ÇANKAYA UNIVERSITY

Speech Emotion Recognition Report

Speech Emotion Recognition

Furkan Duran - 201811027, Elif Aybüke Coşkun – 201811018, Abdullah Özder - 202011410, Şima Kayısı - 201811043, İhsan Bardakcı - 201717007

Table of Contents

İçindekiler

Abstract

Speech emotion recognition is a task of human-computer interaction. People often prefer verbal communication to communicate. It is possible to extract emotion from this communication. Emotion recognition features are extracted from speech signals, features are selected, and emotions are recognized. This report covers our speech emotion recognition project, speech and explains our purpose of doing mood analysis from text. In project development: Python to manage image, audio and text processing. Machine learning algorithms and artificial neural networks will be used to train the model. Librosa, a Python package for music and sound analysis, will also be used.

Keywords: Machine Learning, Librosa, Python, Human-Computer Interaction.

Özet

Konuşma duygu tanıma, insan-bilgisayar etkileşiminin bir görevidir. İnsanlar genellikle iletişim kurmak için sözlü iletişimi tercih eder. Bu iletişimden duygu çıkarmak mümkündür. Duygu tanıma özellikleri konuşma sinyallerinden çıkarılır, özellikler seçilir ve duygular tanınır. Bu rapor, konuşma duygu tanıma projemizi kapsar, konuşma ve metinden ruh hali analizi yapma amacımızı açıklar. Proje geliştirmede: Görüntü, ses ve metin işlemeyi yönetmek için Python. Modeli eğitmek için makine öğrenimi algoritmaları ve yapay sinir ağları kullanılacaktır. Müzik ve ses analizi için bir Python paketi olan Librosa da kullanılacak.

Anahtar Kelimeler: Makine Öğrenimi, Librosa, Python, İnsan-Bilgisayar Etkileşimi.

1. Introduction

The speech emotion recognition project is a machine-human interaction that involves using computers to recognize and interpret emotions expressed in spoken language. This technology has the potential to improve various aspects of human-machine interaction, including customer service in call centers, the effectiveness of virtual assistants, and the diagnosis of mental health conditions.

Different techniques need to be used to analyze the acoustic properties of each speech and the content of the spoken words. Machine learning algorithms must be used to extract certain features of the speech signal, such as pitch, intensity, and spectral features, and classify emotions based on these features.

With speech emotion recognition technology, we can create more natural forms of communication by enabling machines to understand and respond to human emotions. This has a wide range of applications, from improving customer service to increasing the effectiveness of virtual assistants and tracking mood swings.

1.1 Problem Statement

The speech emotion recognition project is a machine-human interaction that involves using computers to recognize and interpret emotions expressed in spoken language. This technology has the potential to improve various aspects of human-machine interaction, including customer service in call centers, the effectiveness of virtual assistants, and the diagnosis of mental health conditions. Besides, since speech analysis uses nonverbal cues to understand the user's emotional state, it is crucial to determine a user's emotional state using human-machine interaction systems. For this reason, data scientists have found it attractive to study human emotions. However, the evaluation of emotional data includes challenges such as collecting appropriate datasets, defining how many emotions need to be recognized, and selecting labeled data. The computer vision and artificial intelligence communities deal with systems of human-machine interaction involving multimodal information. Understanding the current emotional state of its users, due to the large number of challenges evaluated, has long been of interest to those interested in human-machine interaction.

1.2 Background or Related Work

In the projects made in the past, it has been observed that the emotions of Angry, Disgust, Fear, Happy, Sad and Surprise have been concentrated on. IEMOCAP, EMO-DB, RAVDEES and Berlin Emo DB datasets are also the most used datasets. According to the dataset used, a success rate was achieved, with an agglomeration between 60% and 80%. In addition, it has been observed that Machine Learning and Artificial Neural Networks are used while working on the examined projects. Machine learning techniques are used more frequently and a greater percentage of success is achieved. You can find the details of the 8 articles we have reviewed in the "Related Works" section.

1.3 Solution Statement

In this project, a software algorithm will be implemented to extract emotion-related features from speech signals. We will make an emotional state inference by designing a rule-based decision algorithm according to the data we will analyze later. Open-source software algorithms will be used to implement the proposed system.

The planned software language is Python and the planned library is the Librosa library. The project, which will be directed as an interdisciplinary study, will be carried out with a group of students in the Computer Engineering department.

In this context, the students of the Computer Engineering Department will apply the emotional speech recognition part of the proposed system. It is combined with a decision algorithm with algorithms related to the structure of the analyzed sound and speech. For this purpose, a theoretical study on decision algorithms for auditory recognition systems will be directed. After the implementation of this planned system, speech-related features will be returned as an output and the emotional state of the user will be estimated. The developed system will be tested on some English data sets and the performance of the system will be tested.

1.4 Contribution

Different techniques need to be used to analyze the acoustic properties of each speech and the content of spoken words. Machine learning algorithms must be used to extract certain characteristics of the speech signal, such as pitch, intensity, and spectral characteristics, and classify emotions based on these characteristics. This software system will perform auditory emotion recognition using voice data. The user can perform the emotion recognition process by uploading an existing voice recording to the system with the user interface of the system or by typing a text message of his choice into the field on the system. This has a wide range of applications, from improving customer service to increasing the effectiveness of virtual assistants to monitoring mood swings. With speech emotion recognition technology, we can create more natural forms of communication by enabling machines to understand and respond to human emotions.

1.5 Summary

The Speech Emotion Recognition System operates in a straightforward and user-friendly manner, assuming that the login system has been removed. Once the user decides to use the system, they can access the SER's website directly. They have the option to choose between "Text" or "Speech" for analysis. After making their selection, the system will store this process in the Database based on the chosen option. Subsequently, the system will process the data to extract emotional analysis. Upon completion of the data processing phase, classification will be performed. Finally, the system will present the emotional state result as a percentage and display it on the screen. The general system flow is depicted in the figure below as a flowchart.

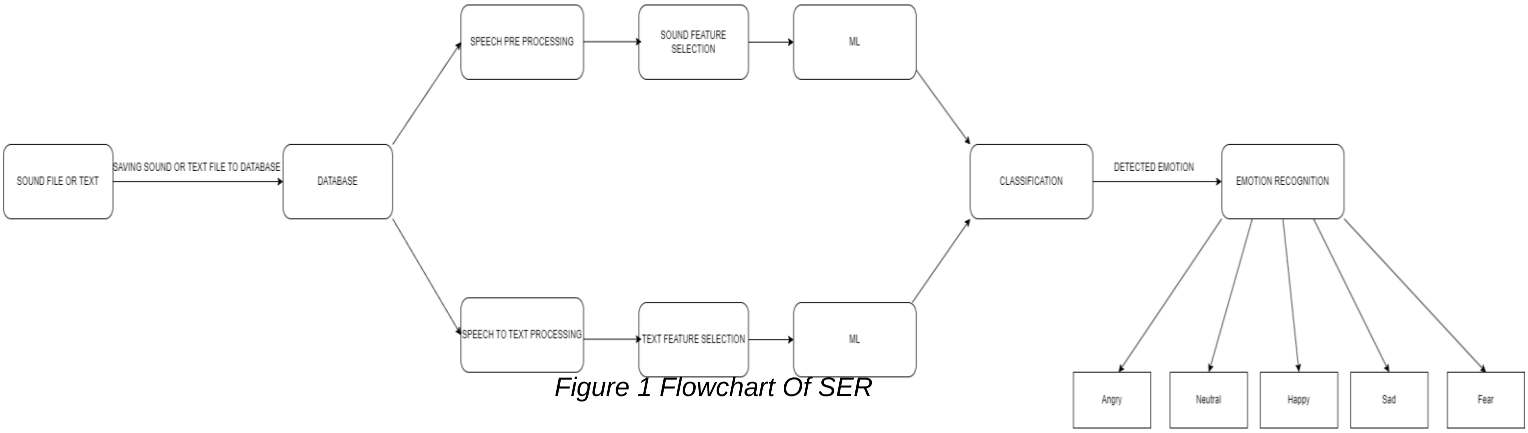


Figure 1 Flowchart Of SER

2. Literature Review

2.1 Speech Emotion Recognition:

Speech is the main and most effective way of human communication. There is no transfer of emotion just by speaking the words in a straight tone. There is a transfer of emotion in every person's speech, even if he is not aware of it. Detecting this emotional state is also a difficult task, because when a person says "awesome" they may be making an allusion to it. But it is a fact that there are physiological changes in people according to their emotional state.

There are also devices that determine mood by transferring data from things such as heart rate and blood pressure a device. But without the need for a device, it is possible to determine the mood from facial expressions and voice. There is also an increase in studies that describe emotion from face and voice.

It is important in the process of recruiting people, in the process of assessing human emotions, in matters such as detecting lies and other issues like this. There are many ways in which the various physical properties of humans can be understood by machines. Among these, facial expressions can be understood by eye movements or facial recognition systems. Or their body language and gestures are among them.

In addition to these, besides the physical movements that people have, speaking is the most effective and easy way of communication. Most of the mood analysis studies and articles are made on the act of speaking that people have rather than physical movements.

In a few articles reviewed and made, there are applications such as changing the presentation style of the instructor by evaluating the situation of the students being uninterested or bored during the education in online education and applications, thanks to the warning received by the system. At the same time, in another study conducted in the field of medicine, it was ensured that the patient's health status was examined by the patient's voice according to the patient's speech. In another way, work has been done on robots that are important for the future. In this study, robots were trained according to human emotions. The basis of this is to provide the most appropriate communication between humans and robots. It has a wide and comprehensive application area, which detects the behavior and emotional states of the people on the other side and enables them to react accordingly, in calls with call centers or in another study included with these.

Education, automobile, security, communication and health can be given as examples of the field of systems made in this field [1].

2.2 Feature Extraction:

Feature extraction plays a crucial role in the speech recognition process, and extracting valuable data from sample speech has been a crucial part of research for many years. Feature extraction helps distinguish one speech from another. It converts the raw speech signal into a dense but effective representation that is more stable than the original, from which it is possible to reconstruct it from the original signal.

Speech features can be grouped into 3 main categories.

These are continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features. This section will examine the pros and cons of each category. In addition, speech emotion recognition can be performed by combining features belonging to different categories for speech signals. The following figure shows examples of properties for each category.

2.2.1 Temporal Features

Most researchers believe that continuous features such as pitch and energy express most of the emotion in the voice.[11,13,15] According to studies by Williams and Stevens [23], low activation versus high activation affects overall energy, and energy distribution affects the duration of the pause frequency. This result has been confirmed by other studies.[18,14] Acoustic connections related to continuous sound are given below.

1) **Time - Domain Features**

   - Amplitude: The instantaneous magnitude of a sound signal.
   - Zero-crossing rate: The rate at which the sound signal crosses the zero level.
   - Energy: The energy content of the sound signal.
   - Duration: The total duration of the sound

2) **Frequency – Domain Features**

   - Mel-frequency cepstral coefficients (MFCCs): Temporal features used to represent the spectral content.
   - Spectral centroid: The spectral center frequency of the sound signal.
   - Spectral flux: The rate of change of spectral content over time.

2.2.2 Spectral Features

Experiments on humans have proven that the quality of a voice plays an important role in conveying emotion. [15,21,16,17] The sound quality seems to complement fully developed emotions in the most regular way, that is, it can lead people to action. [15] Acoustic correlations for sound quality are given below.

1) **Pitch - Related Features**

   - Spectral centroid: The spectral center frequency of the sound signal.
   - Spectral rolloff: The value of the frequency at which a certain percentage of the energy in the sound signal has passed.
   - Spectral harmonicity: A feature representing the harmonic structure of the signal.

2) **Cepstral Features**

   - Mel-frequency cepstral coefficients (MFCCs): Temporal features representing the spectral content of the sound signal.
   - Linear predictive coding (LPC) cepstral coefficients: Features used for signal prediction in the sound signal.

2.2.3 Prosodic Features

Spectral features are often chosen as a short-term representation for the speech signal. It is accepted that the emotional orientation of a sound has an effect on the distribution of spectral energy throughout the speech frequency range.[20] For example, it has been proven by studies that expressions containing happiness have high energy in the high-frequency range, and expressions containing sadness have low energy in the same range.[10,19]

1) **Frequency - Domain Features**

   - Pitch range: The frequency range of the sound signal.

2) **Duration - Related Features**

   - Phoneme duration: Total duration of phonemes in the sound signal.
   - Pause duration: Duration of pauses in the sound signal.
   - Speech rate: Rate of speech in the sound signal.

3) **Intensity - Related Features**

   - Phoneme duration: Total duration of phonemes in the sound signal.
   - Pause duration: Duration of pauses in the sound signal.
   - Speech rate: Rate of speech in the sound signal.

2.3 Related Works:

Related Works: In an emotional speech corpus for the Urdu language was designed and developed in this study. Different machine-learning techniques were used to identify emotions from Urdu speech signals [2]. Five Urdu sentences—one each for happy, sad, angry, disgusted, and neutral—were simulated into five different emotional states. The highest overall recognition accuracy for the disgust emotion for males, females, and the entire dataset was 72.5% with "k-NN," 68.5% with "one-against-rest classifier," and 66.2% with "k-NN." The maximum highest recognition accuracy for the dataset without the disgust emotion was 82.5% with "k-NN," 78.5% with "one-against-rest classifier," and 76.5% with "k-NN" for male, female, and the whole dataset, respectively.

In this paper, The Ryerson Multimedia Lab (RML) emotion database was used, which comprises 241 audiovisual and emotional expression examples in English [4]. There are six fundamental human emotions: anger, disgust, fear, happiness, sadness, and surprise. Began adjusting the basic AE's settings to improve identification rates when the RBF kernel of the SVM classifier was used. We acquire a superior recognition rate equal to 70.37% when the number of units in the hidden layer is 35 after a series of trials in which the number of units is varied.

Leila et al [7] described an automatic speech emotion recognition (SER) system that classifies seven emotions using three machine learning methods (MLR, SVM, and RNN). In order to offer a mix of these properties, two types of features (MFCC and MS) were derived from two separate acted databases (Berlin and Spanish databases). The machine learning models were trained and tested for their ability to distinguish emotional states based on these characteristics. When speaker normalization (SN) and feature selection (FS) are applied to the features, according to SER, the recognition rate of the Berlin database is 83% accurate across the board. These results showed that RNN frequently performs better with more data and that it has the constraint of having very long training cycles. As a result, we came to the conclusion that, as compared to RNN, the SVM and MLR models had a better chance of being used practically with little amounts of data.

In this paper, the IEMOCAP database was used [6]. This dataset was used firstly only for audio tests, then only for text tests and finally by combining audio + text for 6 emotions(Angry, Happy, Sad, Fear, Surprise, Neutral). This dataset is used for RF, XGB, SVM, MNB, LR, MLP, LSTM, ARE and E1. Here only the highest accuracy for Audio was achieved with E1 (Ensemble (RF + XGB + MLP)) at 56.6. For Text only, the highest accuracy was achieved with the TRE (Text-Recurrent Encoders) as 65.5. In the last step, when working on the model again by setting Audio+Text, the highest accuracy was obtained with MDRE (Multimodal Dual-Recurrent Encoders) of 75.3.

Clustering-Based Speech Emotion Recognition (2020) [3]. In this paper, this research outlined an innovative strategy for SER to enhance the Boost recognition accuracy while cutting down on model costs and processing times. On the other hand, we proposed a new method to choose a more effective speech sequence using the K-mean clustering approach based on RBF and changing By using the STFT technique, it may be turned into spectrograms. Hence, We identified the prominent and discriminative characteristics from voice signal spectrograms by using the "FC-1000" Resnet layers in the CNN model are added before it is normalized. Using mean and standard deviation, we can eliminate the variation. After normalization, we feed these discriminative features to deep BiLSTM to learn the hidden information, recognize the final state of the sequence, and classify the emotional state of the speakers. To test the system's resilience, we assessed it using three common datasets: IEMOCAP, EMO-DB, and RAVDESS. We increase the recognition accuracy for the IEMOCAP dataset to 72.25%, the EMO-DB dataset to 85.57%, and the RAVDESS dataset to 77.02%.

Speech emotion recognition with deep convolutional neural networks [5]. In this paper, the IEMOCAP, EMO-DB, and RAVDEES databases were used in this paper. Afterward, we used an incremental approach to enhance classification accuracy by tweaking our baseline model. Unlike some earlier methods, none of the suggested models require translation to visual representations before working with raw audio data. Our best-performing model exceeds previous frameworks for RAVDESS and IEMOCAP, establishing a new state-of-the-art. It performs better than all prior efforts for the EMO-DB dataset, with the exception of one, and compares favorably with that one in terms of generality, simplicity, and application. Specifically, more precise, the suggested framework completes speaker-independent audio classification tasks with 71.61% for RAVDESS with 8 classes, 86.1% for EMO-DB with 535 samples in 7 classes, 95.71% for EMO-DB with 520 samples in 7 classes, and 64.3% for IEMOCAP with 4 classes. Table 1.1 shows the results of different datasets.

Table 1.1 Accuracy results for Related works

Papers	Dataset	Emotions	Technique	Accuracy(%)
An Urdu Speech Corpus For Emotion Recognition (2022) [2]	Urdu Emotional Speech Dataset	Angry, Happy, Sad, Neutral	k-NN (with disgust) k-NN (without disgust)	72.5 82.5
Clustering-Based Speech Emotion Recognition (2020) [3]	IEMOCAP EMO-DB RAVDEES	Angry, Happy, Sad, Fear, Surprise, Neutral	CNN + LSTM	72.25 85.57 77.02

Speech Emotion Recognition with Deep Learning (2020) [4]	RML Dataset	Angry, Disgust, Fear, Happy, Sad, Surprise	Basic AE with SVM	72.83
			Stacked AE with SVM	74.07
Speech emotion recognition with deep convolutional neural networks (2020) [5]	IEMOCAP	Angry, Disgust, Fear, Happy, Sad, Surprise	CNN	64.30
	EMO-DB		LSTM	71.61
	RAVDEES			86.1
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [6]	IEMOCAP (Audio Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF	56.0
			XGB	56.6
			SVM	33.7
			MNB	31.3
			LR	33.4
			MLP	41.0
			LSTM	43.6
			ARE (4-class)	56.0
			E1 (4 -class)	56.3
			E1	56.6
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [6]	IEMOCAP (Text Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF	62.2
			XGB	56.9
			SVM	62.1
			MNB	61.9
			LR	64.2
			MLP	60.6
			LSTM	63.1
			TRE (4-class)	65.5
			E1 (4 -class)	63.1
			E1	64.9
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [6]	IEMOCAP (Audio+Text)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF	65.3
			XGB	62.2
			SVM	63.4
			MNB	60.5
			LR	66.1

			MLP	63.2
			LSTM	64.2
			ARE (4-class)	75.3
			E1 (4 -class)	70.3
			E1	70.1
Leila et al. (2019) [7]	Berlin Emo DB	Angry, Disgust, Joy, Neutral Happy, Sad, Surprise	SVM, MLR	83.0

2.4 Speech To Text - Audio Recognition:

Neural networks[8] play a major role in the use and definition of human behavior and movements in large areas such as artificial intelligence, machine learning, and deep learning. In this study, we will try to describe the behaviors coming from hostile environments with the information we get from neural networks. While studies on speech recognition are the most common in the world, studies on voice are less. In an automatic speech, a sound wave is sent and an inference is made about what language and how it is and converted into text. In this article, we will examine the study of text from a speech on a transcription neural network.

Several suitable datasets are used to investigate the effects of neural networks on speech-to-text. The Mozilla Common Voice dataset[9] is made available to evaluate the effectiveness of attacks made in this article. In the study, randomly sent data (a) was used to find the wrong ones, while (b) was used to find the correct ones. In addition to this, another data set usage was presented to the project. In this dataset, the input waveform is converted to sound 50 times per second and gives the character probability distribution in this project. That means placing it at 50 characters per second. Thus, it is possible to convert short audio clips to text. In such an attack, it is effective even though it requires an average of 18dB distortion. Another dataset, it is concerned with finding the target expression of a random non-speech voice sample. In this study, there was no need to apply technical innovation, instead, only random sound waves were sent. The effectiveness of this sent sound wave was obtained by collecting five-second clips from classical music that does not contain any speech and target sentence parts in the common audio dataset. As a result of this study, it was observed that distortion up to a maximum of 20 dB occurred. In the last part of the study, data was obtained by providing silence. And no data set is used in this part. A hostile noise was added that caused the study not to write any type of text, and no results were obtained.

The thread model was used as a model in the study. In this model, given a sound wave sample a and target y, the goal is to generate another sound wave. Here, if the output completely overlaps with the target without typos, the result is successful. Models and parameters from previous studies are defined as white boxes. It is a kind of competitor threat model. The threats to later work are the black box.

In the study, a distortion metric is used to understand and solve a possible distortion. They measure distortion in decibels. It is intended that the disruption be minimal and that people do not understand it.

3. Software Requirements Specification

3.1 Introduction

3.1.1 Purpose

The purpose of this document is to introduce and detail the system called Speech Emotion Recognition. This system makes an emotion analysis from voice data and produces a result among the options "anger", "excited", "sadness", "frustration", "happiness", "neutral". It is planned to be used both in a system that normal people can use in their daily lives and in a system where priority is given to the emotional state of the customer, especially for those working in the call center. This document also mentions the details and requirements of the project.

3.1.2 Scope of Project

Speech is the main and most effective way of human communication. There is no transfer of emotion just by speaking the words in a straight tone. There is a transfer of emotion in every person's speech, even if he is not aware of it. Detecting this emotional state is also a difficult task, because when a person says "awesome" they may be making an allusion to it. But it is a fact that there are physiological changes in people according to their emotional state. There are also devices that determine mood by transferring data from things such as heart rate, blood pressure, blood pressure to a device. But without the need for a device, it is possible to determine the emotion from facial expressions and voice.

The Speech Emotion Recognition system, on the other hand, aims to perform an emotion analysis only through voice without the need for a machine. Through this system, this system can be used in many areas in daily life. For example, in job interviews, it is aimed to understand the emotional state of a candidate from his voice. It is important in the process of recruiting people, in the process of assessing human emotions, in matters such as detecting lies and other issues like this.

Another area that can be used is education. It can be understood which student is nervous, happy or angry from the voices of the students during the online education. In this way, more or less comments can be made about the efficiency level of the course.

Another area is customer service. Customer service is one of the most important factors that increase or decrease the value of a company on the consumers' side. For this reason, it is aimed to further improve and reach the best levels of communication between customers and businesses around the world. With the increasing e-commerce sector, many people are connected to customer service. Most of the people who connect to customer service are connected to customer service because they have problems. And while waiting in line to connect, many customers get nervous or angry. The Speech Emotion Recognition system is also a candidate to be one of the ideal solutions for this scenario.

It aims to connect angry customers to customer service by measuring the emotional state of the customers and making a ranking. Another of the potential uses of the Speech Emotion Recognition system.

In the Speech Emotion Recognition system, there are 2 actors, the admin and the user whose voice will be analyzed. In order for the system to be used and for the voice analysis to start, the voice data must be processed by the system. The owner of this voice data, usually the user, is aware that his voice is being processed and must give the necessary data privacy permission.

The Speech Emotion Recognition system will produce a result after the required permission and voice processing is complete. The options for these results are "anger", "excited", "sadness", "frustration", "happiness", "neutral".

3.1.3 Glossary

Term	Definition
SER	The main project is an abbreviation of the name Speech Emotion Recognition.
Machine Learning	Machine learning (ML) is a field of understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.
Librosa	Librosa is a Python package for music and audio analysis.
Python	Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.
Human-Computer Interaction	Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology

3.1.4 Overview of the Document

The second part of the document includes titles and details such as External Interface Requirements, Functional Requirements, Software System attributes, Safety Requirement.

The Requirement Specification chapter is written for software developers and details of the functionality of the system are described in technical terms.

The remainder of the document covers the details of the Speech Emotion Recognition system in a technical way.

3.2 Overall Description

3.2.1 Product Perspective

Speech Emotion Recognition is a project that is designed to be used in daily life, education, job interviews and many other areas, and produces a result by analyzing emotion from voice. The basic dynamic of the project is to take a voice file as input and process it with certain methods and produce a result. 6 different results can be produced according to the content of this produced voice file. With this project, it is to do emotion analysis only with sound, without the need for any tool or machine.

3.2.1.1 Development Methodology

Product Features

The software described in this SRS document will be used to help people understand their mood from speech or text that users write. Since it provides the continuation of human and speech life, it is a software that can adapt to any desired area.

3.2.2. User Characteristic

3.2.2.1 Participants

- Student
- Customer
- Worker
- Company
- Lecturer
- Employer

3.3 Requirements Specification

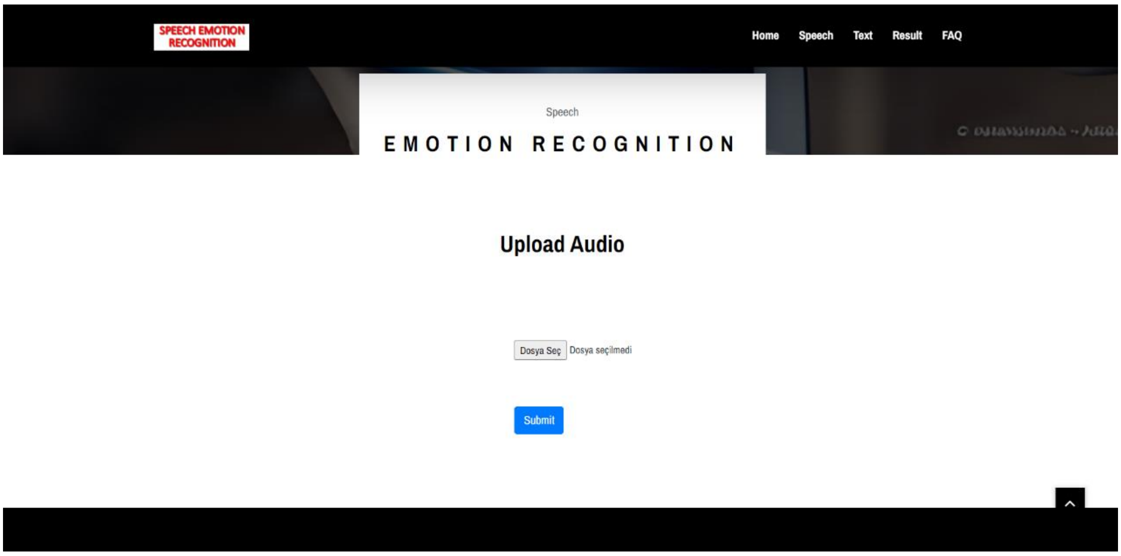
3.3.1 External Interface Requirements

3.3.1.1 User Interfaces

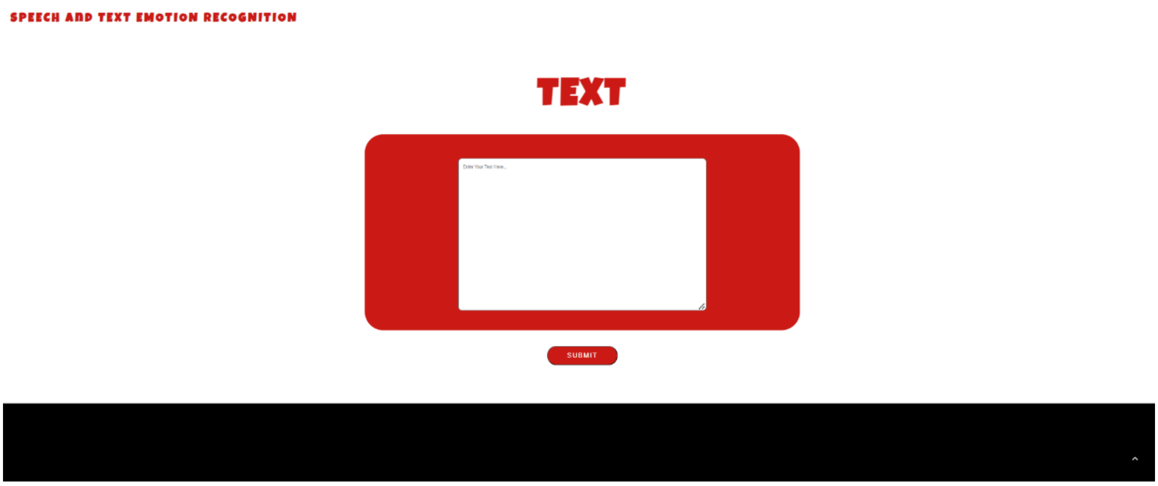
Our software works actively on all platforms with Python installed. Here's what the user can do with the interface:

- **Files can be added externally (voice or text)**

- **Files can be added externally (voice or text)**



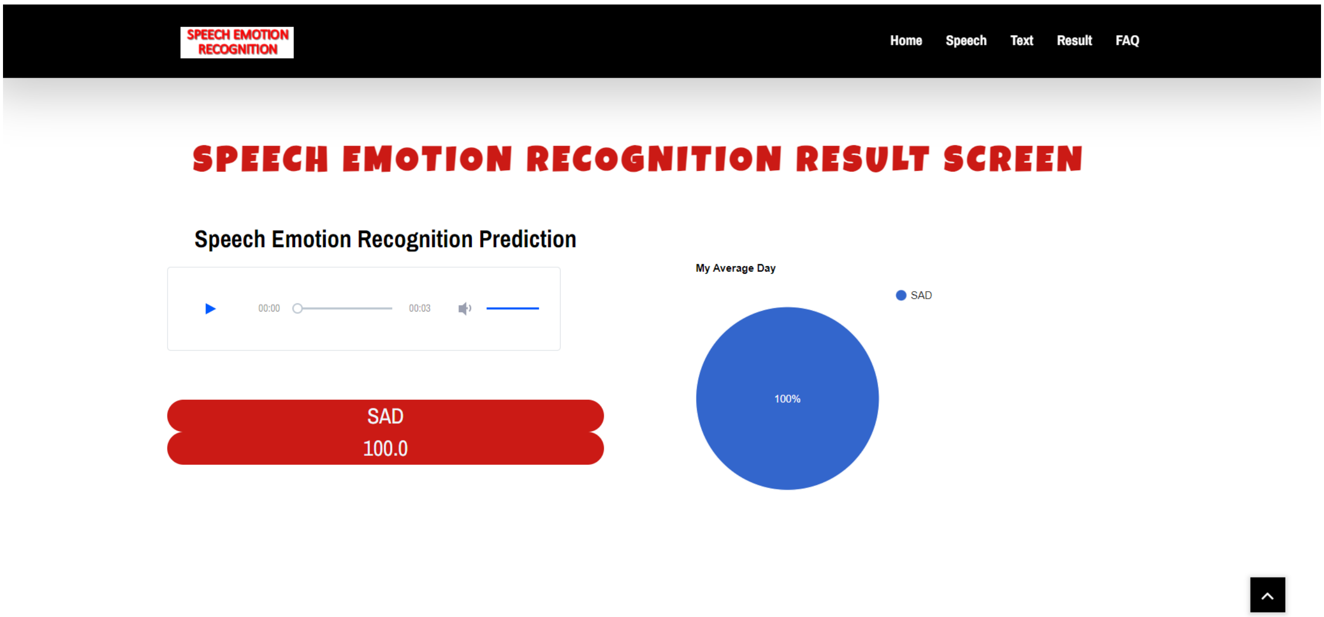
- Can only manually write the text he wants to find his emotion.



SPEECH AND TEXT EMOTION RECOGNITION

TEXT

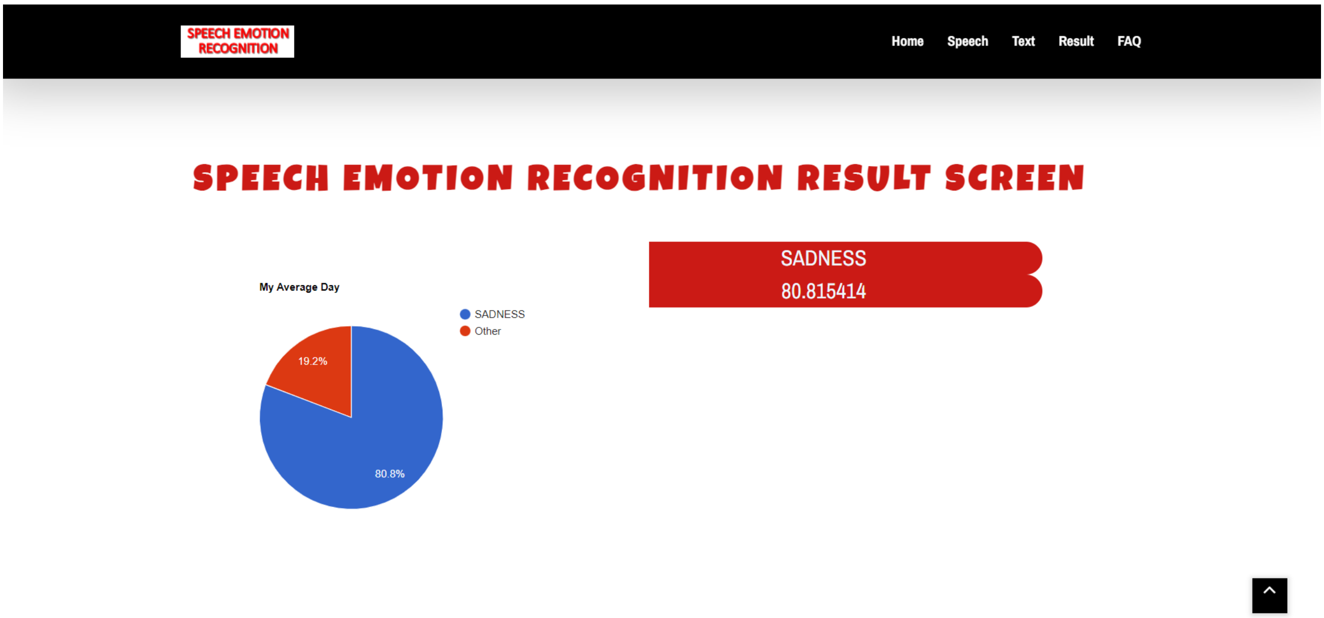
- After people upload the audio files they have previously recorded or after the text they havewritten, they enter this page and select the speech or text page they have applied, and the result is evaluated. As a result of the evaluation, there is a screen that concludes which of the 2 basic emotions.



SPEECH EMOTION RECOGNITION RESULT SCREEN

Speech Emotion Recognition Prediction

SAD
100.0

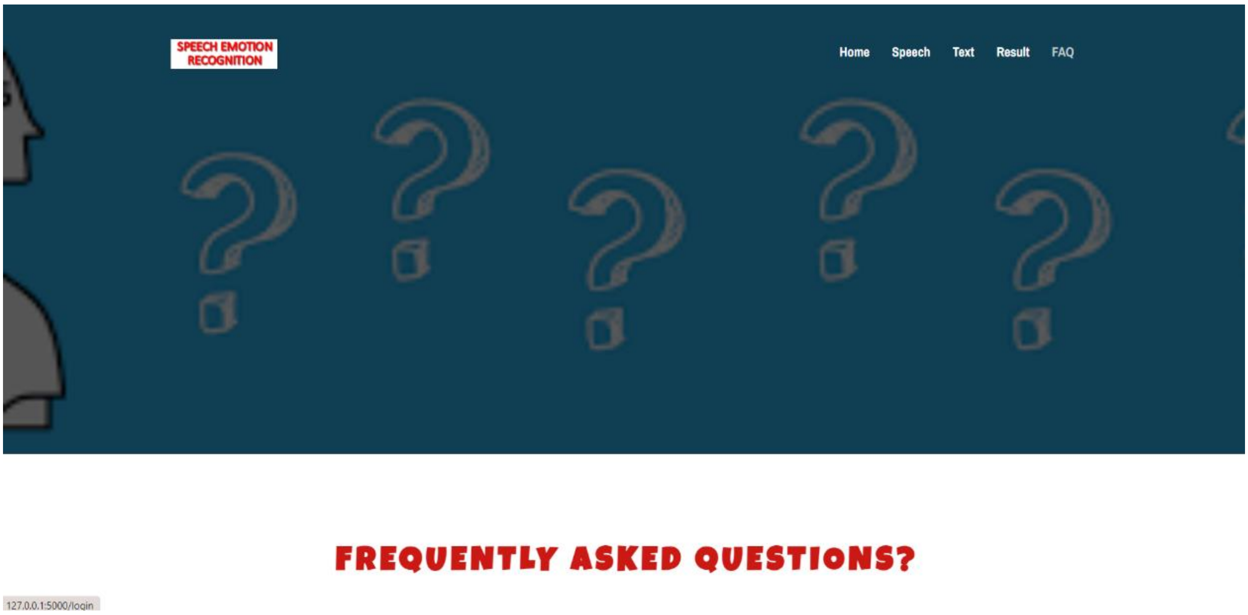


SPEECH EMOTION RECOGNITION RESULT SCREEN

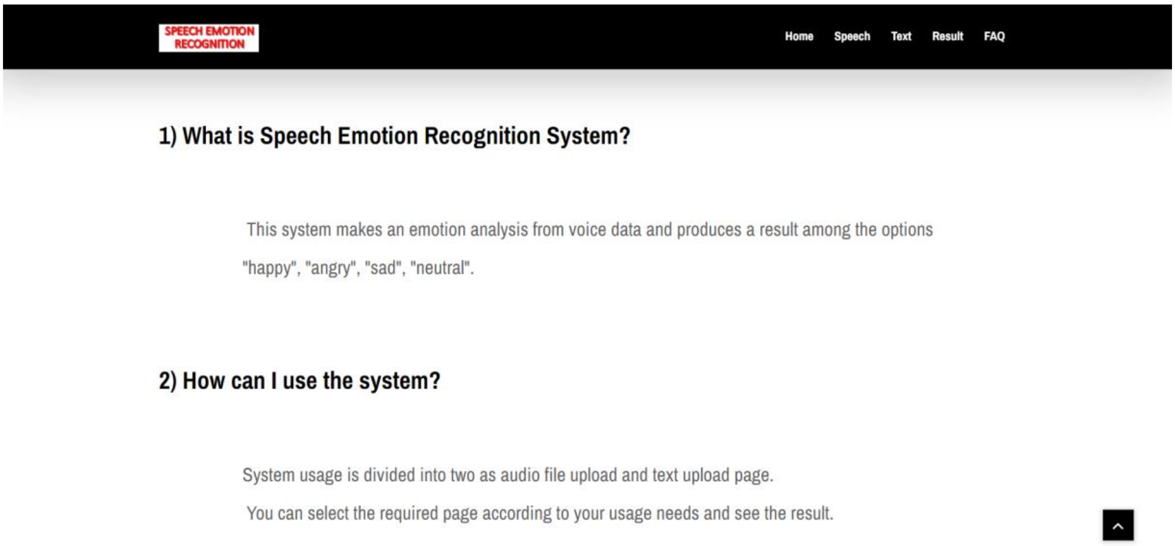
SADNESS

80.815414

- **This page contains information about using the system. When the user has a question about the system, it allows them to take advantage of this page and continue the process.**



FREQUENTLY ASKED QUESTIONS?



1) What is Speech Emotion Recognition System?

This system makes an emotion analysis from voice data and produces a result among the options "happy", "angry", "sad", "neutral".

2) How can I use the system?

System usage is divided into two as audio file upload and text upload page.

You can select the required page according to your usage needs and see the result.

3.3.1.2 Hardware Interfaces

Computer must have one USB port for video recording. It also requires one microphone input for voice recording.

3.3.1.3 Software Interfaces

The Computer must have the libraries associated with Python. Example: Librosa.

Visual Studio Code: To implement the project, we choose HTML, CSS, and Javascript languages.

3.3.1.4 Communications interfaces

Internet connection is required to run this software.

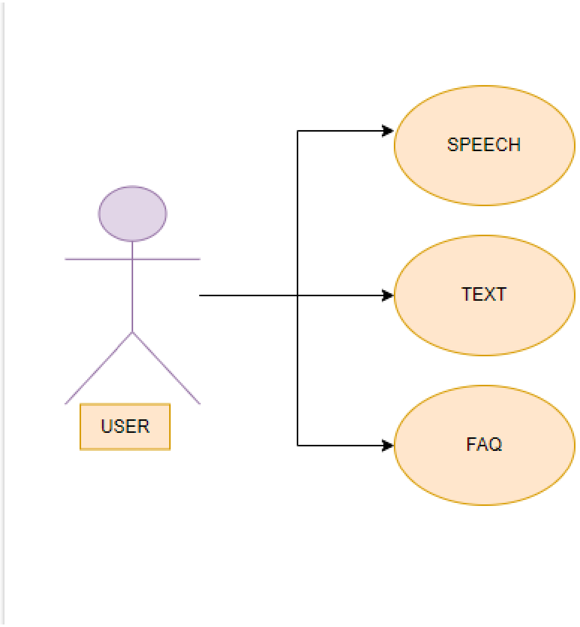
3.3.2 Functional Requirements

3.3.2.1. Home Page Use Case

Use Case:

- Speech
- Text
- FAQ

Use Case Diagram:



Description:

The home page schema has basic operations on what the "User" role can access. The user encounters this menu when user first logs into the system. User can use "Speech", "Text", and "FAQ" operations.

Initial Step-by-Step Description:

1. The user must enter the home screen.

2. The user should choose the theme for emotion recognition. If it is speech emotion recognition, "Speech", if it is to recognize emotion from text, click the "Text" button.

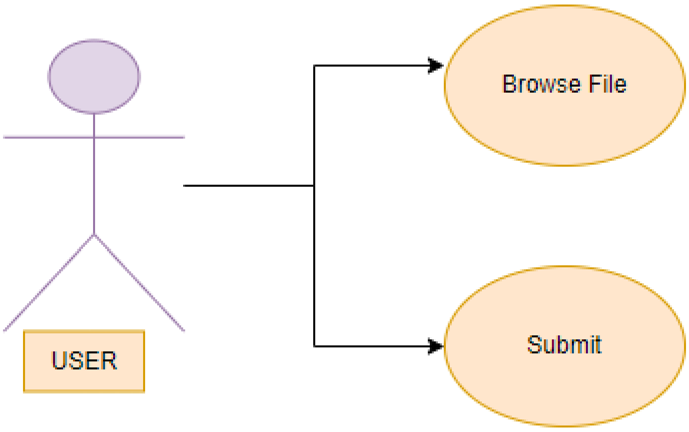
3. The user can use the "FAQ" button for questions that come to mind. This screen contains frequently asked questions.

3.3.2.2. Speech Page Use Case

Use Case:

- Browse File
- Submit

Use Case Diagram:



Description:

User will use this screen for emotion recognition from audio file. From this screen, the user can upload and delete audio files and submit the audio for emotion recognition. From this screen, the user can use "Browse File” and "Submit" operations.

Initial Step-by-step Description:

1. The user must first click the "Browse File" button to select the audio file from his computer. If user does not want to select an audio file from his computer.

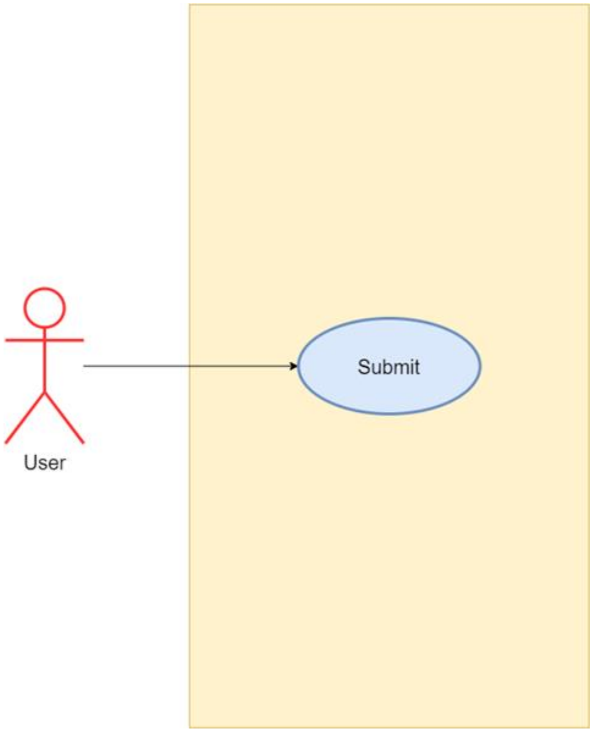
2. After uploading the audio file, the user should click the "Submit" button to see the results.

3.3.2.3. Text Page Use Case

Use Case:

- Submit

Use Case Diagram:



Description:

User will use this screen for emotion recognition from text file. The user can write text from this screen and submit the text for emotion recognition. The user can use the "Submit" operation from this screen.

Initial Step-by-step Description:

1. The user must first fill in the text part with the text for emotion recognition.

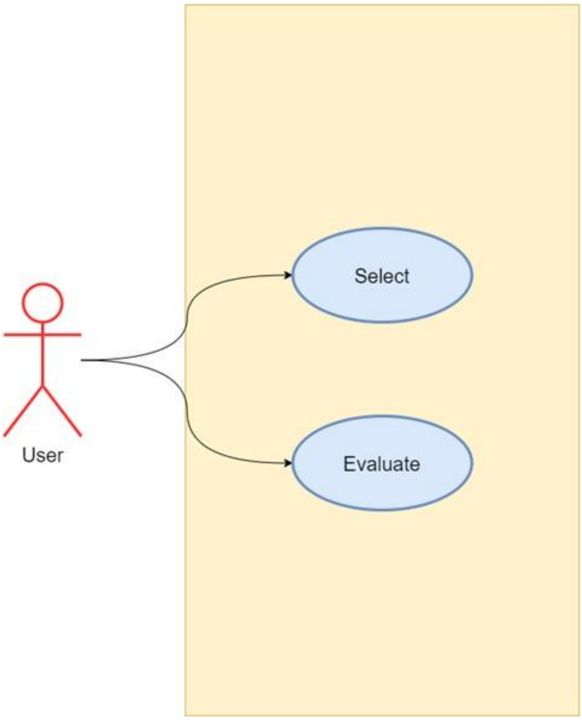
2. After typing the user's text, he should click the "Submit" button to see the results.

3.3.2.4. Result Page Use Case

Use Case:

- Select
- Evaluate

Use Case Diagram:



Description:

Finally, the user uses this screen to see the results of the uploaded text or audio file.

Initial Step-by-Step Description:

1. The user must first select "Text" if he has uploaded a text file, or "Speech" if a user has uploaded an audio file, and then click the "Select" button.

2. After making the selection, the user can click the "Evaluate" button to see the results of the uploaded audio or text file.

3.3.3 Performance Requirement

Since the Speech Emotion Recognition System does not require high-performance hardware, a strong internet connection will be sufficient. In addition, a personal computer suitable for daily use will be sufficient.

3.3.4 Software System Attributes

3.3.4.1 Availability

- The system will work on Windows-based and Android-based operating systems.

3.3.4.2 Ease of Use

- This system is being created as a user-oriented project. It should be simple for the user to use. This simplicity should be the same for the first user and should be the same for the frequent user. The interface we have created will be simple, user - friendly and understandable as a result.

3.3.4.3 Reliability

- The reliability of the System depends on the sound quality or the accuracy of the text to be analyzed.
- The data provided by the user will be used to compare with the result and to measure the reliability.
- If enough data are collected, the most recent machine learning algorithms should be able to reliably utilize the user's data.

3.3.4.4 Maintainability

- The administrator will review and update the training and testing files throughout three months.
- The maintenance period is not a matter because the reliable version is always run on the server which allows users to access past summarization.

3.3.4.5 Security

- Since no data is acquired and saved from the run time, there is no adaptability requirement.

3.3.4.6 Scalability

- Since only one participant uses the system at a time, there is no scalability requirement.

3.3.5 Safety Requirement

Users' data will be retained in the database so it can be utilized for system development, upgrades, and maintenance, which are scheduled to be carried out every three months. The system will be strengthened and its stability will be increased using the data. A pre-approval text will be declared at the stage of system membership, declaring that the data will only be used for system improvement, in order for these data to be utilized as the consent of persons.

4. Software Design Document

4.1 Introduction

4.1.1 Purpose

The purpose of this Software Design Document (SDD) is explaining the system which is called audio-visual emotional recognition. This system's goal is to provide recognition of emotions from speech and text.

We define the SER system as a collection of methodologies that process and classify speech signals to detect embedded emotions. Such systems can be used in speech analytics with interactive visitor agents, as well as in a variety of operational areas. In this study, we try to determine the underlying emotions(happy", "angry", "sad", ”fear” and "neutral.) of recorded speech by examining the acoustic properties of the auditory data of the recordings.

4.1.2 Scope

Emotions are an integral part of human behavior and an inherited characteristic of all modes of communication. We thought of your experience reading detecting different emotions which makes us more rational and understanding. However, while machines can easily understand text, audio, or video information, they are still far behind in accessing the depth of content. Additionally, voice sentiment analysis has many applications in various fields such as healthcare, banking, defense, call center and IT. In this project, user will make the upload as a voice or text file or by taking a voice recording directly with the microphone, and the uploaded audio file and the recording will be the mood in the voice, and the uploaded text will be the mood analysis from the words used.

4.1.3 Glossary of SDD

Term	Definition
SER	The main project is an abbreviation of the name Speech Emotion Recognition
User	People who want to use application
Machine Learning	Machine learning (ML) is a field of understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.
Librosa	Librosa is a Python package for music and audio analysis.
Python	Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.
The remaining chapters and their contents are listed below.	The remaining chapters and their contents are listed below.

4.1.4 Overview of Document

The remaining chapters and their contents are listed below.

Section 2 is an architectural design that describes the development stages of the project. Also included is a project system and architectural design class diagram that describes actors, exceptions, basic sequences, priorities, preconditions, and postconditions. Additionally, this section contains an activity diagram for the Scenario Generator.

Section 3 is realization of the use case. This section presents and describes a block diagram of a system designed according to the use cases in the SRS document.

Section 4 is about the environment. This section showed a sample frame of the environment from the prototype to illustrate the scenario.

4.1.5 Motivation

We are a group of final-year students of computer engineering department. As a group, we chose the project of making mood analysis from sound and text. In this project, we aimed to include technologies in the fields of education, social and humanity. We are interested in speech processing, image processing and artificial intelligence fields.

4.2 Design Overview

4.2.1 Description of Problem

One problem with speech emotion recognition is that it can be difficult to accurately perceive and identify emotions in speech because of the complexity and variability of human emotions. Different people can express the same emotion in different ways, and the same person can express different emotions in similar ways.

Another problem is that speech can be affected by a variety of factors that can distort expressed emotions. For example, background noise, accent, and speaking style can affect perceived emotion in speech.

4.2.2 Technologies Used

This software will communicate with Python to handle image, audio, and text processing. Machine learning algorithms, artificial neural networks will be used to train the model. Librosa, a Python package for music and audio analysis, will also be used. Google Speech Recognition, designed by Google, will be used to translate voice to text. Java Script and React will be used for the design. This software can run on Microsoft Windows, Linux and macOS.

4.2.3 Design Summary

To create a speech emotion recognition system, the following steps are followed:

Collect and preprocess data: The first step in building a speech emotion recognition system is to collect a large dataset of spoken language samples associated with the emotions they are expressing. This may involve labeling audio recordings with appropriate emotion categories (such as happy, sad, angry, neutral). Data may also need to be cleaned and preprocessed to eliminate noise or other defects.

Extract features: After the data has been collected and preprocessed, the next step is to extract relevant features from the audio recordings or transcriptions. This includes continuous speech features (e.g. pitch and energy), voice quality features (e.g. signal amplitude, energy, duration, phrase, phoneme, word, feature boundaries, temporal structures), spectral-based speech features (e.g. expressions containing happiness have high energy in the high frequency range, and expressions containing sadness have low energy in the same range) or nonlinear TEO-Based features (e.g., to find stress in speech).

Train a classifier: With the extracted features, the next step is to train a machine learning or deep neural network classifier to identify the emotions expressed in spoken language samples. In the studies, using the deep neural network classifier, 71.61% success was achieved in the RAVDESS dataset, 86.1% in the EMO-DB dataset, and 64.3% in the IEMOCAP dataset. In another study, using machine learning, 83.0% success was achieved in the Berlin Emo DB dataset and 70.37% in the RML dataset.

Evaluate the system: Once the classifier has been trained, it is important to evaluate its performance on a separate test dataset to ensure it can correctly classify emotions in unseen data.

4.3 Architecture Design

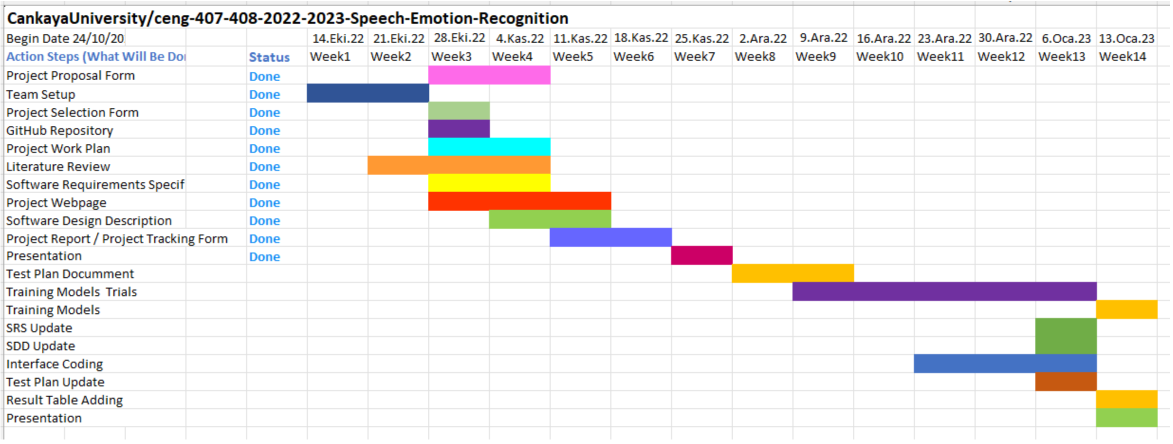
4.3.1 Simulation Design Approach

We preferred Agile Methodology, one of the software development methodologies, to develop the project. The Agile methodology is a way to manage a project by dividing it into several phases. It includes improvement at every stage. The agile methodology encourages ongoing testing and development throughout the project's software development lifecycle. One of the simplest and most efficient methods for translating a vision for a business need into software solutions is the agile software development methodology. The term "agile" is used to describe methods for developing software that involves ongoing planning, learning, and improvement, teamwork, evolutionary development, and early delivery. It promotes adaptable reactions to change.

We chose the Scrum method, one of the Agile methodologies, for our project. Scrum is iterative and incremental. In Scrum, main work is divided into sprints which should be completed within a certain period of time. Every Sprint includes tasks.

The development team and the project's mentor should hold a meeting of at least 40 minutes each week. Scrum has three main roles which are product owner, scrum master and development team. Product owner delivers the requirements, scrum master manages the development team. Development team is the group of developers who work on the project according to schedule. There are some benefits to using Scrum. The first benefit is that short sprints and constant feedback make it simpler to deal with changes. Another benefit is problems can be handled quickly due to weekly meetings. Also, it makes it possible to create quality products in scheduled time.

We also kept the work done and the progress made at the scheduled times on a Project Work Plan table (Figure 1).



CankayaUniversity/ceng-407-408-2022-2023-Speech-Emotion-Recognition

Begin Date 24/10/20		14.Eki.22	21.Eki.22	28.Eki.22	4.Kas.22	11.Kas.22	18.Kas.22	25.Kas.22	2.Ara.22	9.Ara.22	16.Ara.22	23.Ara.22	30.Ara.22	6.Oca.23	13.Oca.23
Action Steps (What Will Be Doı	Status	Week1	Week2	Week3	Week4	Week5	Week6	Week7	Week8	Week9	Week10	Week11	Week12	Week13	Week14
Project Proposal Form	Done			🟪	🟪										
Team Setup	Done	🟦	🟦												
Project Selection Form	Done			🟩											
GitHub Repository	Done			🟪											
Project Work Plan	Done			🟦	🟦										
Literature Review	Done		🟧	🟧	🟧										
Software Requirements Specif	Done			🟨	🟨										
Project Webpage	Done			🟥	🟥	🟥									
Software Design Description	Done				🟩	🟩									
Project Report / Project Tracking Form	Done					🟦	🟦								
Presentation	Done							🟥							
Test Plan Dorcumment									🟨	🟨					
Training Models Trials										🟪	🟪	🟪	🟪	🟪	
Training Models															🟨
SRS Update														🟩	
SDD Update														🟩	
Interface Coding												🟦	🟦	🟦	
Test Plan Update														🟫	
Result Table Adding															🟨
Presentation															🟩

Figure 2 Project Work Plan

4.3.1.1 Class Diagram

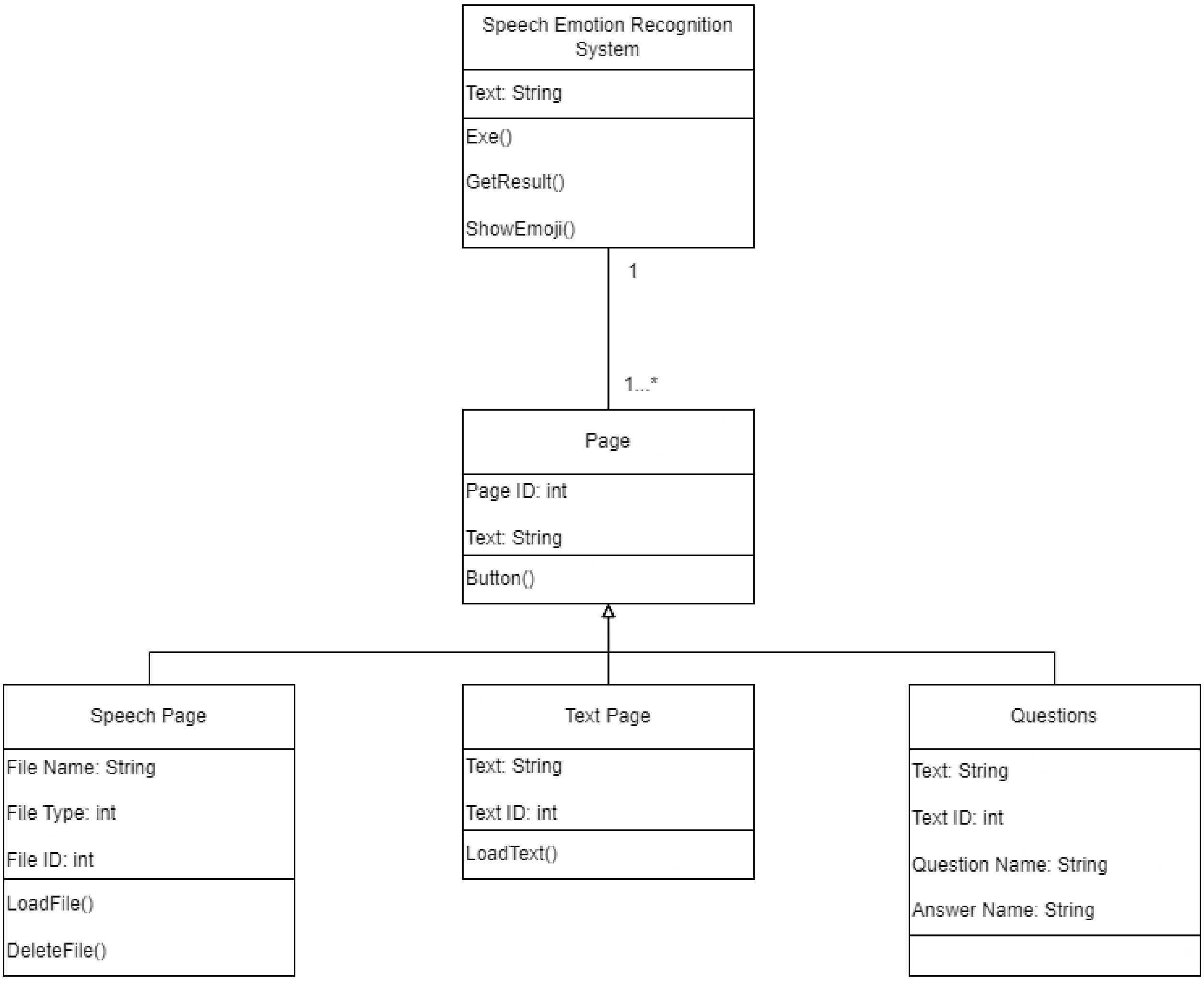


Figure 3 Class Diagram of Speech Emotion Recognition project

The figure (Figure 3) above gives information about the connections and interactions within the system. SER (Speech Emotion Recognition) includes other classes, that is, all systems included in the project. The profile class represents all users using the system. The Text class expresses and contains the text written inside itself. In the Speech section, there is the audio file uploaded by the user. On the Questions page, there are brief information about the use of the system and questions about the use of this system. In the Speech emotion Recognition class, there are expressions containing the result of the text or audio file uploaded to the Speech Emotion Recognition system.

In addition to these, the exact measurement or calculation given by the system is directly indicated with a percentile expression. However, there is also information about the content of the uploaded file.

4.3.2 Architecture Design of Speech Emotion Recognition

4.3.2.1 Main Page

Summary: On the main page, the user can choose between the audio file loading screen or the page to enter text.

Actor: User

Precondition: User must be open the system.

Basic Sequence:

1. The user can choose what he wants from the "Speech" or "Text" section on the main page.
2. The user is directed to the relevant page according to the selection he has chosen.
3. User can exit from the system by selecting the exit button.

Exception: None

Post Conditions: None

Priority: Medium

4.3.2.2 Speech Page

Summary: User should upload the audio file that he wants to analyze to the system.

Actor: User

Precondition: User must be select "Speech" from the main page.

Basic Sequence:

1. The user must upload the desired file to the system by dragging it from the "File Upload" section or selecting it from his own computer.
2. After the installation to the system is finished, he should press the "Submit" button.
3. If he has uploaded a wrong file, he should delete the uploaded file by pressing the trash icon.
4. After pressing the "Submit" button, the system will redirect to the "Result" page.

Exception: None.

Post Conditions: After the file is uploaded and submitted, the system should process this file.

Priority: High

4.3.2.3 Text Page

Summary: The user should write the text that he wants to analyze into the system.

Actor: User

Precondition: User must be select "Text" from the main page.

Basic Sequence:

1. The user should write the text that he wants to write in the relevant part of the "Text" page.
2. After the writing process is finished, he should press the "Submit" button.
3. After pressing the "Submit" button, the system will redirect to the "Result" page.

Exception: None.

Post Conditions: After the text is written, the system must process it.

Priority: High

4.3.2.4 Result Page

Summary: The user will see the result of the text he has written on this page or the result of the audio file he has uploaded.

Actor: User

Precondition: The user has written the text that user wants to be processed from the "Text" section or uploaded the file that he wants to be processed from the "Speech" section.

Basic Sequence:

1. The user should make a choice according to which type he wants to progress on this page.
2. After completing his selection, he should press the "Evaluate" button.
3. After pressing the "Evaluate" button, the system will process this file and generate a result.
4. This result will be displayed to the user on the page with emojis.

Exception: None.

Post Conditions: None

Priority: High

4.3.3 Activity Diagram

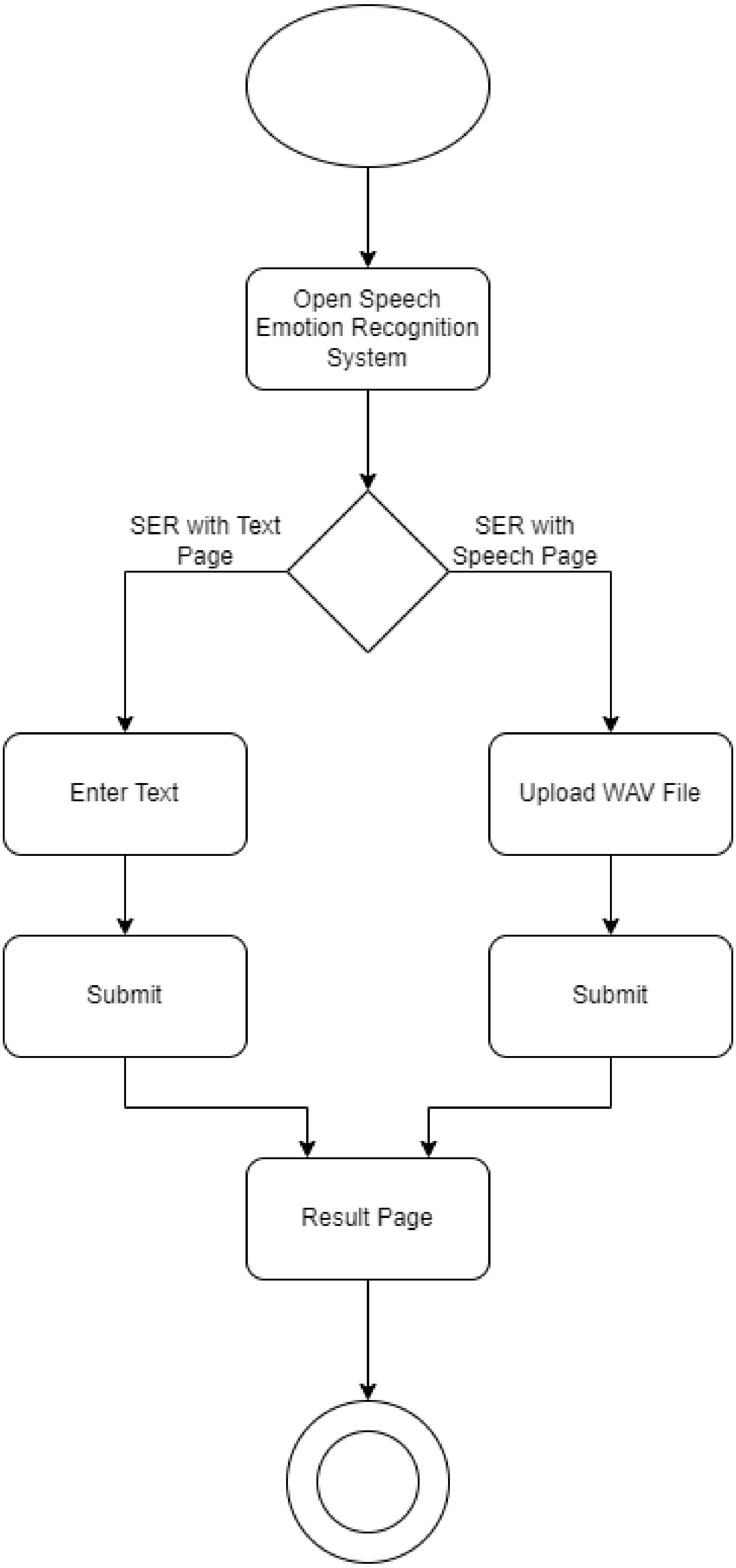


Figure 4 Activity Diagram of SER

Figure 4 shows how the SER works as an activity diagram. The user will be able to use the SER system by following a simple way. The user must decide whether he wants to do sentiment analysis with Text or Voice. After making this decision, the user should either upload an audio file or enter text by following the relevant section. After completing this process, he should go to the result page and evaluate the part he chose. After this process, the sound analysis will be completed and the result will be displayed.

4.4 Use Case Realizations

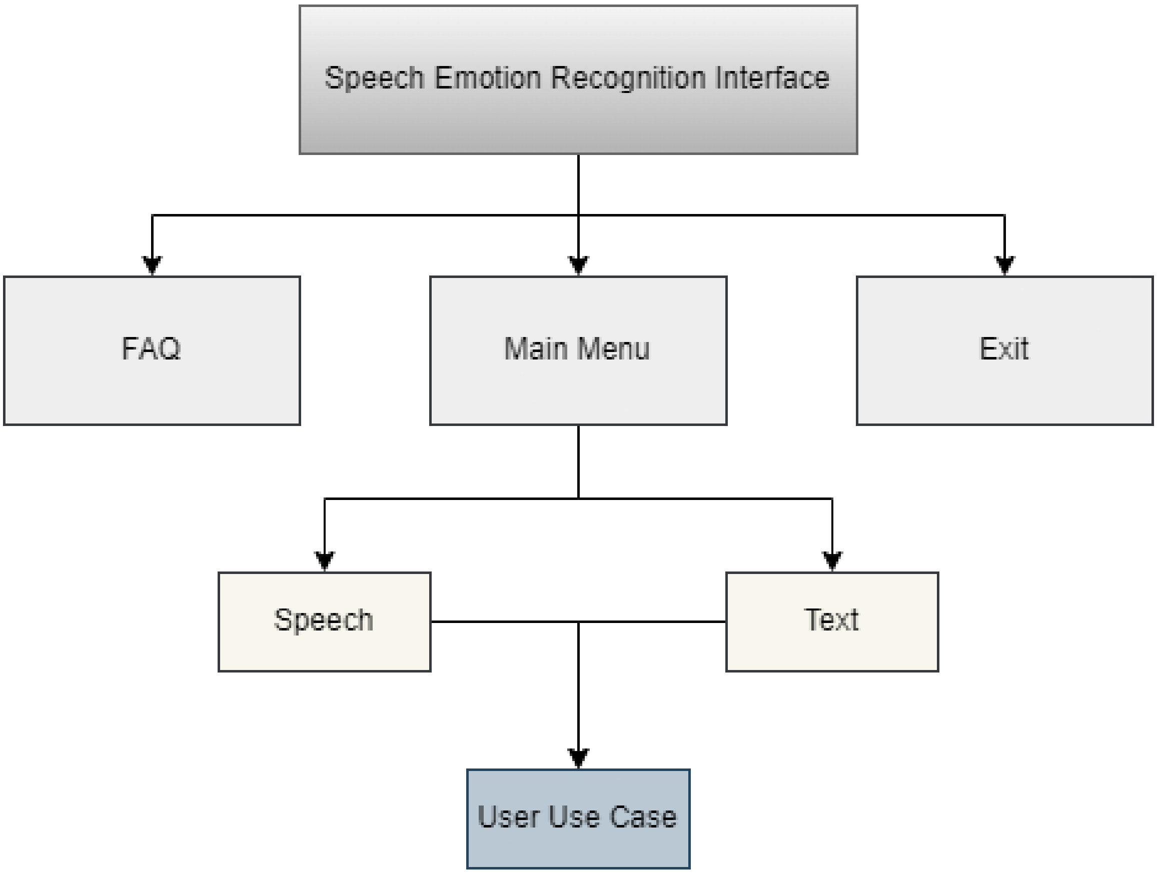


Figure 5 Project Components of Speech Emotion Recognition

4.4.1 Brief Description of *Figure 5*

Speech Emotion Recognition Project Components are shown in Figure 5. All designed systems of the speech Emotion Recognition project are shown in the block diagram in the figure. The system has 4 subcomponents, 1 of which has its subcomponent.

4.4.1.1 Frequently Asked Questions Design

Frequently Asked Questions (FAQ) are designed to contain information about the software and the Main Menu. Here are tips on how to use the software. In addition, information such as on which metrics it was resolved and approximately what accuracy value was found are also included.

4.4.1.2 Main Menu Design

The Main Menu is designed to allow users to easily use the developed Speech Emotion Recognition software. The GUI consists of three main heads. These headings are Main Menu, Frequently Asked Questions and Exit. There are 2 subtitles in the Main Menu. These subheadings are “Speech” and “Text”. Users can upload an audio file and learn emotions from this file. The user can also write an article and learn the emotion from that article.

4.4.1.3 Exit Design

The exit header is used to exit of the software.

4.5 Detection

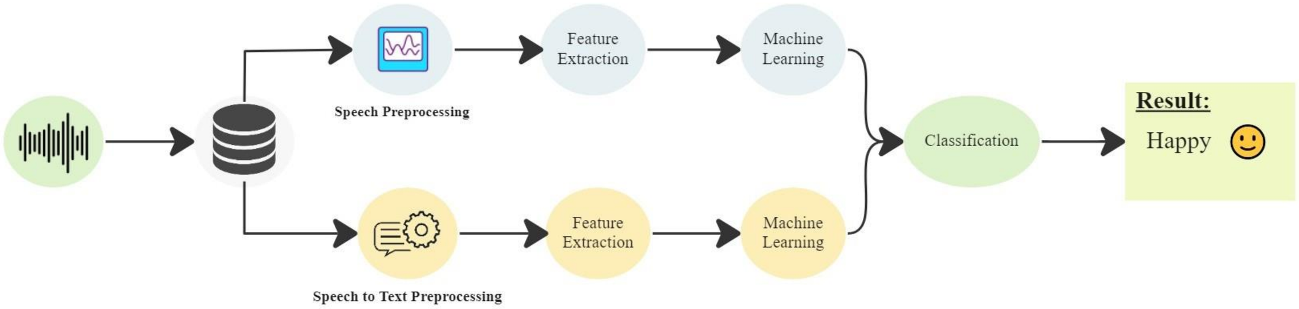


Figure 6 Project System Overview

In this project, the technique of receiving and processing voice data as input and emotion recognition with voice + text was used to create Speech Emotion Recognition.

The data is taken as input by the algorithm. At this stage, sound and text are processed with separate features and extractions. In the next step, the data obtained from both speech and text are subjected to feature extraction. The features obtained as a result of this process are given to the Machine Learning algorithm. In the last stage, the data coming to the classification algorithm shows the emotion according to the "anger", "excited", "sadness", "frustration", "happiness", "neutral" emotions.

4.6 Contributions

When we look at this paper, we see that there are many studies in the field of Speech Emotion Recognition and each study ends their work with a different technique. We aim to design a system that determines 4 different emotional states: Angry, Happy, Sad, Neutral. While detecting these emotions, we are planning a system that extracts emotion analysis from the voices of the speaker. In previous studies, similar systems were designed and the accuracy rates of these systems were calculated by techniques as k-NN, CNN, RF, XGB, and SVM, and an average of 60-65 percent successes was determined. In these studies, datasets such as Urdu Emotional Speech Dataset, RML Dataset, IEMOCAP or Text only were used as datasets. Although we plan to use 4 different situations as Angry, Happy, Sad, Neutral, different emotional analyzes such as Fear, Disgust, Surprise, Joy were also included in these studies.

Although the English language dataset is used in most of these studies, many different languages, including Turkish, are used in a substantial part.

The main operation of the system we want to do is to first make an emotion analysis from the voice and then develop a system that makes speech to text and makes emotion analysis from that text. Then, to combine these two different analyzes and produce a result.

5. Future Works

In this study, we conducted researches that enable people to understand their emotional states from their speech or text. We designed a site that allows people to upload the audio files they have recorded to the system or to upload the text files they have written to the system and measure their mood. We designed a system that allows a person to measure 5 basic emotional states and included happy, neutral, sad, angry and fearful emotions. This study can be useful in many areas. For example, the emotions arising from phone calls can be effectively supported in calls made with call centers and customer satisfaction can be increased. Or it can be helpful in measuring team collaboration and evaluation of work in teamwork. In addition to these, there are some difficult parts of this study.

Since it is a generalizable study, the limited number of data sets prevents large and large studies from being carried out, and it is difficult to adapt the records from different populations in terms of language and culture. Accordingly, future studies will help to find descriptors that best represent emotional state. For example, besides the audio file from the file, it can be more comprehensive through changes in heart rate or facial expressions.

6. Conclusions

At the end of the CENG407 - Innovative System Design and Development I course, the development stages of a software project were experienced. There was an opportunity to work on a real-world project on Literature Review, where previous studies on the subject to be studied were meticulously examined, Software Requirements Specifications, where the main business processes to be supported for the software, features, basic performance parameters were examined, and the writing of Software Design Documents, where more technical details and risks were concerned about the software.

You can reach this Literature Review, Software Requirements Specification and Software Design documents by examining the 2nd, 4th and 5th sections.

7. Acknowledgment

We are grateful for the assistance and guidance we have provided so far from Assist. prof. Dr. Ayşe Nurdan SARAN. The help we received from him was a valuable gain not only for developing the project and increasing its quality but also for improving our theoretical and practical knowledge about ourselves, crisis management and the steps to bring a project to its final stage.

8. References:

[1] Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A. Empathy and emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1), 98-105,2013 [1]

[2] Asghar, A., Sohaib, S., Iftikhar, S., Shafi, M., & Fatima, K. (2022). An Urdu speech corpus for emotion recognition. *PeerJ Computer Science*, *8*, e954.

[3] Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, *8*, 79861-79875.

[4] Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. *Procedia Computer Science*, *176*, 251-260.

[5] Harár, P., Burget, R., & Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In *2017 4th International conference on signal processing and integrated networks (SPIN)* (pp. 137-140). IEEE.

[6] Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*.

[7] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In *Social media and machine learning*. IntechOpen.

[8]A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connection- ´ ist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376. ACM, 2006.

[9] Mozilla. Project deepspeech. https://github.com/mozilla/DeepSpeech, 2017.

[10] R. Banse, K. Scherer, Acoustic profiles in vocal emotion expression, J. Pers. Soc. Psychol. 70 (3) (1996) 614–636.

[11] L. Bosch, Emotions, speech and the asr framework, Speech Commun. 40 (2003) 213–225.

[12] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (4) (2000) 429–442.

[13] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Language Process. 17 (4) (2009) 582–596.

[14] R. Cowie, R.R. Cornelius, Describing the emotional states that are expressed in speech, Speech Commun. 40 (1–2) (2003) 5–32.

[15] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.

[16] J.R. Davitz, The Communication of Emotional Meaning, McGraw-Hill, New York, 1964.

[17] C. Gobl, A.N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.

[18] T. Johnstone, K.R. Scherer, Vocal Communication of Emotion, second ed., Guilford, New York, 2000, pp. 226–235.

[19] L. Kaiser, Communication of affects by single vowels, Synthese 14 (4) (1962) 300–319.

[20] T. Nwe, S. Foo, L. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (2003) 603–623.

[21] K.R. Scherer, Vocal affect expression. A review and a model for future research, Psychological Bull. 99 (2) (1986) 143–165 cited by (since 1996) 311.

[22] H. Teager, S. Teager, Evidence for nonlinear production mechanisms in the vocal tract, in: Speech Production and Speech Modelling, Nato Advanced Institute, vol. 55, 1990, pp. 241–261.

[23] C. Williams, K. Stevens, Vocal correlates of emotional states, Speech Evaluation in Psychiatry, Grune and Stratton, 1981, pp. 189–220.

[24] G. Zhou, J. Hansen, J. Kaiser, Nonlinear feature based classification of speech under stress, IEEE Trans. Speech Audio Process. 9 (3) (2001) 201–216.