International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

Smart Multi-Model Emotion Recognition System with Deep learning

Dr BJD Kalyani[1], Kopparthi Praneeth Sai[2], N M Deepika[3], Shaik Shahanaz[4], G Lohitha[5]
[1]Associate Professor, Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Dundigal, Hyderabad, India.
kjd_kalyani@yahoo.co.in
[2]Student of Master of Science, Department of Computer Science
Lamar University, USA
kopparthypraneeth47@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Dundigal, Hyderabad, India.
deepikaneerupudi390@gmail.com
[4]Assistant Professor, Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Dundigal, Hyderabad, India.
shaikshahanaz.in@gmail.com
[5]Assistant Professor, Department of Information Technology
Institute of Aeronautical Engineering
Dundigal, Hyderabad, India.
g.lohitha@iare.ac.in

Abstract—Emotion recognition is added a new dimension to the sentiment analysis. This paper presents a multi-modal human emotion recognition web application by considering of three traits includes speech, text, facial expressions, to extract and analyze emotions of people who are giving interviews. Now a days there is a rapid development of Machine Learning, Artificial Intelligence and deep learning, this emotion recognition is getting more attention from researchers. These machines are said to be intelligent only if they are able to do human recognition or sentiment analysis. Emotion recognition helps in spam call detection, blackmailing calls, customer services, lie detectors, audience engagement, suspicious behavior. In this paper focus on facial expression analysis is carried out by using deep learning approaches with speech signals and input text.

Keywords- Deep Learning, Facial expressions, Multi-modal emotion recognition, Speech,Text.

I. INTRODUCTION

Human computer interaction become popular with recent trends [1] in the field of computer science and engineering includes AI, Machine learning, big data, natural language processing and computer enabled tools for recognition with robotics. Many researchers focused on human computer interaction, human activity recognition, speech recognition, human emotional recognition and sentiment analysis. As a result, the efficiency and accuracy of fake news detection, suicidal tendency recognition [2] and emotion recognition is greatly increased. Affective computing with multi-model recognition transforms information from variety of sources includes text, speech and facial recognition. The limitations of unimodal recognition like flow of time, lack of performance and accuracy is reduced with the multi-model recognition.

Generally multi-model emotion recognition systems utilize feature extraction methods like Gaussian Model, Natural language processing, automata and Hidden Markov Model [3]. This paper concentrates on deep learning-based approaches, convo-luted neural networks in emotion recognition. Humans will naturally communicate their personal emotions in social contexts. Mutual understanding and trust can be built through having a clear knowledge of one other's emotions. Humans require the ability to express and comprehend their emotions, primarily communicate personal feelings through words, voice, and facial expressions. Facial expressions are the most essential way of expressing human emotion information, according to re-searchers. Facial expression data accounts for around 55% of the data transmitted by the experimenters, voice data for 38%, and language data for only 7% of the overall data [4]. When

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

compared to language and music, it's evident that facial expression in-formation is more significant for emotional to understand and recognition.

II. RELATED WORK

Aasthajoshi [5] et al., describes that speech is an interactive interface medium as it is possible to express emotions and attitude through speech. A hybrid or the combi-nation of Hidden Markov Models (HMMs) and Support Vector Machines (SVM) has been proposed for better classification of emotions like happy, angry, sad and aggressive. The combining advantage on capability to the dynamic time warping, which export the likelihood probabilities and optimal state sequences, have been used to model the speech feature sequences.

Ankursupra [6] et al., carried his research done on audio recordings from Ryerson Audio-Visual Database of Emotional Song and Speech (RAVDESS). The features considered in research are Mel-Frequency Cepstral Coefficients (MFCCs), Log-Mel Spectrogram, pitch and energy. The comparative study is done with CNN, LSTM and deep neural networks, achieved an accuracy of 68%. The features of literature review are described in Table 1.

Table 1: Features of Literature Review

Title	Author	Description
Emotion Recognition and Acoustic Analysis from Speech Signal	Chang-Hyun Park	In this paper, it is a pitch detection algorithm that is particularly robust for telephone speech
Speech Emotion Recognition Using Combined Features of HMM	Aastha Josh	In this paper, a hybrid of Hidden Markov Models (HMMs) has been proposed to classify four emotions viz. happy, angry, sad and aggressive.
Emotion Recognition from Speech	Nikhil Panwar, Sohan Panwa	The analyses were carried out on audio recordings from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). After pre-processing Mel-Frequency Cepstral Coefficients (MFCCs)

III. GAP ANALYSIS

The existing system cannot be applied on large datasets, not produced accurate analysis of input taken and give many unpredictable answers. In speech-based system, Speech sample is first passed through a gender reference database which is maintained for recognition of gender before it gets into the process. Statistical approach is followed taking pitch as feature for gender recognition. Based on the distance calculation of the analysis frame from the reference database, one can classify the frame as happy, anger or normal.

The proposed work, will have a multi modal emotion recognition system, where in that platform can accurately identify the emotion of a particular input. Applica-tion take speech, text, facial expression as input. In this system for speech recognition 1D CNN, 2D CNN [8] are used. Then the input will be converted into frames of each frame size 60ms for every 50ms which means there is overlapping of data for 10ms. This is because to mitigate data loss and frequencies are calculated based on pitch. Then those frequencies are going to be compared with the actual databases to identify the emotion. Text analysis is done based on 1D CNN, RNN, Long-Short-Term Memory (LSTM) [9], facial expression analysis is done by point wise convolution and depth wise convolution. The architecture of proposed application as in Figure 1.

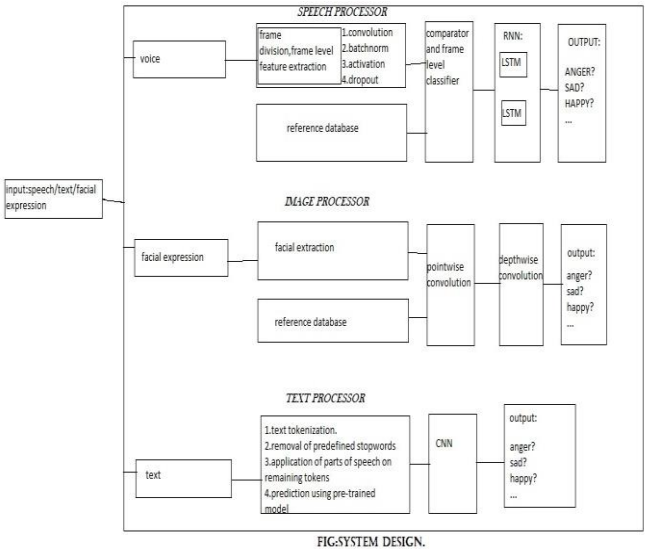

Figure 1: Architecture of Proposed System

IV. IMPLEMENTATION

The proposed system is a multi-modal emotion recognition modal, where the identifying of emotion will be accurate, less ambiguity, less error rate. Huge amounts of data can be taken as input. Inputs can be of 3 forms I.e., speech, text, facial expressions.as Modal is using Deep Neural Networking approaches it will be fast in identifying the correct emotion.

This emotion recognition helps in spam call detection, blackmailing calls, customer services, lie detectors, audience engagement, suspicious behavior by tracking identity, age, gender and current emotional state. By developing a virtual platform for interview preparation through which facilitates easy detection of emotions. The software used is flask, writing codes for each model like HTML, CSS, speech code, text, facial recognition code [10] and then loads the modal, later compares with the actual data base and recognizes the emotion. The data flow is illustrated in Figure 2.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

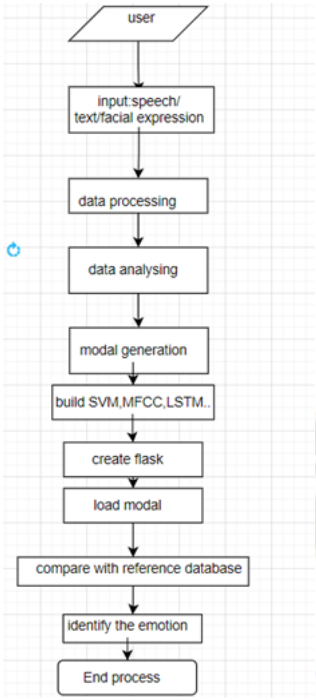

Figure 2: Data Flow Diagram

The following datasets are used for this research are:

- Stream-of-consciousness dataset from a study by Pennebaker and King [1999] [11] for the text input. There are a total of 2,468 daily writing submissions from 34 psychology students in this collection (29 women and 5 men whose ages ranged from 18 to 67 with a mean of 26.4).
- FER2013 Kaggle [12] Challenge data set for the video data sets. The data consists of grayscale images of faces at a resolution of 48x48 pixels.
- Ryerson Audio-Visual Database of Emotional Speech and Song for audio data sets (RAVDESS) [13]. There are 7356 files in this database (total size: 24.8 GB).

The analysis of three emotions considered in this research are carried out with the help of following steps:

4.1 Text Analysis

- Retrieving text data
- Preprocessing [14] of natural language on a per-case basis
- The document is tokenized.
- Regular expressions are used to clean and standardize compositions.
- Punctuation has been removed.
- Tokens in a lowercase
- Stop words that have been pre-defined have been removed.
- Part-of-speech tags are applied to the remaining tokens.
- To improve accuracy, tokens are lemmatized using part-of-speech tags.
- Using pre-trained model, can make predictions.

4.2 Video Analysis

- Start the webcam.
- Using a Histogram of Oriented Gradients [15], identify the face.
- Zoom in close to the face.
- Reduce the size of the face to 48 * 48 pixels.
- Make a face prediction using our pre-trained model.

4.3 Audio Analysis

- Recording of voice
- Discrimination of audio signals
- Extraction of the log-mel-spectrogram [16]
- A moving window was used to split the spectrogram.
- Use our pre-trained model picture to make a forecast.

V. RESULTS AND DISCUSSIONS

In this paper the proposed system is implemented using flask framework to create a web-application by combining all three models to make a multi-modal emotion detection application, after running the program in command prompt an URL link is generated as in Figure 3 which is uploaded in the google and then web-application will be displayed. The user's input is accepted in each stream for the detection of emotion.

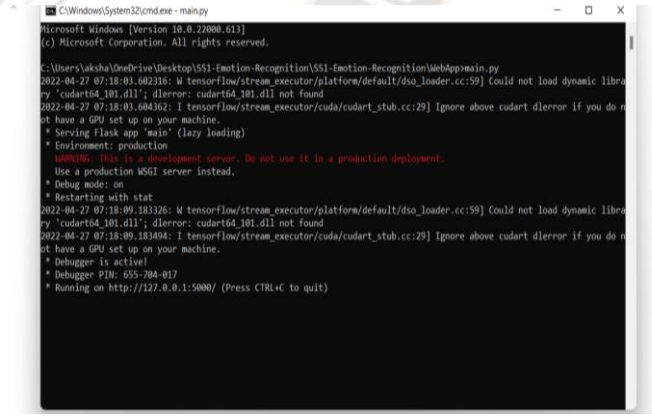

Figure 3: flask framework implementation

Multi-level emotion recognition is a web application, helpful for conducting virtual interviews. In this application audio emotion recognition is processed by using ensembled CNN and SVM [17] algorithms by taking audio speech as input. The text emotion detection is carried out by with CNN algorithm by taking raw text input and finally recognition based on facial expressions is done by CNN algorithm [18] by plotting

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

landmarks on the face. The Python libraries for the application implementation are as in Figure 4.

```
#!/usr/bin/python3
# -*- coding: utf-8 -*-

### General imports ###
from __future__ import division
import numpy as np
import pandas as pd
import time
import re
import os
from collections import Counter
import altair as alt

### Flask imports
import requests
from flask import Flask, render_template, session, request, redirect, flash, Response

### Audio imports ###
from library.speech_emotion_recognition import *

### Video imports ###
from library.video_emotion_recognition import *

### Text imports ###
from library.text_emotion_recognition import *
from library.text_preprocessor import *
from nltk import *
from tika import parser
from werkzeug.utils import secure_filename
import tempfile
```

Figure 4: Import of Library Files

The results from the deployed multi-modal webapp are very natural just like a hu-man interviewing with another human. It is able to recognize every emotion based on user inputs. As it is developed for organizational purpose, the application consists of functionalities as in Figure 5.

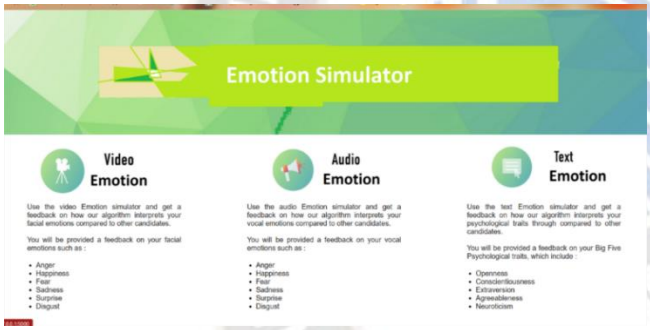

Figure 5: Functionalities of Application

5.1 Functionalities of Application

5.1.1 Emotion recognition based on speech (Audio)

In order to get the emotion of the candidate who is appearing for interview in a virtual mode based on his pitch, frequencies, speech signals will be processed by this option of the application as in Figure 6 and audio emotion recognition is illustrated by Figure 7.

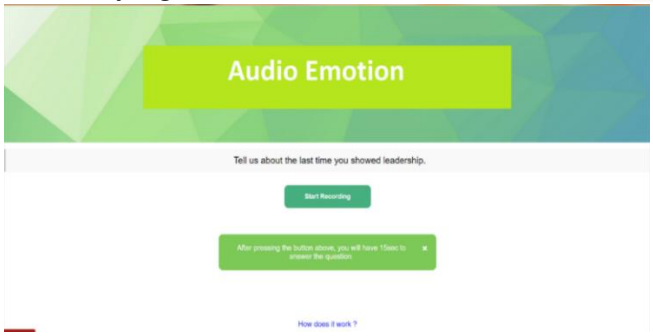

Figure 6: UI of Audio Emotion

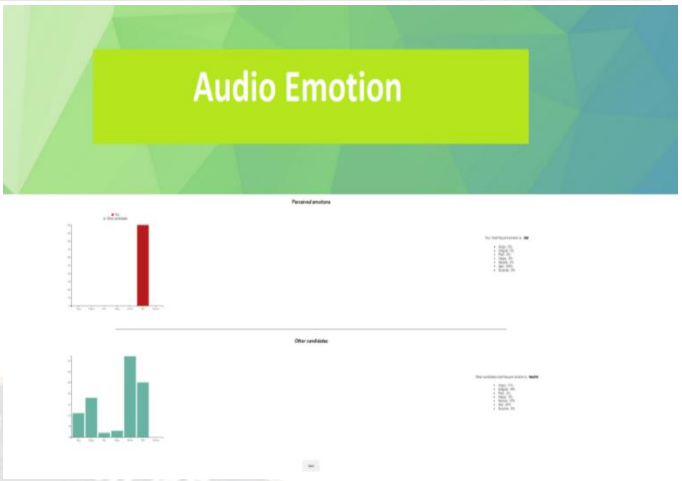

Figure 7: Audio Emotion Recognition

5.1.2 Emotion recognition based on facial expressions

In order to get the emotion of the candidate who is appearing for interview in a virtual mode based on his facial expressions that is by having landmarks plotted on his/her face as in Figure 8 by selecting through face option in the application. If user is not sitting under good light conditions or if the video quality [19] is poor then application warns to sit in well lighted area.

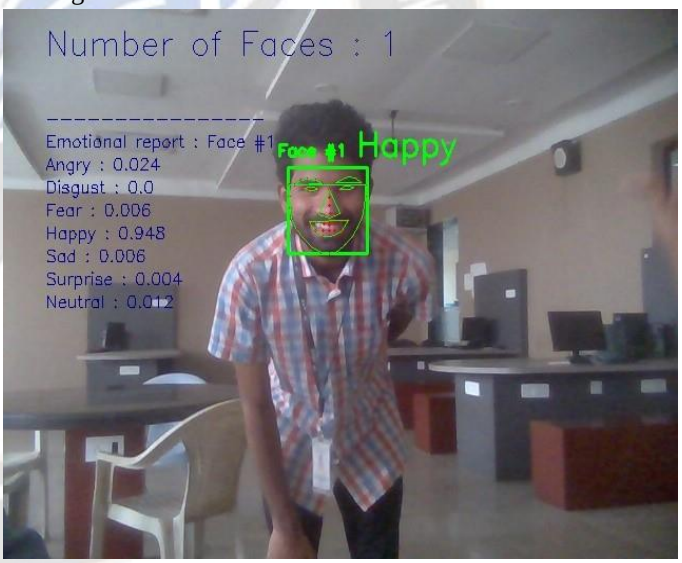

Figure 8: Facial Expression Recognition

5.1.3 Emotion recognition based on text

In this option the user upload or write a text live as in Figure 9 and get the results as in Figure 10 the emotions of the candidate giving the interview by undergoing some natural language processing [20], tokenizing [21] and all.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

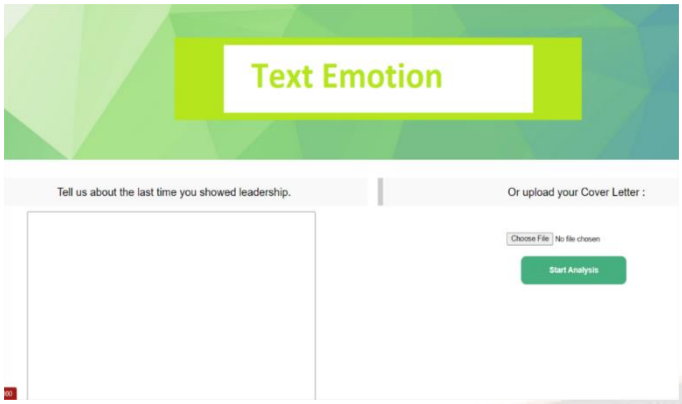

Figure 9: UI for Text Emotion

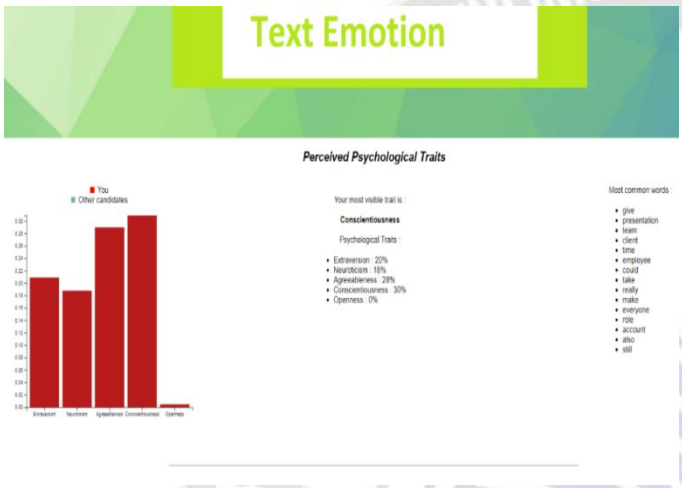

Figure 7: Text emotion recognition

The proposed system is trained using HMM algorithm [3] for emotions that are considered the three emotions as audio, video and facial expressions, while SVM has been introduced for making a decision for classification of gender. The recognition result of the hybrid classification has been compared with the isolated SVM and the maximum recognition rates have reached 98.1% and 94.2% respectively.

VI. CONCLUSION

Multi-Model emotion recognition is a web application using deep learning approaches. There is tremendous increase of data from all the different resources in real time are not supported by unimodal systems. Hence the multi-modal human emotion recognition platform is developed to handle three verities of data. This platform helps candidates who are taking virtual interviews and is more accurate than Unimodal systems.

The current study also opens up new avenues for future research, such as expanding the use these multi-modal emotion recognition in various sector like in lie detectors which may help police department at the time of punishing and getting truth from culprit, in customer services includes many e-commerce websites may use this to know the feedback from customers and many more. This will also expand the security in many areas where they are using artificial intelligence and machine learning.

REFERENCES

[1] AasthaJoshi, National Conference on August 2013, "Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm."

[2] AnkurSapra, Nikhil Panwar, and SohanPanwar, "Emotion Recognition from Speech," Volume 3, Issue 2, pp. 341-345, February 2013.

[3] Anagnostopoulos, C. N., Iliou, T., and Giannoukos, I. Anagnostopoulos, C. N., Iliou, T., and Giannou A survey from 2000 to 201 1[J] on features and classifiers for emotion recognition from speech. 155-177 in Artificial Intelligence Review, vol. 43, no. 2, 2015. (this is in the NETHERLANDS).

[4] BjörnSchuller, Manfred Lang, and Gerhard Rigoll, "Automatic Emotion Recognition by Speech Signal," National Journal, Volume 3, Issue 2, pp. 342-347, 2013.

[5] Bourlard, H., Konig, Y., Morgan, N., and others. [C]/ / 1996 8th European Signal Processing Conference. For hybrid HMM/ANN speech recognition systems, a new training strategy has been developed. IEEE, Trieste, Italy, 1996:1-4. (in Italy).

[6] Chang-Hyun Park and Kwee-Bo Sim, "Emotion Recognition and Acoustic Analysis from Speech Signal," in Chang-Hyun Park and Kwee-Bo Sim, "Emotion Recognition and Acoustic Analysis from Speech Signal," in Chang-Hyun Park and K Q2003 IEEE, International Journal on 2003, volume 0-7803-7898-9/03.

[7] Chao Wang and Stephanie Seneff's paper "Robust Pitch Tracking For Prosodic Modeling In Telephone Speech" was presented at the 2003 National Conference on "Big Data Analysis and Robotics"

[8] Chiu Ying Lay and Ng Hian James, "Gender Classification from Speech," in Chiu Ying Lay and Ng Hian James, "Gender Classification from Speech," in Chiu Ying Lay and N (2005).

[9] "Evaluation of expression recognition techniques," by Ira Cohen and colleagues. Retrieval of images and videos. 184-195 in Springer Berlin Heidelberg, 2003.

[10] IEEE Conference on 2004, Margarita Kotti and Constantine Kotropoulos, "Gender Classification In Two Emotional Speech Databases."

[11] In September 2001, Sony CSL Paris published "The creation and recognition of emotions in speech: characteristics and algorithms."

[12] Jason Weston, "Support Vector Machine and Statistical Learning Theory," International Journal, pp. 891-894, August 2011.

[13] M. El Ayaadi, F. Karrae, and M. S. Kamal are the first to mention M. El Ayaadi, F. Karrae, and M. S. Kamal. Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011. "Survey on emotion recognitions: Features, classification method, and database".

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 1
DOI: https://doi.org/10.17762/ijritcc.v11i1.6061
***Article Received**: 26 November 2022 **Revised**: 29 December 2022 **Accepted**: 05 January 2023*

[14] Mohammed E. Hoque1, Mohammed Yeasin1, and Max M. Louwerse2, "Robust Recognition of Emotion from Speech," International Journal, Volume 2, pp. 221-225,October 2011.

[15] Nobuo Sato and Yasunari Ohbuchi, "Emotion Recognition Using MFCCs," Information and Media Technologies 2(3):835-848 (2007), published from Journal of Natural Language Processing 14(4): 83-96. (2007).

[16] P. Ekman and W. Friesen, Consulting Psychologists Press, 1978, Facial Action Coding System: A Technique for the Measurement of Facial Movement.

[17] "Recognizing Emotion In Speech Using Neural Networks," IEEE Conference on "Neural Networks and Emotion Recognition" in 2013, Keshi Dai1, Harriet J. Fell1, and Joel MacAuslan2.

[18] Representing facial pictures for emotion classification, Padgett, C., and Cottrell, G. Advances in Neural Information.

[19] S.Wu, F. Li, and P. Zhang. DNN-based Emotional Recognition based on Weighted Feature Fusion for Variable-length Speech [C]/2019 15th International Conference on Wireless Communications and Mobile Computing (IWCMC). 674-679 in IEEE, 2019. (in the United States).

[20] T.L. Nwe', S W Foo, and L C De Silva, "Detection of Stress and Emotion in Speech Using Traditional And FFT Based Log Energy Features," in T L Nwe', S W Foo, and L C De Silva, "Detection of Stress and Emotion in Speech Using Traditional And FFT Based Log Energy Features," in T L Nwe', S 0-7803-8185-8/03 IEEE 0-7803-8185-8/03 IEEE 0-7803-8185-8/03 (200).

[21] T.L. Nwe, S. W. Foo, and L. C. De Silva. Hidden Markov models for speech emotion recognition[J] 603-623 in Speech Communication, vol. 41, no. 4, 2003. (this is in the NETHERLANDS).