-Learn- Scikit-learn is an open-source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

Numpy- NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices.

JupyterLab- JupyterLab is an open-source, web-based UI for Project Jupyter and it has all the basic functionalities of the Jupyter Notebook, like notebooks, terminals, text editors, file browsers, rich outputs, and more. However, it also provides improved support for third-party extensions.

MongoDB- MongoDB is a source-available cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas.

ExpressJS- Express.js, or simply Express, is a back-end web application framework for building RESTful APIs with Node.js.

React- React is a free and open-source front-end JavaScript library for building user interfaces based on UI components.

NodeJS- Node.js is an open-source, cross-platform, back-end JavaScript runtime environment that runs on a JavaScript Engine and executes JavaScript code outside a web browser, which was designed to build scalable network applications.

Minimum System requirements-

- Pentium 200 MHz Processor
- 64 MB of RAM
- Microphone
- Sound Card

Best System requirements-

- 1.6 GHz Processor
- 128 MB or more of RAM
- Sound cards with very clear signals
- High-quality microphones

3.4 Communications Interfaces

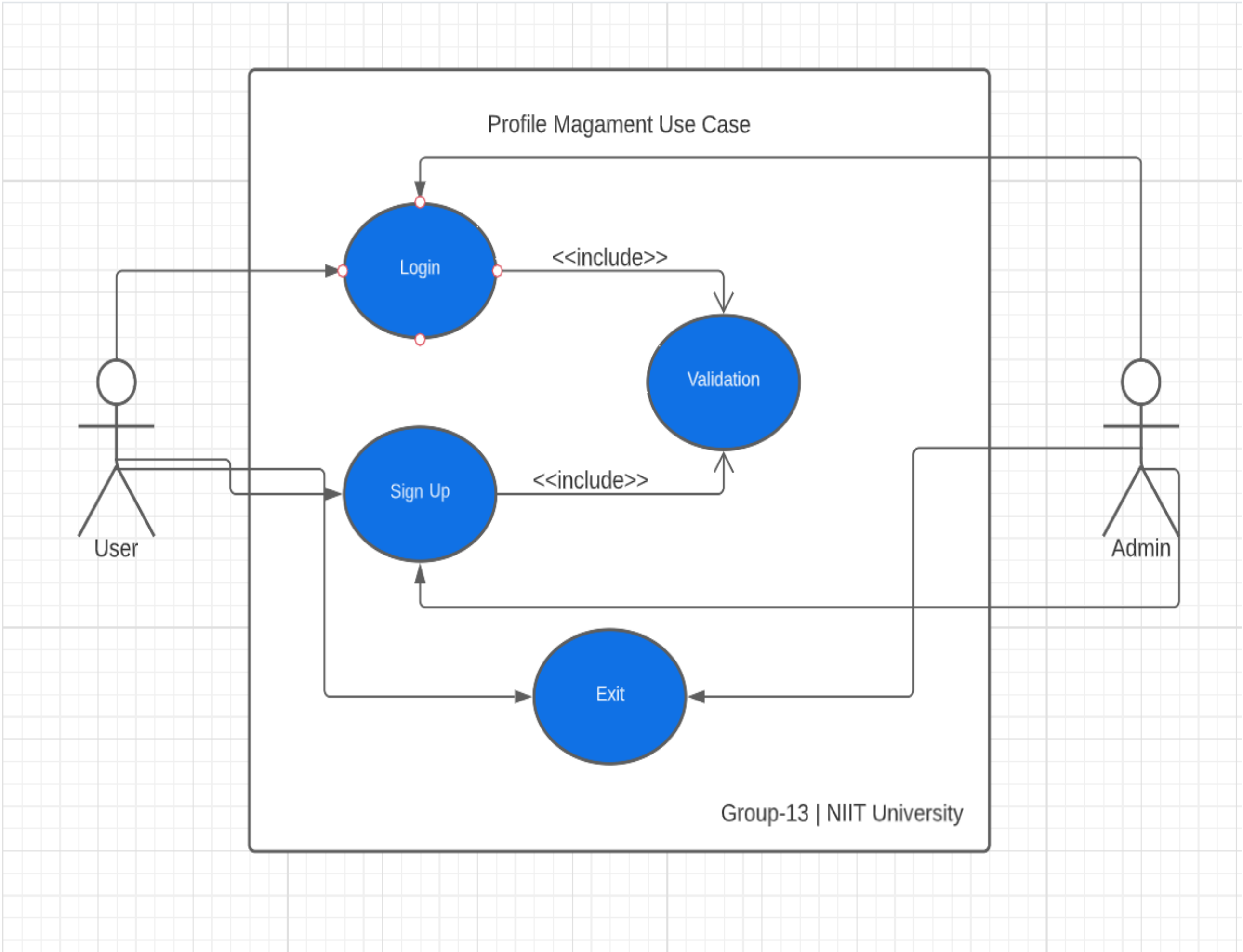
The HTTP or HTTPS protocol(s) will be used to facilitate communication between the client and server.

4. System Features

The main features of this system are Authentication and the application of Recognition of speech emotion upon uploading the input audio file or recording the audio and then showing the content information and exiting the system. From the admin point of view, the admin has privileges like manipulating the data. The following are detailed explanations of each feature with the use of Use case diagrams.

4.1 Profile Management Use Case

Profile management-use case diagram: Figure 4.1



4.1.1 Description and Priority

Profile Management has activities like Login, Signup, Validation, and Exit. The above figure 4.1 says a detailed view of profile management and its flow. At first, either the admin or the user has to go through the authentication menu when entering the system. Then the user or the admin has to use functions like Sign Up, and Sign In or Sign In to enter the main working page where recognition of speech emotion is carried out. Later if the user wants to exit the system then the user can choose to Exit.

4.1.2 Stimulus/Response Sequences

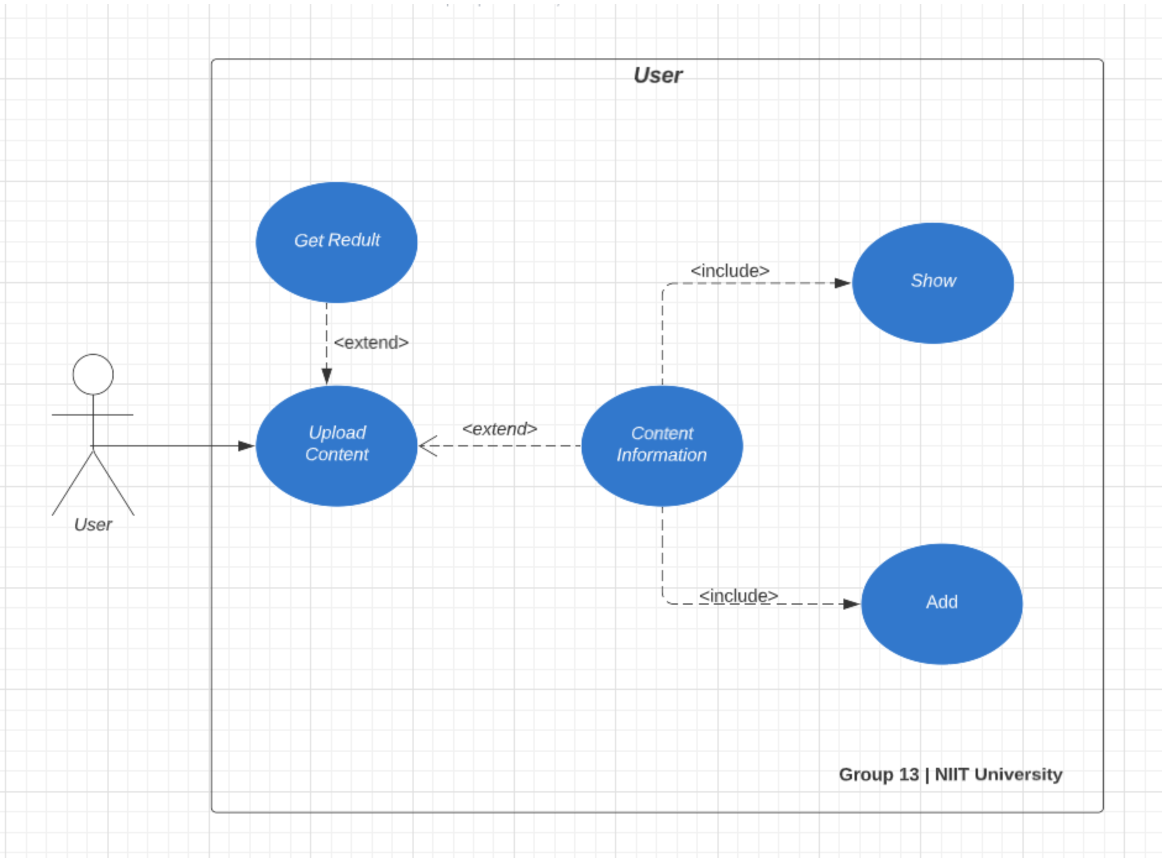
- Response step-1: If the user is new to the system then the user must sign up and then log in or if the user already exists then the user can log in to the system.
- Response step-2: If the username and password combination is true then enter the next page else if the combination is invalid, then the user must re-login using the correct credentials.
- Response step-3: If the user chooses to exit the system, then Exit.

4.1.3 Functional Requirements

- Sign-up and register page
- Logging in with Google and its verification
- An “I forgot my password” section that sends a link to the verified email
- Navigating the website
- Clicking on the add audio button
- Choosing the option of either uploading an audio file from the system or recording it right now.

4.2 User Use Case

User-use case diagram: figure 4.2



4.2.1 Description and Priority

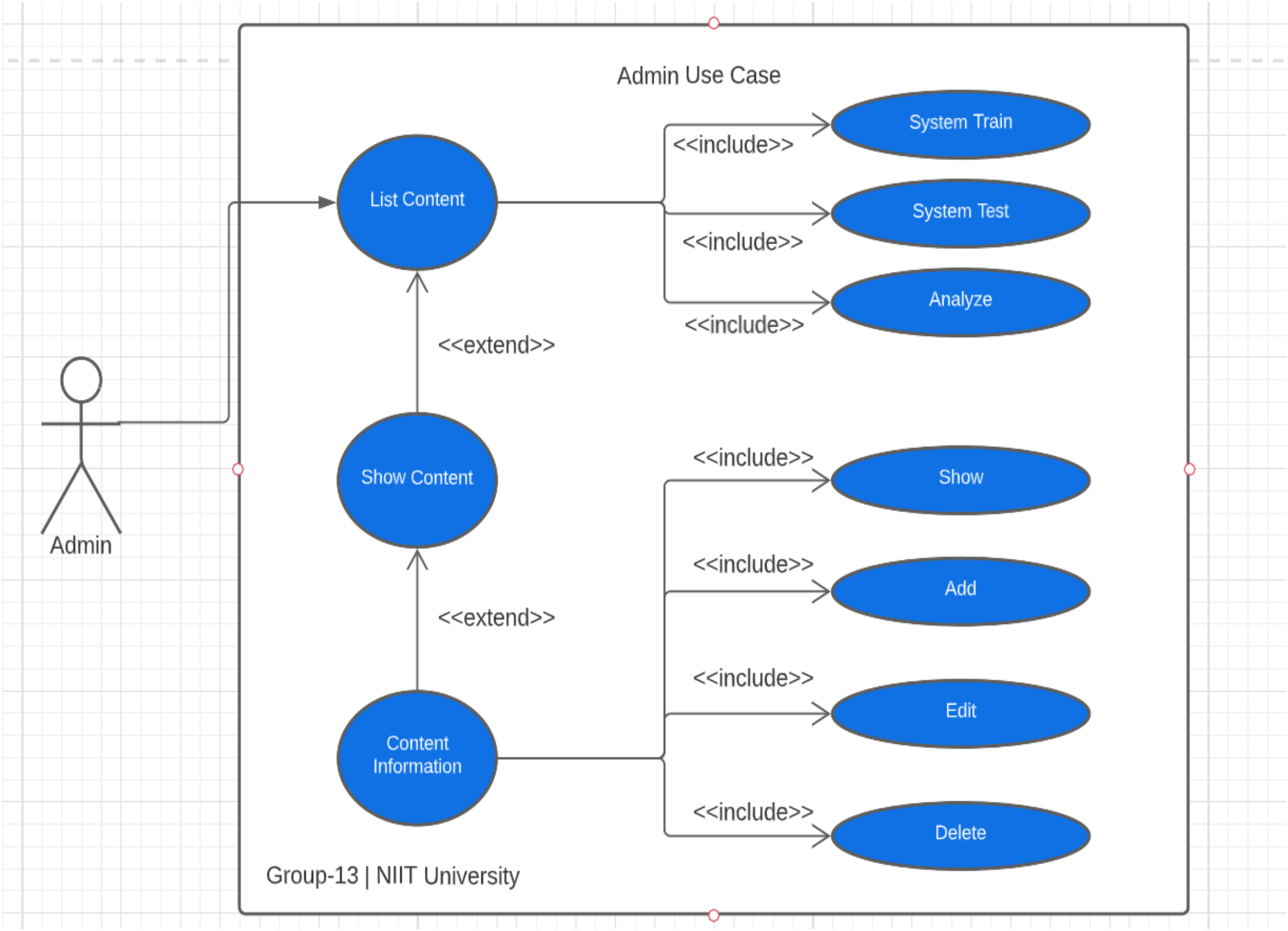
The above use case diagram figure 4.2 of the User has activities like Upload Content(audio file), Get the result (Emotion Category), Content Information, Add and Show. This defines and explains the type of action the user has performed on the system to get a result. Also, users are able to Add and Show content information and look at all uploaded content in the system.

4.2.2 Stimulus/Response Sequences

- Response step-1: The user has to select Upload Content/Record content, and the system waits for the user to submit the audio file from the user's computer which only supports waw format files.
- Response step-2: After the user submits the audio file, the user should enter, add and view content information.
- Response step-3: When the user uploads the recorded or uploaded audio file, the system processes the given audio file and gives the user the result as an emotion category as output.

4.3 Admin Use Case

Admin-use case diagram: figure 4.3



4.2.1 Description and Priority

The above Admin-use case diagram figure 4.3 explains and shows the activities like list content, system train, system test, analyze, show content, content information, add, edit and delete. The admin is authorized to intervene in the system. The figure also explains the admin’s privileges.

4.2.2 Stimulus/Response Sequences

- Response step-1: The admin is able to see all uploaded content in the system.
- Response step-2: The admin can train and test the system using those contents.

- Response step-3: The admin can analyze the content as in data statistics or any other factor analytics to modify the model further to increase the model performance.
- Response step-4: The admin can view all the content and edit, delete and view added content's information.

5. Other Nonfunctional Requirements

5.1 Performance Requirements

The minimum system requirements for the computer to be used are as follows:

1. Processors: Intel® Core™ i3 processor
2. Disk Space: 1 GB
3. Operating Systems: Linux,macOS, and Windows7 or later
4. Python versions: 3.6.X or higher
5. Included development tools: Anaconda
6. Compatible tools: Jupyter, VScode

5.2 Safety Requirements

System reliability will improve as long as the audio's sound quality is good and the person's voice is clearly audible. Since the upload size and type of the file to be uploaded are limited, no system crashes will be allowed.

5.3 Security Requirements

In order to improve the software, we will store the data in our system and then we will use the data to develop our system. This data will be used to increase the stability of the system. Therefore, the data will only be used for system improvement.

5.4 Software Quality Attributes

The system will work on all operating systems. In order to increase the stability of the software, the testing and training files of the software will be updated once a month by the administrator. Since the developed application is a user-oriented project, it should provide simple usage to the user. Therefore, the interface that we will be developing will be understandable and user-friendly.

6. Other Requirements

The computer to be used must have microphone input for voice recordings. It must have the libraries attached to python. Some of these are Librosa, Scikit-Learn, and Numpy, and also the computer must have MongoDB, ExpressJS, React, etc.
There is an internet connection required to run this software.

Appendix A: Glossary

Term	**Definition**
SRS	Software Requirement Specifications
Admin	Person who manages the system
API(Application Programming Interface)	It allows two applications to talk to each other.
Personal Assistant	A virtual assistant that works for a particular person upon asking , Example : Alexa, Siri etc.
User	Person who wants to know the situation of emotion
Mp4	A file format created by the Moving Picture Experts Group(MPEG) as a multimedia container format designed to store audio data
Waw	A file format for speech.
Content	Speech-Audio
Certainty	How likely an event or a statement is supposed to be true
Bot	User-friendly chat bot or personal assistant assigned for customized application.
Probabilistic method	Proving the existence of combinatorial

	objects with specified properties
Neural Network	A computer system modeled on the human brain and nervous system.
Multi Layer Perceptron(MLP)	A type of neural network. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers.
Recurrent Neural Network(RNN)	Recurrent neural networks recognize data's sequential characteristics and use patterns to predict the next likely scenario.

Appendix B: Analysis Models

Speech emotions have improved a lot over the past experiences. The focus of predicting certain emotions based on a snippet of information is human predictable but when it comes to the field of automation, the field of Machine learning and Neural networks play a major role in identifying the results and predicting the sample classification. Here in this project, the use of neural networks or machine learning in any case does not give certainty. Instead we deal with a probabilistic approach. But when working with probabilistic models the task of improving the accuracy and precision plays a key role in order to predict almost correct answers to the data. The neural network that we are using to predict the emotions of the human voice is Multi layer perceptron (MLP) and to be precise it follows Recurrent neural network (RNN) type, where neural network is capable of taking audio signals as input and process them through its layers and predict the emotion category. The scope of this product lies wherever the speech emotion is a factor of a conversation. Our application can be paired with modern day personal assistants, can be used in automated fields like robotics and where robots can be used in field psychology treatments, where a bot plays a key role in identifying the emotions of a human.

Appendix C: To Be Determined List

None at this time.