IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

Speech Emotion Recognition in Machine Learning and IoT

Prathamesh Shinde[1], Sufiyan Gawandi[2], Atharva Baxi[3], Aman Pathan[4]

Student, Department of Computer Engineering, Trinity College of Engineering and Research, Pune, India[1,2,3,4]

Abstract: In the past decade a lot of research has gone into Automatic Speech Emotion Recognition (SER). 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 time, speech emotion recognition also finds its applications in medicine and forensics. In this paper 7 emotions are recognized using pitch and prosody features. Majority of the speech features used in this work are in time domain. Support Vector Machine (SVM) classifier has been used for classifying the emotions. Berlin emotional database is chosen for the task. A good recognition rate of 81% was obtained. The paper that was considered as the reference for our work recognized 4 emotions and obtained a recognition rate of 94.2%. The reference paper also used hybrid classifier thus increasing complexity but can only recognize 4 emotions.

Keywords : Artificial Intelligence, Machine Learning, Voice Recognition, Speech Recognition, Speech Emotion Recognition.

1.INTRODUCTION

Human emotions are very difficult to comprehend from a quantitative perspective. Facial expressions are one of the best ways of guessing the emotional state of a person. Speech is another modality that can be used. Speech is a complex signal which contains information about the message, speaker, language and emotions. There are various kinds of emotions which can be articulated using speech. Emotional speech recognition is a system which basically identifies the emotional state of human being from his or her voice; speech is very misleading even for humans to judge the emotion of the speaker. A major motivation comes from the desire to improve the naturalness and efficiency of human-machine interaction. The reference paper that was chosen has been able to successfully recognize only 4 emotions. The work presented here has classified 7 emotions with a overall good recognition rate. In general, the systems for speech analysis uses various techniques for the extraction of characteristics from the raw signal. Concerning emotions, the relevant information is in the Pitch, Prosody and in the Voice quality. The next step in this strategy is to discover the features which discriminate the speech data (to the training labels) and to discard the non-discriminative features. This is achieved by calculating the cross validation between parameters after which grid of parameters is created; the one with highest cross validation is selected. The Emotional profiles (EP) are constructed using SVM with Radial Basis Function (RBF). Emotion-specific SVMs are trained for each class as self-versus others classifiers. Each EP contains n-components, one for the output of each emotion-specific SVM. The profiles are created by weighting each of the n-outputs by the distance between the individual point and the hyperplane boundary. The final emotion is selected by classifying the generated profile. This is done by one vs one comparing of each emotion to the existing profile of the emotion. Fig.1 comprehensively explains the methodology followed in this paper. Emotion recognition is done using two modules. The first module is the feature extraction module and the second is the classifier module. In the feature extraction module, we have used a feature set comprising pitch, prosody and voice quality features. Several classifiers exist for the task of emotion recognition. The different classifiers are SVM, MLP (Multilayer Perceptron), HMM (Hidden Markov Model), GMM (Gaussian Mixture Model), ANN (Artificial Neural Networks) etc. The SVM classifier yields good results even from small test samples and hence it is widely used for speech emotional recognition. The SVM classifier is therefore used for the proposed work. Because of the Structural Risk Minimization, SVM classifiers usually have better performance than others.

2.LITERATURE SURVEY

Emotion recognition and its production are effortless abilities of human beings, as opposed to machines. Getting machines endowed with the ability to recognize emotions is a tremendous challenge, which, if achieved will determine the naturalness of human-machine interaction. The first and foremost requirement for achieving automatic emotion recognition is the availability of databases. Database collection of emotional speech requires spontaneity and genuineness, because emotional speech is usually exhibited when a person is not in Neutral state. Often when a database is collected, it is usually in a controlled environment, so the emotional speech tends to sound contrived. Due to this factor, speech recordings from conversations of call centers and interviews of TV shows are often used, as in. Basic emotion states such

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

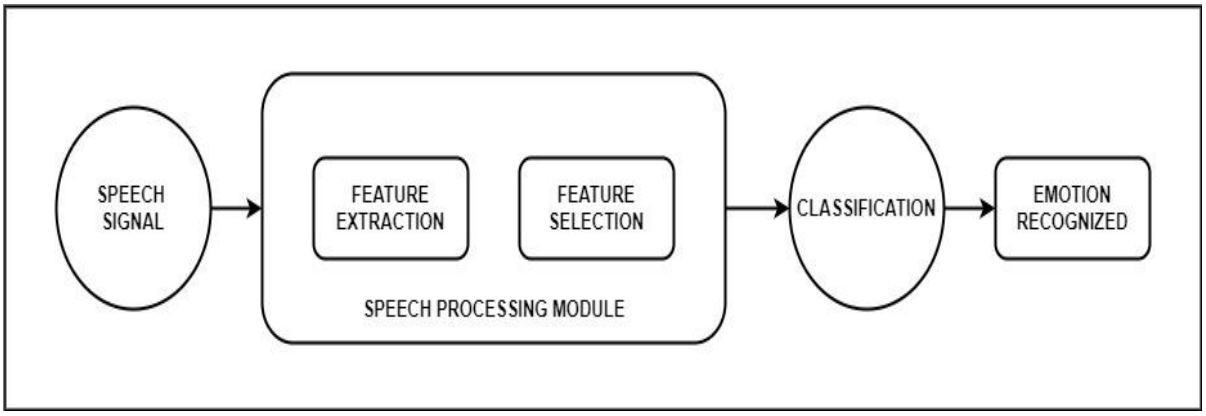
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

as Anger, Happy, Sad, Fear, and surprise are the most common case of studies. Automatic determination of these emotions from a speech signal has been attempted and successful in several researches in today's fast advancing speech technology. Classification of human speech into 4 emotional states (Neutral, Anger, Happy, Sad) have been seen in, where sub-segmental features like loudness, energy of excitation, detection of voiced and. Block diagram to illustrate the basic outline of emotion recognition through speech signal. unvoiced region, instantaneous F0, and strength of excitation were used for analysis. Spectral band energy ratio and strength of excitation were examined for distinguishing Anger and Happiness of two databases, with accuracy of 75% for IIITH Telugu Database and 68% for German Emotional Database. For recognition of 7 emotions, an overall recognition rate of 91.6% was obtained in, using modulation spectral features (MSF), short-term spectral features (MFCC and PLP) and prosodic features. Instantaneous fundamental frequency (F0) is mostly used for robustness emotion recognition. Formants are quantitative characteristics of the vocal tract and are characterized by unique center frequencies and bandwidth. Estimation of Formants can be done using a technique called Linear Prediction Analysis (LPA). For emotion recognition, classifiers such as Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN), k-Nearest Neighbor (kNN), etc. are commonly used. Apart from the basic emotions, several prominent studies related to non-verbal speech which depict emotions have also been seen. Few prominent mentions include detection and analysis of shouted speech (indicating high-arousal or angry speech), and analysis of laughter. Another study has made extensive experimentation and analysis on the non-verbal speech sounds which were examined into three distinct categories - paralinguistic sounds, emotional speech and expressive voices.

3. EXISTING SYSTEM



Phase 1 – Training phase
System learns reference patterns which represent different speech sounds (e.g. phrases, words, phones) that constitute the vocabulary of the application.

Phase 2 – Recognition phase
Unknown input pattern is identified using set of references.
Speech Recognition System works in following stages -

- Speech Analysis

Speech data is analyzed which includes speaker specific information due to vocal tract, excitation source and behavior feature which is important for speaker recognition.

- Feature Extraction

Different individual characteristics of speech embedded in utterances are extracted.

- Modelling

Hidden Markov Model (HMM) is used to create models for each letter.

- Testing

Feature testing of the dataset is done.

4.PROPOSED SYSTEM

Emotion recognition systems based on digitized speech is comprised of three fundamental components: signal preprocessing, feature extraction, and classification. Acoustic preprocessing such as denoising, as well as segmentation, is carried out to determine meaningful units of the signal. Feature extraction is utilized to identify the relevant features available in the signal. Lastly, the mapping of extracted feature vectors to relevant emotions is carried out by classifiers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

The system architecture given below depicts a simplified system utilized for speech-based emotion recognition. In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, and feature selection. The required features are extracted from the preprocessed speech signal and the selection is made from the extracted features. Such feature extraction and selection are usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers are utilized for classification of these features. Lastly, based on feature classification different emotions are recognized. The system proposed here will have one more stage where lights will be automated on the basis of emotion recognized in the earlier stage.

Phase 1 – Training phase
System learns reference patterns which represent different speech sounds (e.g. phrases, words, phones)
Phase 2 – Recognition phase
Unknown input pattern is identified using set of references.
Speech Recognition System works in following stages -

- Speech Analysis

Speech data is analyzed which includes speaker specific information due to vocal tract, excitation source and behavior feature which is important for speaker recognition.

- Feature Extraction

Different individual characteristics of speech embedded in utterances are extracted.

- Modelling

Then we test the model and recognize the emotion

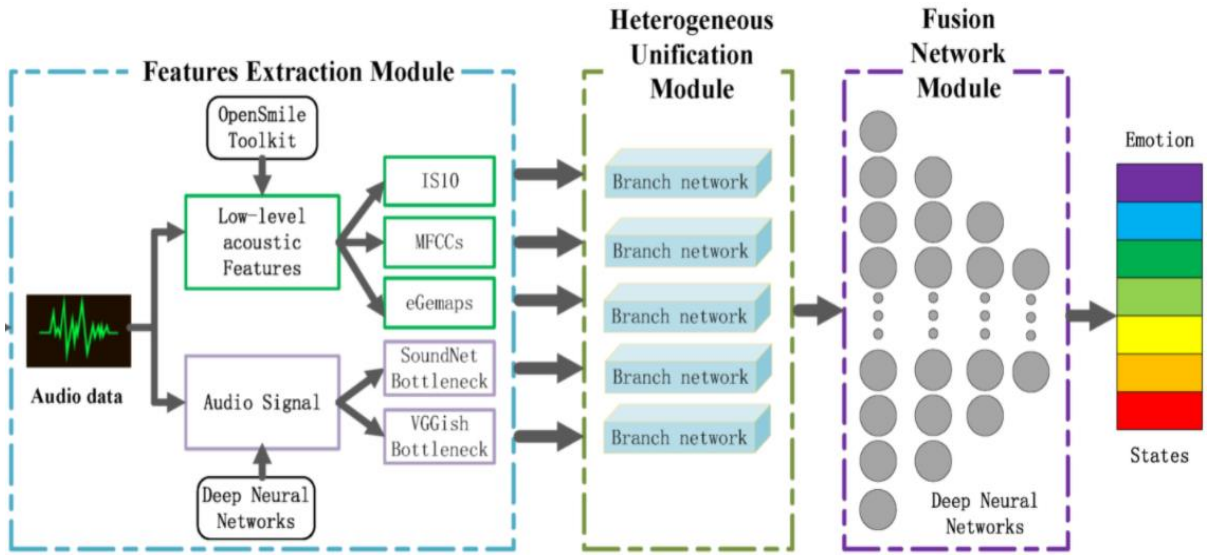
- Testing

Feature testing of the dataset is done.
Lights are changed according to Emotions recognized

- According to the emotions recognized we change the lights of the room using the recognized emotion as field input
- Then the decision control is done and signal is sent to the sensors and lights are automated accordingly

5.ML SYSTEM ARCHITECTURE

For Emotion Detection



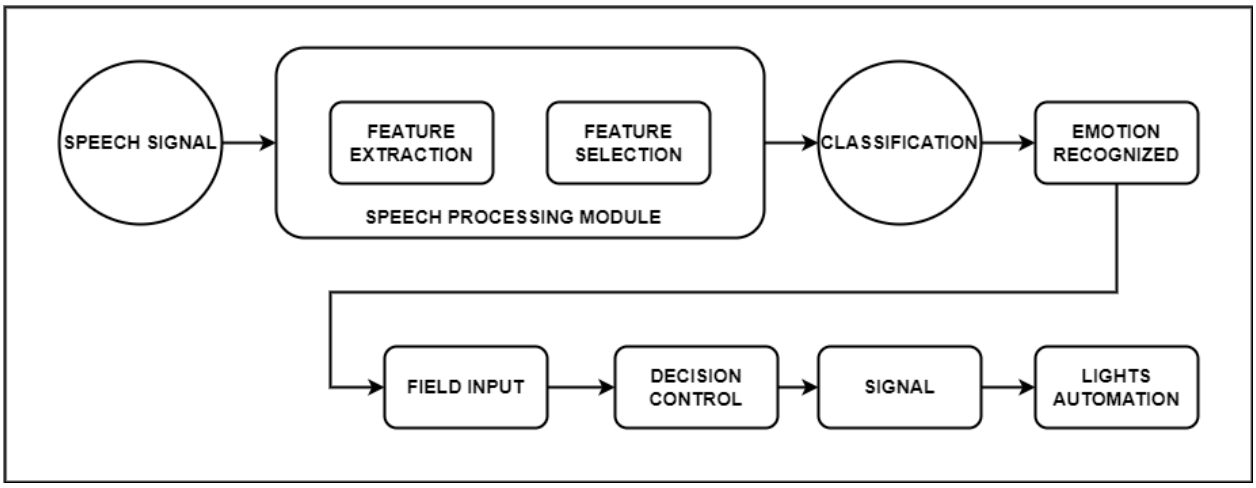
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

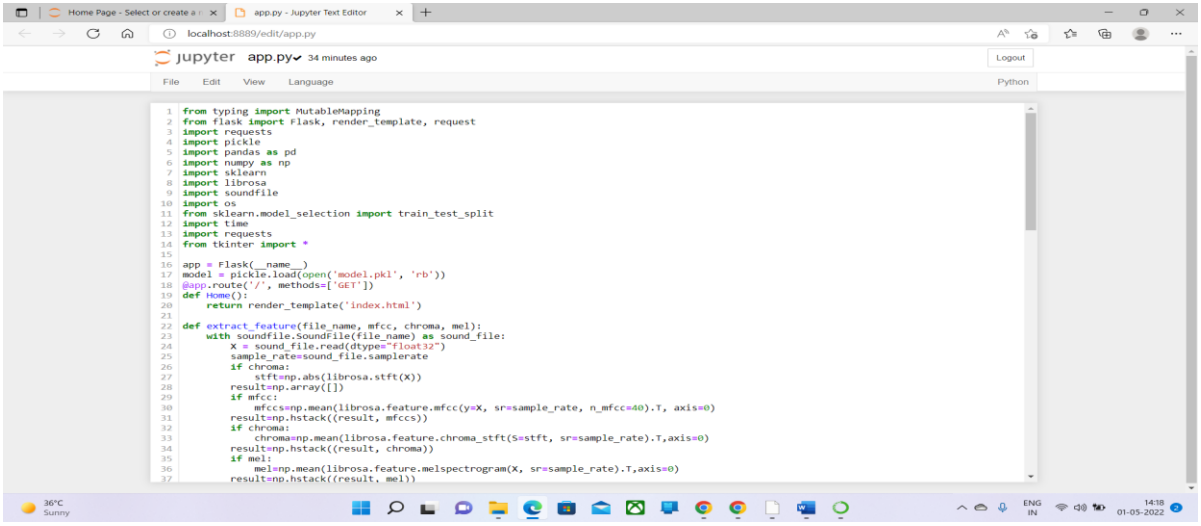
Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

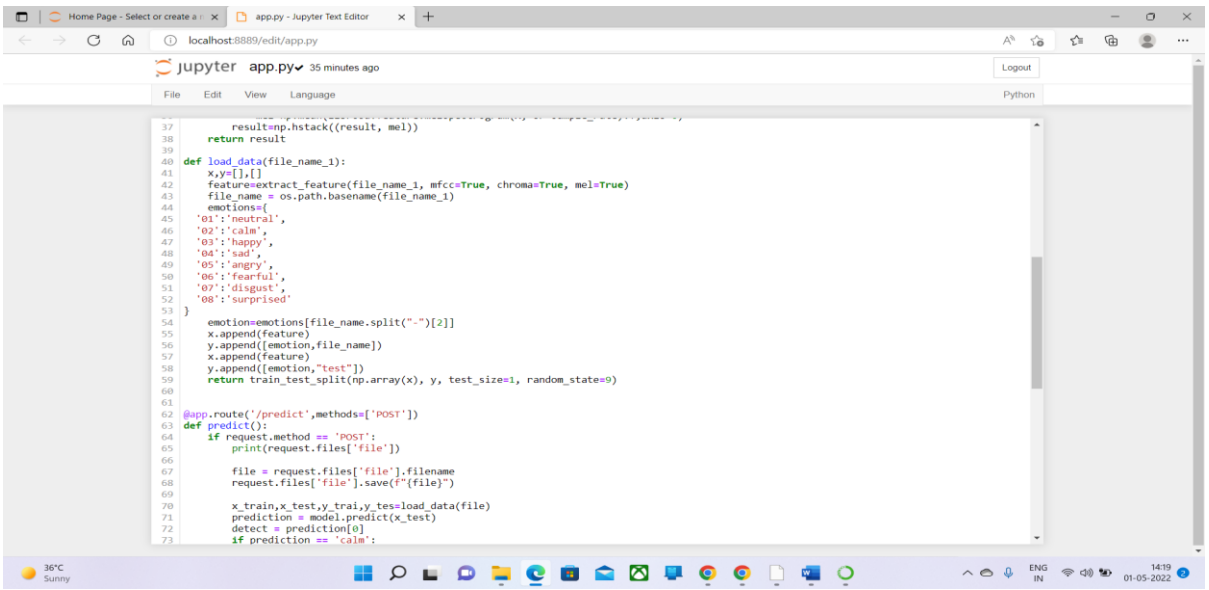
Our IoT addition funtionality



6.OUR ALGORITHM



```
from typing import MutableMapping
from flask import Flask, render_template, request
import requests
import pickle
import pandas as pd
import numpy as np
import sklearn
import librosa
import soundfile
import os
from sklearn.model_selection import train_test_split
import time
import requests
from tkinter import *

app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))
@app.route('/', methods=['GET'])
def Home():
    return render_template('index.html')

def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
            stft=np.abs(librosa.stft(X))
        result=np.array([])
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
        result=np.hstack((result, mfccs))
        if chroma:
            chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
        result=np.hstack((result, chroma))
        if mel:
            mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
        result=np.hstack((result, mel))
```



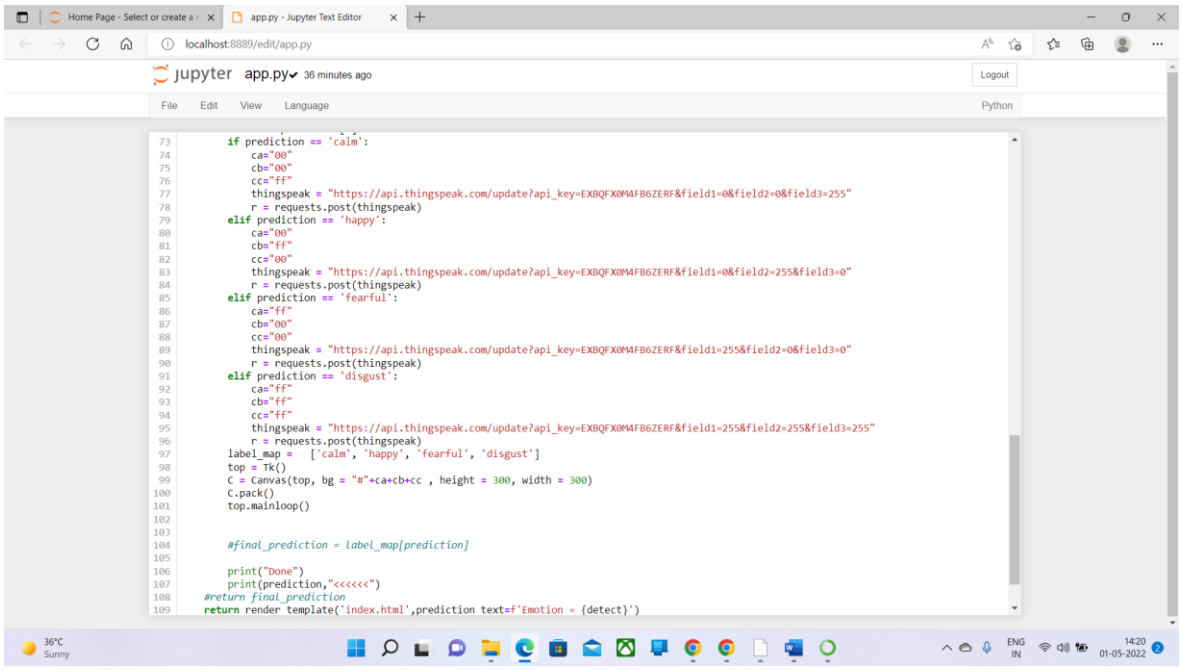
```
        result=np.hstack((result, mel))
    return result

def load_data(file_name_1):
    x,y=[],[]
    feature=extract_feature(file_name_1, mfcc=True, chroma=True, mel=True)
    file_name = os.path.basename(file_name_1)
    emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}
    emotion=emotions[file_name.split("-")[2]]
    x.append(feature)
    y.append([emotion,file_name])
    x.append(feature)
    y.append([emotion,"test"])
    return train_test_split(np.array(x), y, test_size=1, random_state=9)

@app.route('/predict',methods=['POST'])
def predict():
    if request.method == 'POST':
        print(request.files['file'])

        file = request.files['file'].filename
        request.files['file'].save(f"{file}")

        x_train,x_test,y_trai,y_tes=load_data(file)
        prediction = model.predict(x_test)
        detect = prediction[0]
        if prediction == 'calm':
```

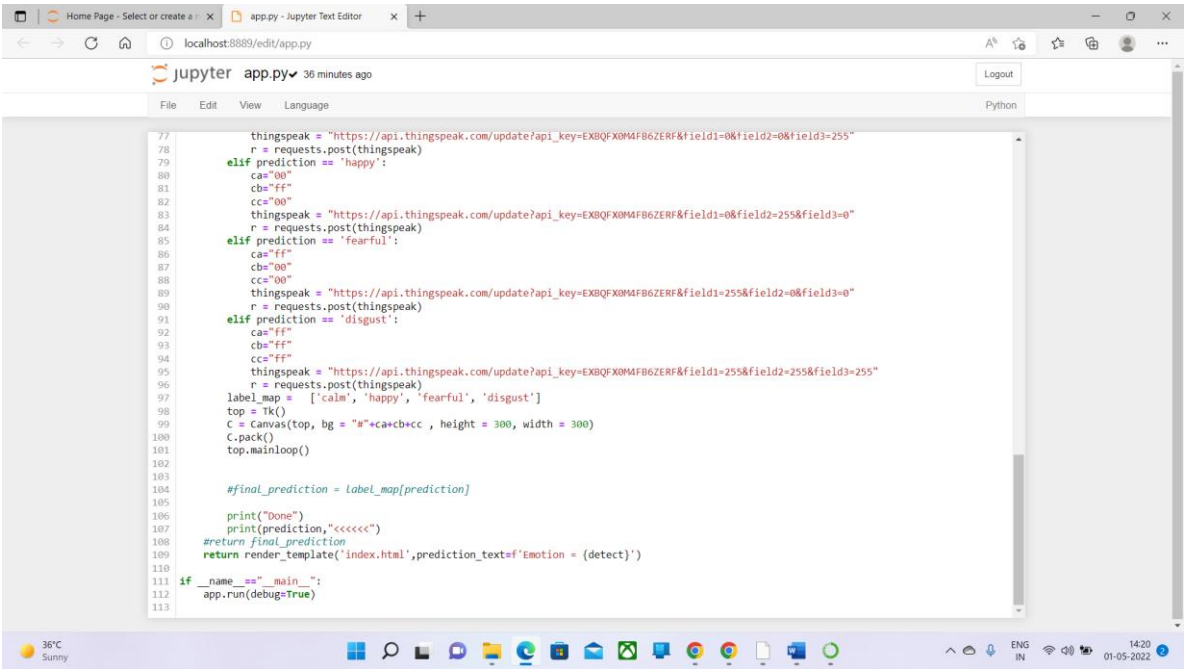
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

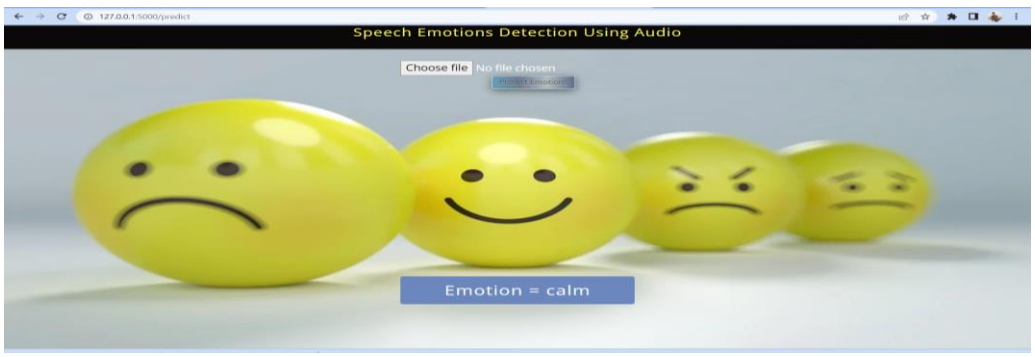
Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456





OUTPUT

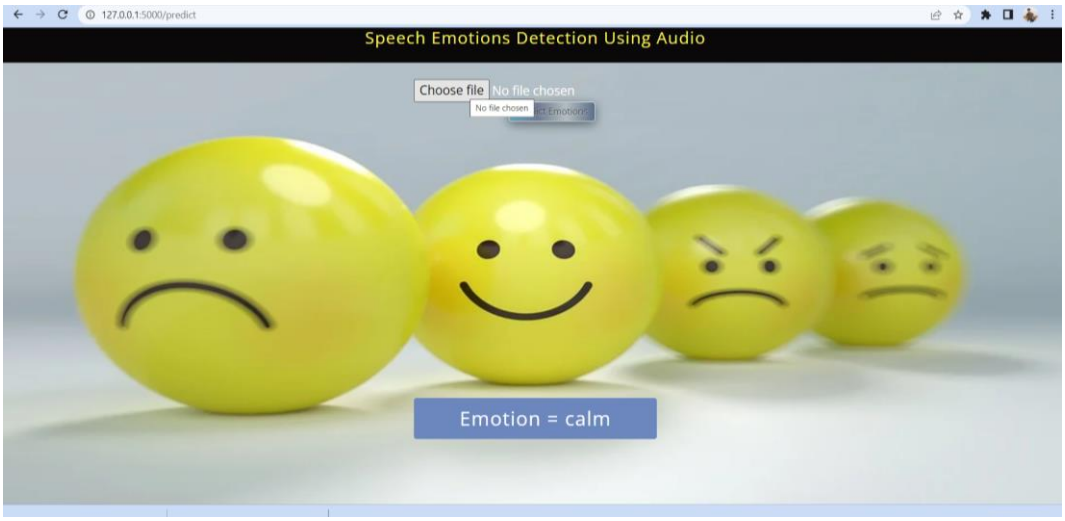


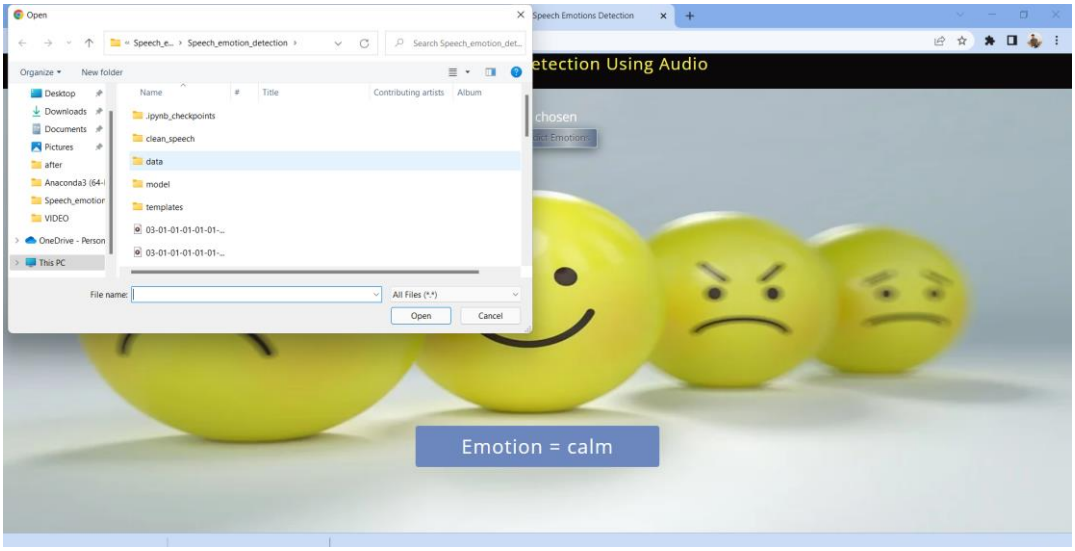
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

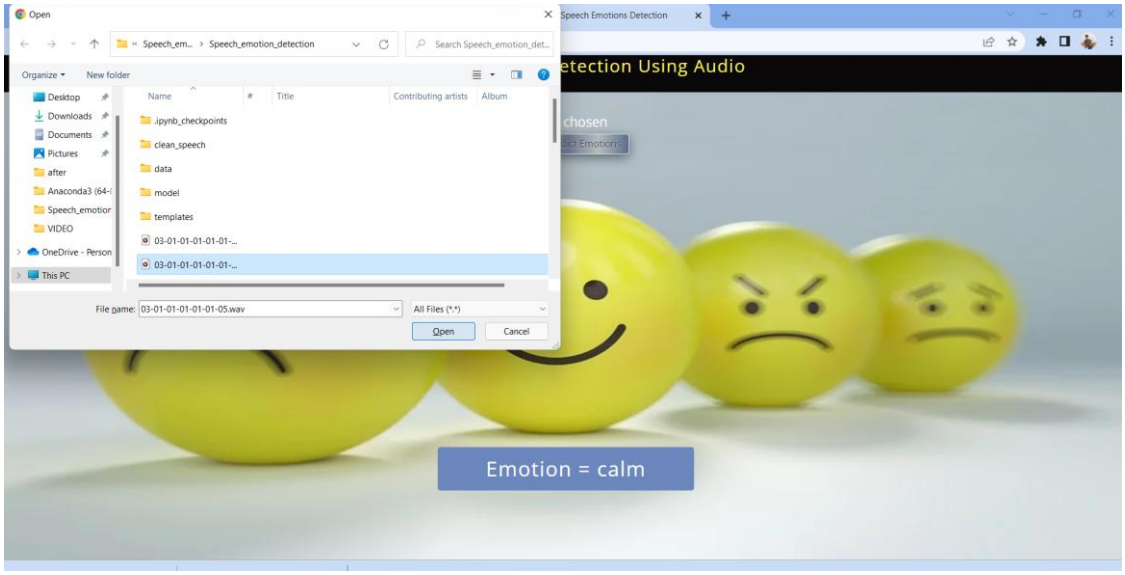
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456





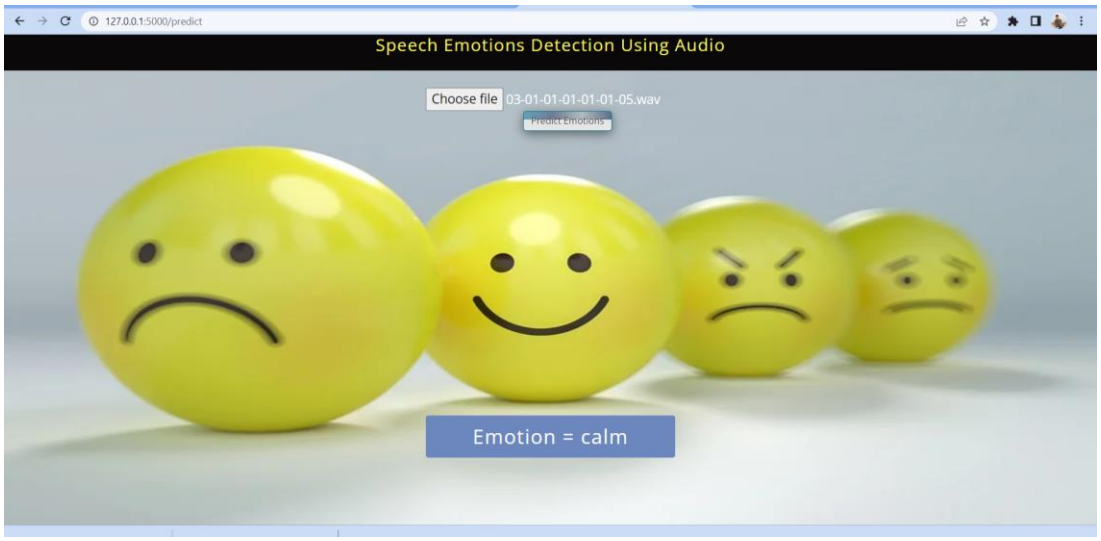


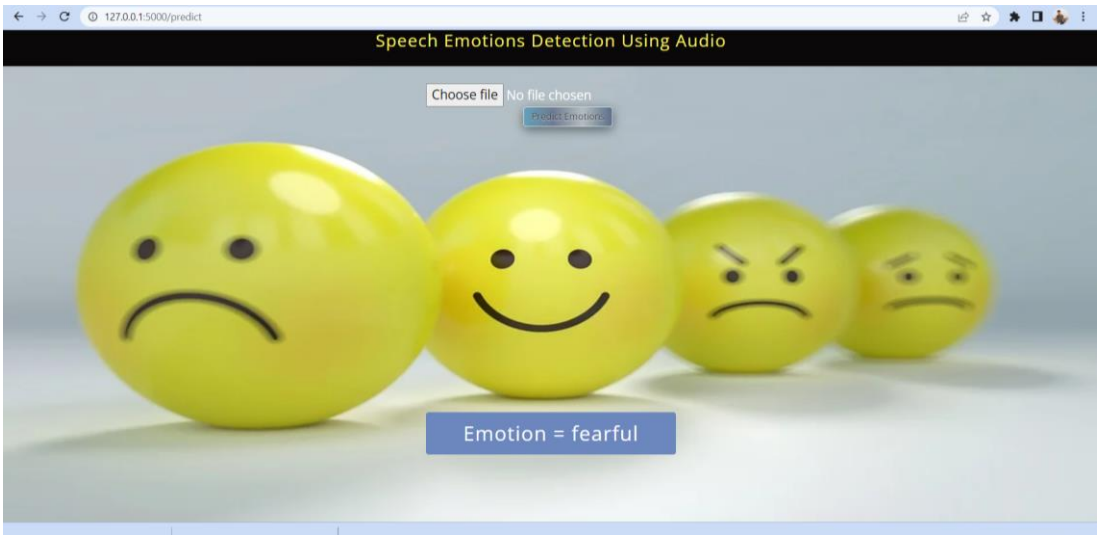
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

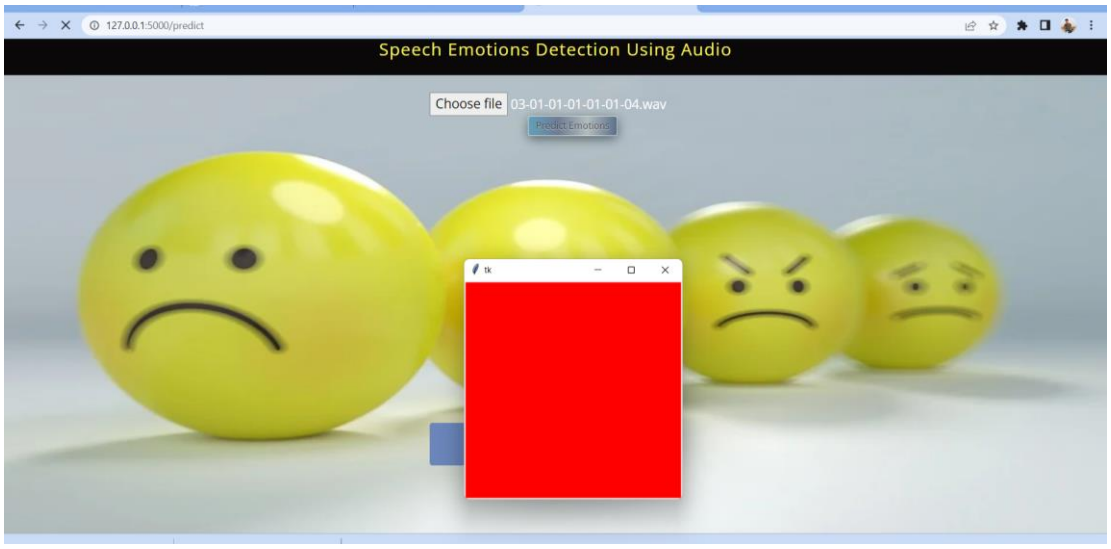
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456







IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

7.CONCLUSION

This paper gives a descent approach for Speech Emotion Recognition after studying various researches done by multiple researchers in this field. We have given a brief idea about how our system is going to work. Our proposed system aims to be very useful for the people suffering from Alexithymia or for the people suffering mild depression and understands ones emotion in a better way and deal with in a best way possible and create a ambience to improve or enhance a person's emotion.

We are hoping to do more research in this field and try and implement this system with more functionalities for helping more and more people we can.

8.REFERENCE

[1] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in INTERSPEECH, pp. 1916– 1920, 2013.

[2] S. R. Kadiri, P. Gangamohan, V. Mittal, and B. Yegnanarayana, "Naturalistic audio-visual emotion database," in 11th International Conference on Natural Language Processing, 2014, pp. 206.

[3] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speecha review," in Toward Robotic Socially Believable Behaving Systems, vol. 1, Springer, pp. 205–238, 2016.

[4] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using gmms," in Ninth International Conference on Spoken Language Processing, 2006.

[5] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech communication, vol. 48, no. 9, pp. 1162–1181, 2011.

[6] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[7] P. Gangamohan, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Excitation source features for discrimination of Anger and Happy emotions," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[8] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech communication, vol. 53, no. 5, pp. 768–785, 2011.

[9] F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," Journal of Experimental Psychology: Human Perception and Performance, vol. 12, no. 3, pp. 302, 1986.

[10] A. Bombatkar, G. Bhoyar, K. Morjani, S. Gautam, and V. Gupta, "Emotion recognition using speech processing using k-nearest neighbor algorithm," International Journal of Engineering Research and Applications (IJERA) ISSN, pp. 2248–9622, 2014.

[11] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. "Support vector machines using GMM supervectors for speaker verification," IEEE signal processing letters, vol. 13, no. 5, pp. 308-311, 2006.

[12] J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, Jun. 1998, pp. 121-167.

[13] X. Cheng and Q. Duan, "Speech emotion recognition using gaussian mixture model," in The 2nd International Conference on Computer Application and System Modeling, 2012.

[14] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," Pattern Recognition, ICPR 2004, Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004.

[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[16] A. P. Varga, and R. K. Moore, "Hidden Markov model decomposition of speech and noise," Acoustics, Speech, and Signal Processing, 1990. ICASSP-90, 1990 International Conference on. IEEE, 1990.

[17] J. J. Hopfield, "Artificial neural networks," IEEE Circuits and Devices Magazine, vol. 4, no. 5, pp. 3-10, 1988.

[18] V. K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," in Consumer Communications and Networking Conference (CCNC), 2013 IEEE. IEEE, 2013, pp. 106–111.

[19] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," The Journal of the Acoustical Society of America, vol. 133, no. 5, pp. 3050–3061, 2013.

[20] V. K. Mittal and B. Yegnanarayana, "An automatic shout detection system using speech production features," in International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction. Springer, 2014, pp. 88–98.

[21] V. K. Mittal and A. Vuppala, "Changes in Shout Features in Automatically Detected Vowel Regions," in Proc. IEEE 11th International Conference on Signal Processing and Communication (SPCOM 2016), 12-15 Jun. 2016, IISc, Bangalore.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11456

[22] V. K. Mittal and A. K. Vuppala, "Significance of Automatic Detection of Vowel Regions for Automatic Shout Detection in Continuous Speech," in Proc. IEEE/ISCA 10th International Symposium on Chinese Spoken Language Processing (ISCSLP 2016), Tianjin, China, 17-20 Oct. 2016.

[23] V. K. Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," Computer Speech & Language, vol. 30, no. 1, pp. 99–115, 2015.

[24] V. K. Mittal and B. Yegnanarayana, "Study of changes in glottal vibration characteristics during laughter," in INTERSPEECH, 2014, pp. 1777–1781.

[25] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in Noh voices," The Journal of the Acoustical Society of America, vol. 137, no. 6, pp. 3411–3421, 2015.

[26] V. K. Mittal and B. Yegnanarayana, "An impulse sequence representation of the excitation source characteristics of nonverbal speech sounds," in Proc. SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies, 2016, pp. 69–74.

