SOFTWARE REQUIREMENT SPECIFICATION

For

Identification of speech emotion to detect a user's mood

10 Dec 2022

Submitted By

Specialization	SAP ID	Name
AIML	500087027	Nidhish Pandey
AIML	500087764	Om Agarwal
AIML	500087330	Rupesh Kumar

Department of Informatics

School of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

Table of Contents

Revision History

Date	Change	Reason for Changes	Mentor Signature
16/09/2022	Start collecting several research paper regarding our project	So that we have deep knowledge of our project.	
2/10/2022	Increase the size of dataset	To increase the accuracy of our model.	
27/10/2022	Planned to introduce GUI in our project	So that our model will become user friendly.	
3/11/2022	Increased the size of training data.	To increase the accuracy of our model on different speeches.	
16/11/2022	Planned to introduce GUI in our project	So that the project can be accessed on web.	
01/12/2022	Planned to create our own dataset for SER.	To improve accuracy of model on Indian Regional Languages.	

1. Introduction

Automatic speech recognition is basically the process of converting spoken words into text form, basically transcribing what someone is speaking. It is a challenging problem to solve, but you can see various examples of this technology at work nowadays. Since emotions help us to understand each other better, a natural outcome is to extend the understanding to computers. Speech recognition is already in our everyday life, thanks to smart mobile devices that are able to accept and reply to synthesized speech. Speech emotion recognition could be used to enable them to detect our emotions as well.

If you have an android phone and just say 'OK Google", you will see a window open up at the bottom. If you now speak more words, you will see that the app will try to identify what you are speaking. It won't get perfect the first time, and it probably will show some intermediate words as you continue to speak. But, in the end, the technology will recognize your speech.

Another example you can see in YouTube videos is if you turn on CC or Closed Captions. In a lot of videos, the text is actually being recognized on the go by using the speech recognition models built by Google. There are lots of other examples available online like this.

1.1. Purpose of Project

Emotion detection has become one of the biggest marketing strategies in which the mood of the customer plays an important role. So to detect the current emotion of the person and suggest to him the appropriate product or help him accordingly will increase the demand of the production company.
Humans have the natural ability to use all their available senses for maximum awareness of received messages. Emotional detection is natural for humans but it is a very difficult task for machines.

Detecting emotions is one of the most important strategies in today's world.
For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications.

Some examples could be -

- Slowing down a smart car when one is angry or fearful.
- Including call centres to play music when one is angry on the call.

1.2. Target Beneficiary

Emotion detection has become one of the biggest marketing strategies in which the mood of the customer plays an important role. So to detect the current emotion of the person and suggest to him the appropriate product or help him accordingly will increase the demand of the production company.

Detecting emotions is one of the most important strategies in today's world.

For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications.

Some examples could be -

→ Slowing down a smart car when one is angry or fearful.

→ Including call centres to play music when one is angry on the call.

1.3. Project Scope

Emotion recognition is the process of **identifying human emotion**. People vary widely in their accuracy at recognizing the emotions of others. Use of technology to help people with emotion recognition is a relatively nascent research area.

The primary objective of SER is to improve man-machine interface. It can also be used to monitor the psycho physiological state of a person in lie detectors. In recent times, speech emotion recognition also finds its applications in medicine and forensics.
The literature in speech emotion detection is not very rich and researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotion to class together.

This model can be used by various apps, online shopping websites and so onto know about the user's emotions. Further improvements can be made to the model so that it can perform well in real time. For improving the accuracy of the model, we can increase the size of the dataset. The classifier can be embedded in a software or an app so that it can work in real time.

1.4. References

•https://ijesc.org/upload/bc86f90a8f1d88219646b9072e155be4.Speech%20Emotion%20Recognition%20using%20MLP%20Classifier.pdf

•https://eudl.eu/pdf/10.4108/eai.7-12-2021.2314726

•https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

•https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

•https://www.kaggle.com/datasets/ejlok1/cremad

•https://www.kaggle.com/datasets/dejolilandry/asvpesdspeech-nonspeech-emotional-utterances

2. Project Description

Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents.

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.

Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.

SER(Speech Emotion Recognition) is used in call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on.. for helping companies improving their services

It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen

2.1. Reference Algorithm

Since the project is a classification problem, Multilayer Perceptron seems the obvious choice. We chose this model to predict the right emotions.

This classifier connects to a neural network. Unlike other classification algorithms such as Support Vector or Naïve Bayes classifier, MLP classifier rely on an underlying neural network to perform the task of classification.

2.2. Data Structure

In this project the data structures used are one dimensional and multidimensional numpy arrays because NumPy arrays are faster and more compact than Python lists. An array consumes less memory and is convenient to use. NumPy uses much less memory to store data and it provides a mechanism of specifying the data types. This allows the code to be optimized even further.

2.3. SWOT

Strength:

- •Current need of the market
- •Useful in managing other AI application

Weakness:

- •Fault intolerant.
- •Accuracy may depend on the accent of the user.

Opportunity:

- •A new boost for the automotive industry.
- •Boon for product based companies

Threat:

- •Some research is ongoing to achieve this objective more efficiently.

2.4. Project Features

- The software program described on this SRS may be used to come across people's feelings.
- This project can be used in numerous areas that want to measure client pleasure in a marketing platform, assisting advertisers to promote products extra efficiently.The soft

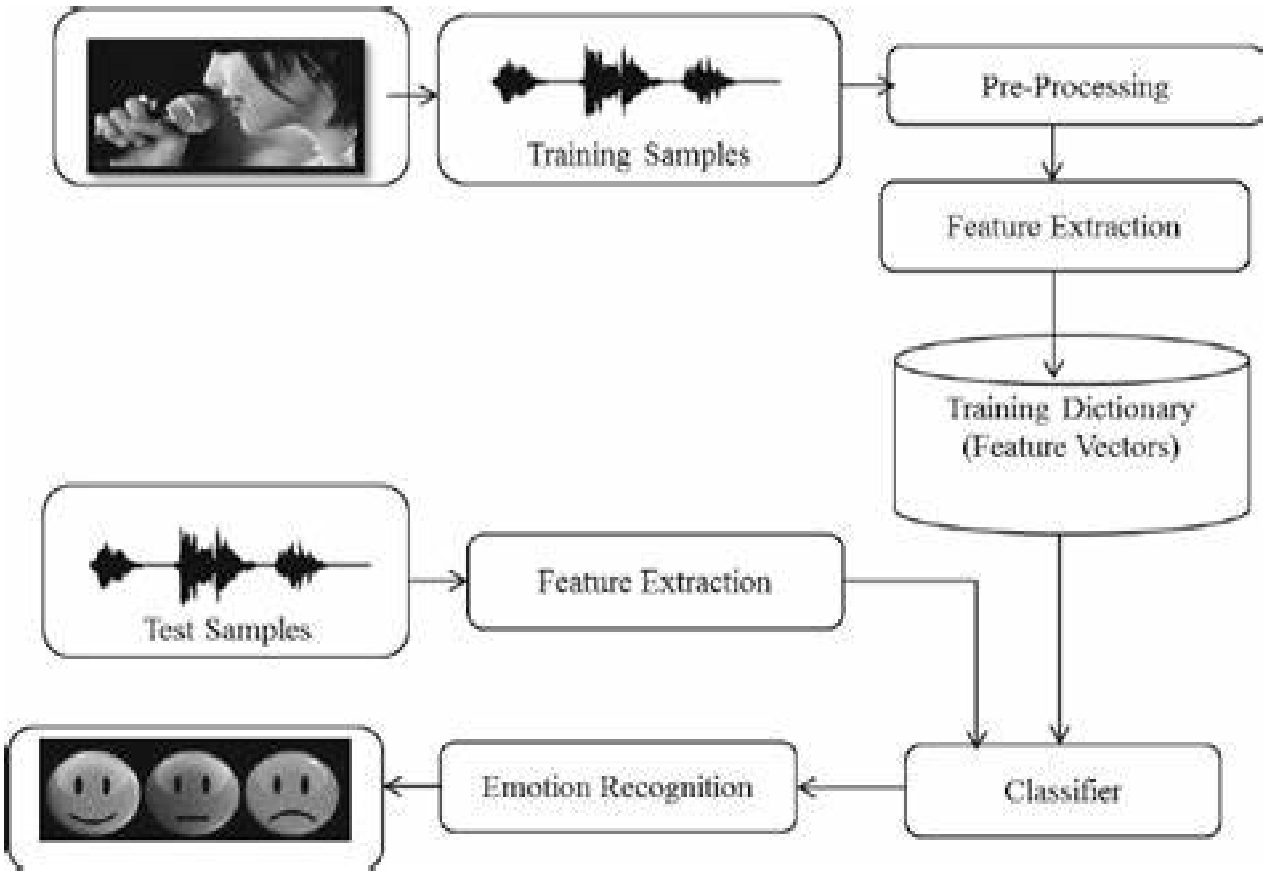
2.5. User Classes and Characteristics

Here the users are basically divided into 6 classes based on six different emotions which are:

- Happy
- Sad
- Neutral
- Fear
- Disgust
- Angry

The users are classified in these classes on the basis of the characteristics of their voice samples like chrome, mfccs, mel etc.

2.6. Design Diagrams



2.7. Assumption and Dependencies

The main assumption of this project is that the user will be provided with voice input without any fake accent or emotion. If the emotion was faked by the user the model can't predict the correct emotion precisely.

3. System Requirements

3.1. User Interface

The user can interact with our project via our GUI

.

3.2. Software Interface

Following are the softwares for our SER GUI.

Software used	Description
Operating System	We have chosen the Windows operating system for its best support and user-friendliness.
Database	To save the audio files we used .wav format and the predictions made by model we used .csv formats of files.
Visual Studio Code	To implement the project we have chosen python language and its more efficient for machine learning based projects.

3.3. Database Interface

For training of our model we use a custom dataset that we created by using the three already existing datasets. Our dataset contains 11,178 audio files which are divided into 6 emotions: Happy, Angry, Disgust, Fear, Neutral and Sad.

1. RAVDESS:

This dataset includes around 1500 audio files input from 24 different actors. 12 male and 12 female where these actors record short audios in 8 different emotion.

2. TESS:

The TESS Dataset is a collection of audio clips of 2 women expressing 7 different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

3. CREMA:

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences.

4. SAVEE:

The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories.

5. ASVP-ESD

The Audio, Speech, and Vision Processing Lab Emotional Sound database (ASVP-ESD).

This dataset contains audio files regrouped in 130 folders; The data are organized as follows: Meanwhile some are mixed, odd folder numbers are mainly for females, and even for males (total size: 2 GB).

As it's a realistic dataset some folders contain dialog or several people interacting in the audio; Speech and non-speech Emotional sounds include boredom(sigh,yawn), neutral, happiness (laugh, gaggle), sadness(cry), anger, fear (scream, panic), surprise(amazed,gasp), disgust(contempt), excite(Triumph,elation), pleasure(desire), pain(groan), disappointment; A total of 12 different emotions plus breath. 2 levels of intensity were used for the database (normal and high).

3.4. Protocols

Not Applicable

4. Non-Functional Requirements

4.1. Performance Requirement

The project must meet the end user requirements. Accuracy and fast must be imposed in the Project. The project is development as easy as possible for the sake of the end user. The project has to be developed with a view of satisfying the future requirements and future enhancement. The project has been finally implemented satisfying the needs specified by the company. As per the performance is concerned this system is performing This processing as well as time taken to generate well reports even when large amounts of data were used.

4.2. Security Requirement

Web applications are available via network access, it is difficult. If not possible, to limit the population of the end-user who may access the applications? In order to product sensitive connect and provide secure mode be implemented throughout the infrastructure that supports

the web application and within the application itself. Web Applications have become heavily integrated with critical corporate and database. E-commerce applications extract and then store sensitive customer information.

4.3. Software Quality Attributes

A software component that is developed for reuse would be correct and contain no defects. In reality, formal verification is not carried out routinely, and defects can add to occur. However, with each reuse, defects are eliminated, and a component qualifies improve as a result. Over time the components virtually defect free. This project will be developed on a daily basis for further logical and performance-based enhancements. Software reliability is defined in statically term as" the probability of faultier-free operation of a computer program in a specified environment for specified time". The software quality and reliability, failure is non-conformance to software requirements. Failure can be only anything or catastrophic. One failure can be corrected within 12 seconds while another requirements week even mouths to correct. Complicating the issue even further, the correction of the one failure may in fact result in the introduction of the errors that ultimately result in other failure.

5. Other Requirements

5.1. Appendix A: Glossary

Emotion recognition can be used to understand how candidates feel during interviews and to measure how they react to certain questions. This information can be used to optimize interview structure for future candidates and streamline the application process.

It can also be used to understand feelings of a student while learning something, this will enable us to improve the quality of teaching in more efficient ways.

5.2. Appendix C: Issue List

- Existing Datasets-

  There are several open and paid datasets available for Speech Emotion Recognition. But there are shortcomings related to them.

These datasets are extraordinarily confined in covering languages, audio system, genders, a while, dialects, etc. The trouble of training, validating, and checking out on a restrained dataset is the abundance of overconfidence (overfitting) and tough failure within the area.

There isn't any standard set of labels for human emotions! These datasets all use a barely one of a kind set of emotions.

Being angry in a web recreation differs from being indignant with a customer service agent. A worldwide definition of what an emotion seems like seems really futile. Check the picture of Feelings Wheel beneath for a visual rationalization.

- Curating a Dataset-

The direct approach is to collect and label the training examples. It is best to keep collecting data beyond the first version of the dataset because certainly we need more data after your first training run. We also have to invest in building a learning loop so you can keep gathering and labelling data from models in production. This method is expensive and time-consuming.

If don't have the means to collect a dataset, we need to get creative and can use crowdsourcing, which is much cheaper than building an in-house labelling team. But still, we pay per label. The last measure is to use an already-trained model to create labels for us.

- Training the Model-

Assuming you have enough data, you can train a classifier end-to-end that maps from a speech signal to one of the labels. In practice, we won't be able to, and likely overfit to train and fail to generalize.