

Performance levels based on speech emotion recognition using emoji detection by deep learning method

- 1.Podugu saikiran, III B.tech, Department of Electronics and Communication Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad; 19951a04d6@iare.ac.in
2. Dr.V.Padmanabha Reddy,Professor,Department of Electronics and communication Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad; v.padmanabhareddy@iare.ac.in
- 3.Dr.S. China Venkateswarlu ,Professor, Department of Electronics and Communication Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad; c.venkateswarlu@iare.ac.in

Abstract:

In our daily life we see different types of expressions and emotions in human communications and they are conveyed by using different expressions of human emotions. It could be body language, facial expressions, eye contact, laughter, and tone of voice. The languages of the world's peoples are different but we can identify the emotions of people by using their body language .People can almost understand part of the message that the other partner wants to say/Convey with emotional expressions

Human emotional expression is the most used expression of emotions through voice. This article presents our research on speech emotion recognition using emojiusing deep neural networks such as CNN, CRNN, and GRU. By using the Interactive Emotional Dyadic Motion Capture (IEDC) corpus for the study with four emotions:

Anger, happiness, sadness, clam, disgust, neutrality, the feature parameters used for recognition include the Mel spectral coefficients and other parameters related to the spectrum and the intensity of the speech signal. The data augmentation was used by changing the voice and adding white noise. The results show that the GRU model gave the highest average recognition accuracy of 74.47%. This result is superior to existing studies on speech emotion recognition using emojiusing emojiwith the IEDC corpus.

KEYWORDS: Emotion speech recognition, IEDC, CNN, CRNN, GRU, data augmentation.

1.1 Introduction:

The main theme of this paper is summarised as follows. The fact that people have emotional expressions is one of the measures showing that human civilization is the highest. It can be said that only humans have very diverse emotional expressions. The expression of emotions can be through body language, eyes, facial expressions, voice, laugh, etc. Just one of them also corresponds to many different emotional forms. In direct or indirect communication, even if there is no corresponding communication image, the human voice both carries the content to be conveyed and at the same time expresses the emotional state of the person towards the communication content. Robots can do many things better than humans, but currently, the expression of emotions of robots, especially through voices, is far behind that of humans. Therefore, the research on speech emotion recognition using emoji plays an important role in promoting advances in human–machine interaction. A significant amount of emotional data with different languages has been built and, emotion-recognition studies have been conducted. Among the emotional corpus, IEDC is multimodal emotional dataset in English and has been used as data for research on emotion recognition. For emotion recognition, multimodal recognition can be combined—for example, by combining speech-signal recognition with image recognition (face recognition and body-language recognition) and natural language recognition with noting exclamation words. In the case of such a combination, for better recognition efficiency will be achieved. It can be said that human interaction is marked by affects (attitudes and emotions).

In this study, we limited ourselves to emotion recognition based only on speech signals, and we will present new research results on using deep neural networks for speech emotion recognition using emoji with IEDC. The remainder of the article is organized as follows. An overview of relevant studies in this section is our conclusion.

1.2 Literature Survey:

This research is surveyed and evaluated quite a significant number of studies on speech emotion recognition using emoji for different corpora including IEDC .

IEDC was a corpus collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). IEDC launched in 2008, and since then there have been many studies on emotions using this corpus. In general, for convenience of comparison, most studies performed recognition for the same four emotions even though IEDC has data for nine emotions (happiness, anger, sadness, frustration, surprise, fear, excitement, other, and a neutral state). Those four emotions are anger, happiness, sadness, and neutrality.

For the happiness emotion, some studies consider excitement as happiness or combine excitement and happiness into a common emotion called happiness. On the other hand, according to happiness and excitement are close in the activation and valence domain. In our study, excitement is considered happiness. To be able to compare the performance of recognition systems using the IEDC corpus, we used four emotions of IEDC as other studies

have done. The construction of a system to identify all emotions by IEDC will be reserved for another study

(at the end of the article for convenience), They listed emotion-recognition studies with the IEDC corpus, and in the limited scope of this article, we only focus on speech emotion-recognition studies using IEDC speech data. The studies were listed mainly in recent years (2019 to 2021), with the remainder being a small number of studies from 2014 to 2018. From Table 14, we gave the models, the feature parameters, and the achieved recognition accuracy for each study. For the models that were used for emotion recognition, the vast majority of IEDC emotion-recognition studies have used neural-network models. The commonality of the studies listed is that there is no data augmentation for IEDC.

In, the authors used SVM to recognize four emotions from IEDC with an average accuracy of 74.9%. The studies listed from 2015 to now all used neural-network models. Studies using the LSTM model account for a fairly large proportion of the total number of studies. Besides, studies were using CNN in combination with LSTM. CNN, DCNN, and multi-channel CNN models were used. A combination of CNN and RNN models to get the CRN. The model used was based on attention-based convolutional neural networks (ACNN).

For the feature parameters that have been used for emotion recognition, some studies combine the features of speech and textual data. There are a large number of studies that have used a spectrogram, a Mel-spectrogram, or a combination of a spectrogram and a MFCC as feature parameters. For feature parameters are log spectra of short-time Fourier transforms. Besides the feature parameters mentioned above, several other features are used in combination such as chromagram, tonnetz, spectral contrast, pitch, spectral centroid, energy, zero-crossing rate, spectral flux, and spectral roll-off

1.3 Existing System:

In exiting method present the IEDC corpus for experiments, data augmentation, feature parameters, and deep neural network (DNN) models for our research. The last part of the section is a brief description of the performance parameters used to evaluate the research results. IEDC is a multimodal emotional corpus. Ten actors were recorded in dyadic sessions (five sessions with two subjects each). In total, the database contained approximately twelve hours of data. With this database, the authors hoped to be able to expand and generalize their results about the relationship and interplay between speech, facial expressions, head motion, and hand gestures during an expressive speech and conversational interactions. The distribution of the sample number for nine emotions is given in Figure 1. The sampling frequency of IEDC wav files was 16 KHz. With a frame width of 256 samples and a frame shift of 128 samples, the average number of frames per wav file was 372 for IEDC wav files. The frame shift was changed according to the sample number of the file. The smaller the

number frame in the file, the smaller the frame shift. For the critical case, i.e., where the minimum frame shift was 0, the duration of the corresponding file will then be equal to $256 \times 372.0 / \text{Sampling Frequency} = 5.92$ s. Wav files with a duration less than this value were disqualified. One such case is a wav file whose waveform is shown in Figure 2, the duration of which was 0.764 s.

2.2 Problem Statement:

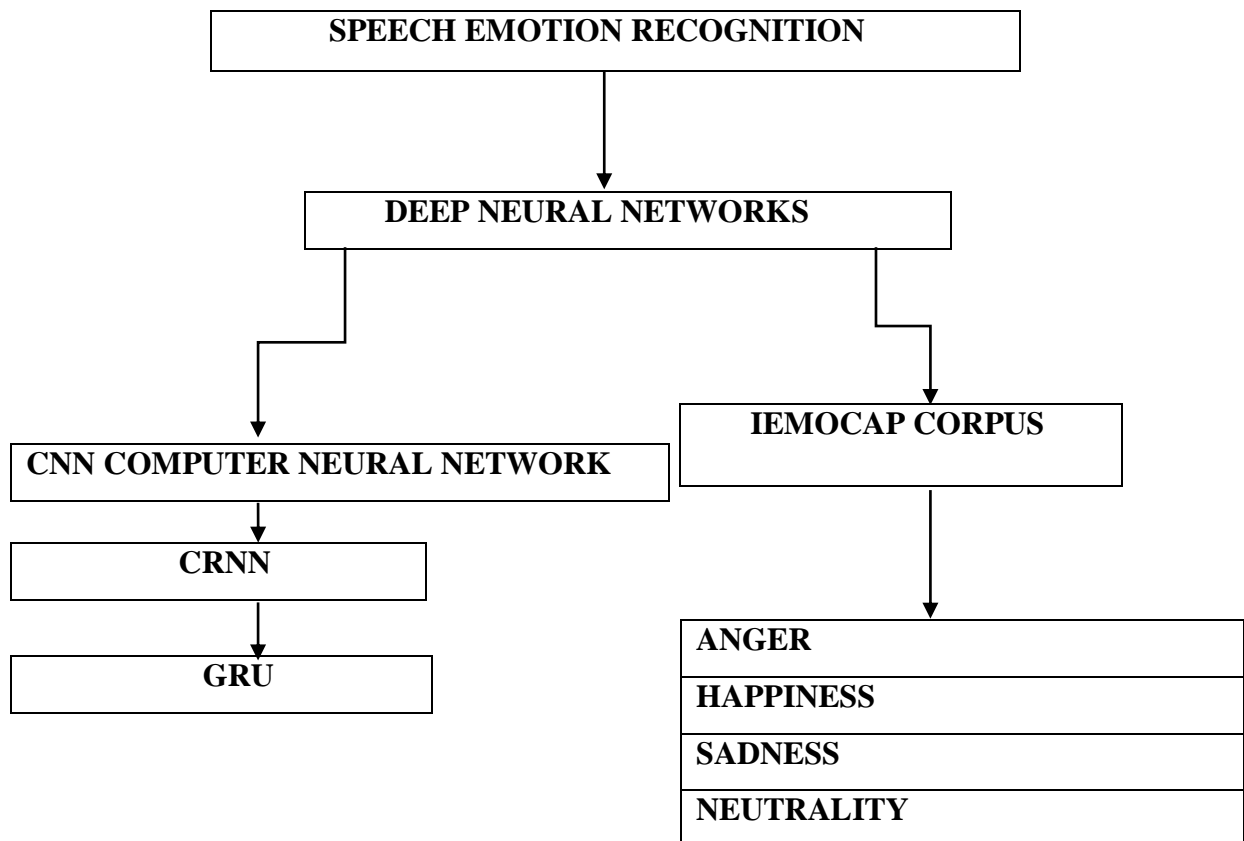
Among the studies listed have higher recognition accuracies. In the following,

- Here I present more closely these three studies.
- The recognition accuracies for four emotions (anger, happy, neutral, and sadness) in were 95.90%, 83.8%, and 81.75%, respectively.
- The common point of these studies is that they used CNN, and feature parameters were based on a spectrogram.
- The authors in assumed that individuals may use different means to express emotions and then that Speech emotion recognition using emoji(SER) should be conditioned on the speaker identity information.
- So, one of the major contributions of is that the authors have conditioned emotion classification to speaker identity by using a key-query-value attention called Self Speaker Attention (SSA)
- Which allows computing both self and cross-attribute (relation between speaker identity and emotions) attention scores to focus on the emotion-relevant parts of an utterance. For feature parameters, used the 3-D Log-Mel spectrogram that consists of a three-channel input.
- The first channel is the static of the Log-Mel spectrogram from 40 filter banks; the second and third channels are deltas and delta-deltas, respectively, which can be considered as approximations of the first and second derivatives of the first channel.
- For evaluations, a 10-fold cross-validation technique was performed. There was no data augmentation in. The main contributions of are that the authors proposed an algorithm using a DCNN to extract emotional features for SER and a Correlation-based Feature Selection (CFS) algorithm, which led to improved accuracy for SER.

For data, used a supervised resampling filter to oversample the minority class (oversampling increases the number of samples in the minority class).

- The authors in applied a ten-fold cross-validation technique to their evaluations. The data were randomly split into 10 equal parts for training and testing processes with a splitting ratio of 90:10.

2.3 BLOCK DIAGRAM OF EXISTING METHOD:

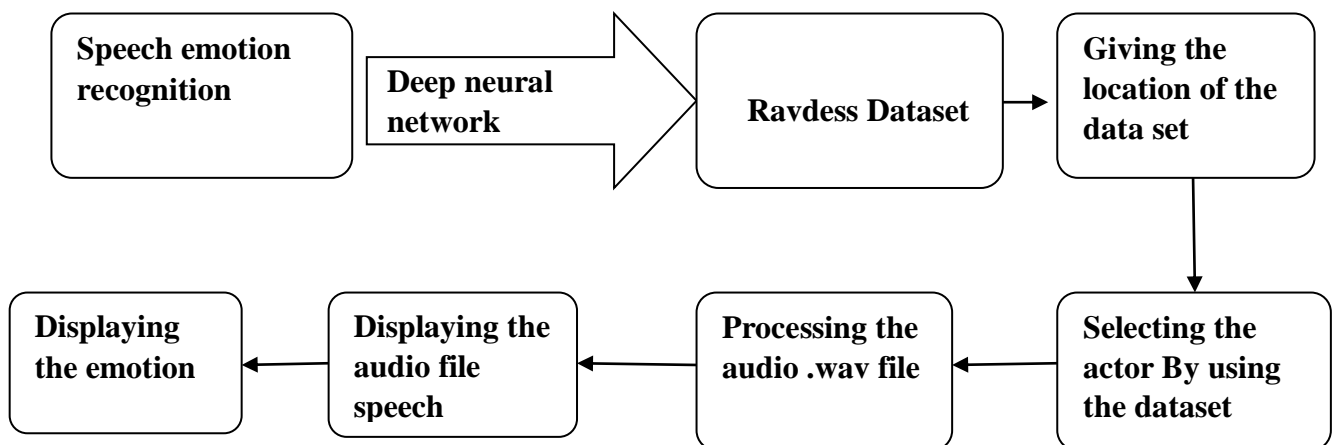


2.4 Proposed System:

Researching emotions recognition on data with a greatest number of emotions by using the different actor voices and emotions here I am using 25 actor's voices to determine the speech emotion by using various pitches to determine the voice emotion and representing the voice

and speeches by using python speech recognition method. In this research in using the Ravdess speech emotion data set which contains 2000 different types voices with n number of speech emotions.

2.5 block diagram



In the proposed method the input is given as a .wav file. The python converts the given data into the audio signal then calculated. After computing the .wav audio signal of the actor time and frequency domain plots are calculated and stored. Then this signal is will generate the emotions based on the frequency of the actor. Then the audio file is analysed. Then filters the emotions and unfiltered signals are obtained in both frequency and then gives the output of the emotion.

3.1 SOFTWARE USED:

This project is done by using the Google Colab in Python language

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

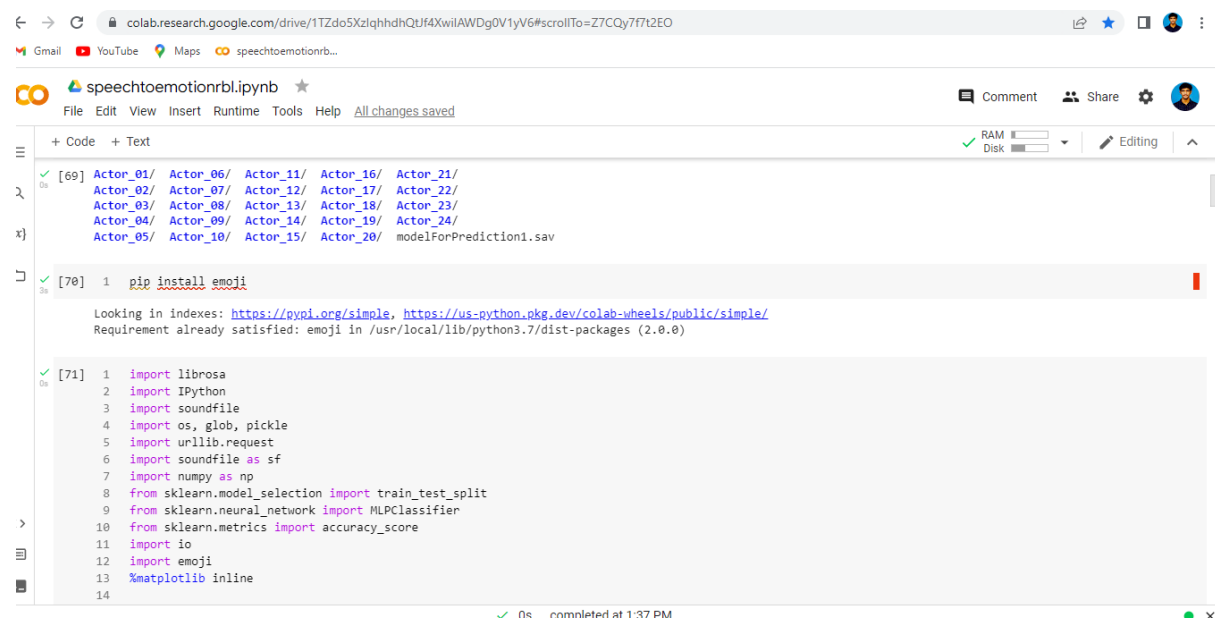
- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier.

Colab notebooks allow you to combine **executable code** and **rich text** in a single document, along with **images**, **HTML**, **LaTeX** and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them. To learn more, see Overview of Colab. To create a new Colab notebook you can use the File menu above, or use the following link: create a new Colab notebook.

Colab notebooks are Jupyter notebooks that are hosted by Colab. To learn more about the Jupyter project, see jupyter.org.

3.1.1 PRACTICAL SETUP:



The screenshot shows a Google Colab notebook interface. The browser address bar displays the URL: `colab.research.google.com/drive/1TZdo5XzqlqhdhQtUf4XwIiAWDg0V1yV6#scrollTo=Z7CQy7f7t2EO`. The notebook title is "speechtoemotionrbl.ipynb". The code editor shows the following code:

```
[69] Actor_01/ Actor_06/ Actor_11/ Actor_16/ Actor_21/
      Actor_02/ Actor_07/ Actor_12/ Actor_17/ Actor_22/
      Actor_03/ Actor_08/ Actor_13/ Actor_18/ Actor_23/
      Actor_04/ Actor_09/ Actor_14/ Actor_19/ Actor_24/
      Actor_05/ Actor_10/ Actor_15/ Actor_20/ modelForPrediction1.sav

[70] 1 pip install emoji

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: emoji in /usr/local/lib/python3.7/dist-packages (2.0.0)

[71] 1 import librosa
      2 import IPython
      3 import soundfile
      4 import os, glob, pickle
      5 import urllib.request
      6 import soundfile as sf
      7 import numpy as np
      8 from sklearn.model_selection import train_test_split
      9 from sklearn.neural_network import MLPClassifier
     10 from sklearn.metrics import accuracy_score
     11 import io
     12 import emoji
     13 %matplotlib inline
     14
```

The status bar at the bottom indicates "0s completed at 1:37 PM".

Figure1: practical setup using python – Google Colab

The above image represents about the practical setup of the system. In this image we can see Google Colab code editor with the code in it. The left side of the image indicates about the

files which are being used the right side of the image contains code in it. We can find the run option or execution option at the top right corner as Runtime on the screen.

4.1.1 CASE 1:

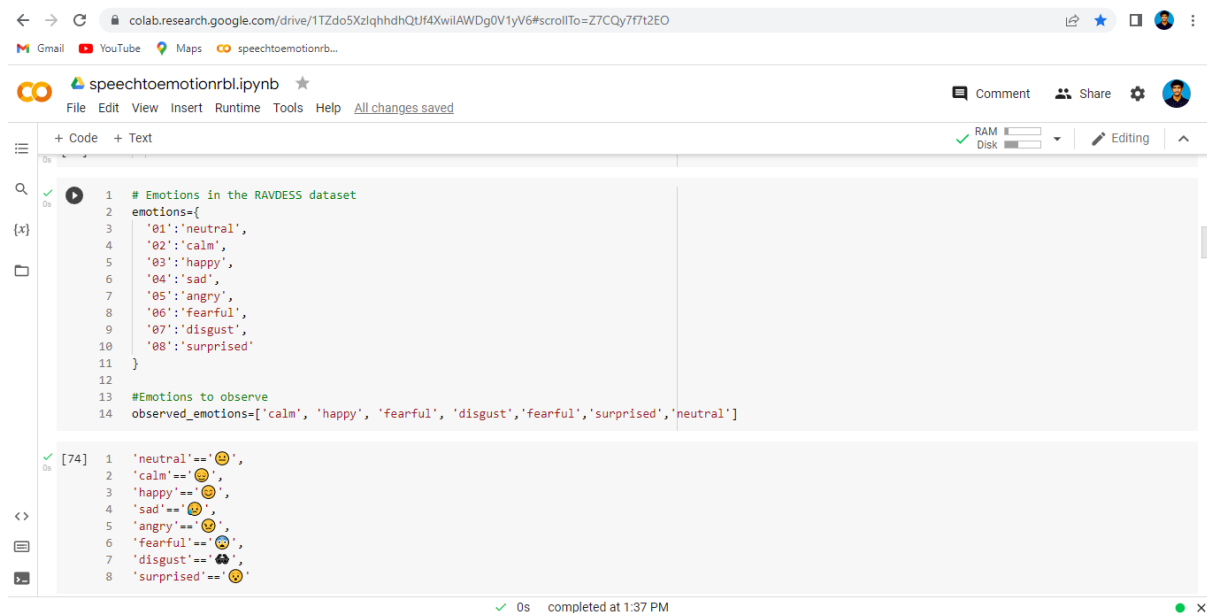
The below figure represents about the obtained output value for the given input and dataset. The output can be seen in the terminal which is in the down dialog box of the image represents after entering the code in the editor and installing the required libraries we run or execute the code. At down of the image a dialog box has opened which indicates the execution of the code and in the terminal section we can observe the output.

```
[72] 1 #Extract features (mfcc, chroma, mel) from a sound file
2 def extract_feature(file_name, mfcc, chroma, mel):
3     with soundfile.SoundFile(file_name) as sound_file:
4         X = sound_file.read(dtype="float32")
5         sample_rate=sound_file.samplerate
6         if chroma:
7             stft=np.abs(librosa.stft(X))
8             result=np.array([])
9         if mfcc:
10            mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
11            result=np.hstack((result, mfccs))
12        if chroma:
13            chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
14            result=np.hstack((result, chroma))
15        if mel:
16            mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
17            result=np.hstack((result, mel))
18    return result

1 # Emotions in the RAVDESS dataset
2 emotions={
3     '01':'neutral',
4     '02':'calm',
5     '03':'happy',
```

Figure2: Preparation of source code and simulation process started

The above image represents after entering the code in the editor and installing the required libraries we run or execute the code. At the bottom of the image a dialog box has opened which indicates the execution of the code and in the terminal section we can observe the output



```
1 # Emotions in the RAVDESS dataset
2 emotions={
3     '01':'neutral',
4     '02':'calm',
5     '03':'happy',
6     '04':'sad',
7     '05':'angry',
8     '06':'fearful',
9     '07':'disgust',
10    '08':'surprised'
11 }
12
13 #Emotions to observe
14 observed_emotions=['calm', 'happy', 'fearful', 'disgust','fearful','surprised','neutral']

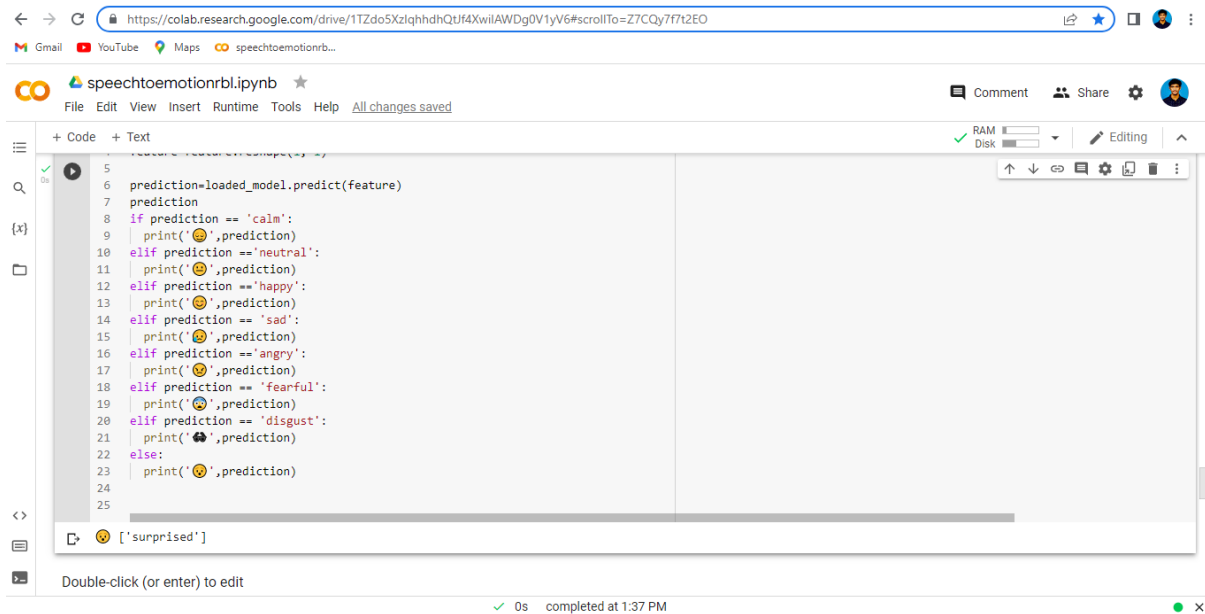
[74] 1 'neutral'==😐,
2 'calm'==😌,
3 'happy'==😄,
4 'sad'==😞,
5 'angry'==😡,
6 'fearful'==😱,
7 'disgust'==😔,
8 'surprised'==😲
```

Figure 3: different types of emotions mentioned in the dataset

The above image represents about the practical setup of the system. In this image we can see visual code studio editor with the code in it. The left side of the image indicates about the files which are being used the right side of the image contains code in it. We can find the run option or execution option at the top right corner of the screen.

Here in the above figure I mentioned different types of emotions using Ravdess dataset for recognizing the speech emotion using different actors dataset here emotions included are neutral, calm, happy, sad, angry, fearful, disgust, surprise.

All this expressions are observed using the sound file detection in the ravdess dataset.



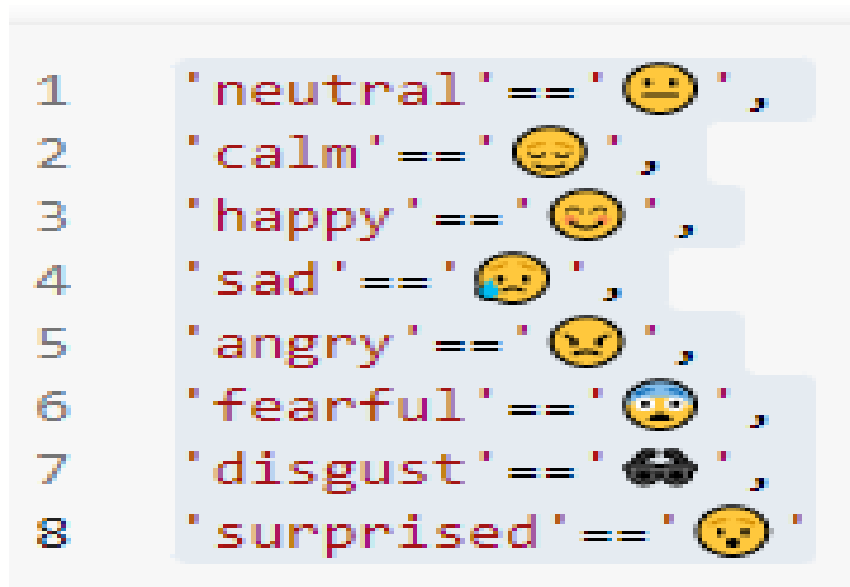
```
5 prediction=loaded_model.predict(feature)
6 prediction
7
8 if prediction == 'calm':
9     print('😊',prediction)
10 elif prediction == 'neutral':
11     print('😐',prediction)
12 elif prediction == 'happy':
13     print('😄',prediction)
14 elif prediction == 'sad':
15     print('😞',prediction)
16 elif prediction == 'angry':
17     print('😡',prediction)
18 elif prediction == 'fearful':
19     print('😱',prediction)
20 elif prediction == 'disgust':
21     print('🤢',prediction)
22 else:
23     print('😲',prediction)
24
25
```

Output: 😲 ['surprised']

Double-click (or enter) to edit

0s completed at 1:37 PM

Figure 4. The above image represents the output emotion obtained for the given input. The obtained emotion emoji are in ravdess dataset domain. The emoji represents emotion of the .wavfile signal.



1	'neutral' == '😐',
2	'calm' == '😊',
3	'happy' == '😄',
4	'sad' == '😞',
5	'angry' == '😡',
6	'fearful' == '😱',
7	'disgust' == '🤢',
8	'surprised' == '😲',

Figure 5. Here the above figure shows the different emots for different speech expressions

5. CONCLUSION:

This method is very helpful to detect the speech emotion using the trend like using the emoji's to express the emotions , and the accuracy of finding the emotion is quite good for finding the different expressions based on the audio files by using this technology we can command the robots and other AI machines to know the voice expressions while we are interacting the machines .

REFERENCES:

1. Scherer, K.R.; Ellgring, H. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion* 2007, 7, 158–171. [CrossRef] [PubMed]
2. Delattre, P. Les dix intonations de base du français. *Fr. Rev.* 1966, 40, 1–14.
3. Mac, D.K.; Castelli, E.; Aubergé, V.; Rilliard, A. How Vietnamese attitudes can be recognized and confused: Cross-cultural perception and speech prosody analysis. In *Proceedings of the 2011 International Conference on Asian Language Processing*, Penang, Malaysia, 15–17 November 2011; pp. 220–223.
4. Scherer, K.R.; Banse, R.; Wallbott, H.G. Emotion inferences from vocal expression correlate across languages and cultures. *J. Cross-Cult. Psychol.* 2001, 32, 76–92. [CrossRef]
5. Danes, F. Involvement with language and in language. *J. Pragmat.* 1994, 22, 251–264. [CrossRef]
6. Shigeno, S. Cultural similarities and differences in the recognition of audio-visual speech stimuli. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP98)*, Sydney, Australia, 30 November–4 December 1998; p. 1057.
7. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmúlik, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 2021, 10, 1163. [CrossRef]
8. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 2008, 42, 335–359. [CrossRef]
9. Chen, S.; Jin, Q.; Li, X.; Yang, G.; Xu, J. Speech emotion classification using acoustic features. In *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing*, Singapore, 12–14 September 2014; pp. 579–583.

10. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Schuller, B.W. Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition. In Proceedings of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 2327–2331.
11. Valter Filho, M.; Souza, M. Interaffection of Multiple Datasets with Neural Networks in Speech Emotion Recognition. In Proceedings of the 17th National Meeting on Artificial and Computational Intelligence, Porto Alegre, Brasil, 20–23 October 2020; pp. 342–353.
12. Yu, Y.; Kim, Y.J. Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database. *Electronics* 2020, 9, 713. [CrossRef].
13. Krishna, D.N.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In Proceedings of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 4243–4247.
14. Lu, Z.; Cao, L.; Zhang, Y.; Chiu, C.C.; Fan, J. Speech sentiment analysis via pre-trained features from end-to-end asr models. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7149–7153.
15. Chen, F.; Luo, Z.; Xu, Y.; Ke, D. Complementary fusion of multi-features and multi-modalities in sentiment analysis. In Proceedings of the 3rd Workshop of Affective Content Analysis, New York, NY, USA, 7 February 2020; pp. 82–99.
16. Li, R.; Wu, Z.; Jia, J.; Bu, Y.; Zhao, S.; Meng, H. Towards Discriminative Representation Learning for Speech Emotion Recognition. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 5060–5066.
17. Cai, R.; Guo, K.; Xu, B.; Yang, X.; Zhang, Z. Meta Multi-Task Learning for Speech Emotion Recognition. In Proceedings of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 3336–33418.
18. Dangol, R.; Alsadoon, A.; Prasad, P.W.C.; Seher, I.; Alsadoon, O.H. Speech Emotion Recognition Using Convolutional Neural Network and Long-Short Term Memory. *Multimed. Tools Appl.* 2020, 79, 32917–32934. [CrossRef]
19. Tripathi, S.; Beigi, H. Multi-modal emotion recognition on IEMOCAP with neural networks. *arXiv* 2008, arXiv:1804.05788.

20. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
21. Zheng, S.; Du, J.; Zhou, H.; Bai, X.; Lee, C.H.; Li, S. Speech Emotion Recognition Based on Acoustic Segment Model. In Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–26 January 2021; pp.
22. Tripathi, S.; Ramesh, A.; Kumar, A.; Singh, C.; Yenigalla, P. Learning Discriminative Features using Center Loss and Reconstruction as Regularizer for Speech Emotion Recognition. In Proceedings of the Workshop on Artificial Intelligence in Affective Computing, Macao, China, 10 August 2019; pp. 44–53.
23. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* 2020, 59, 101894. [CrossRef]
24. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep learning-based emotion recognition system using speech features and transcriptions. In Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle.
25. France, 7–13 March 2019. 25. Li, X.; Song, W.; Liang, Z. Emotion Recognition from Speech Using Deep Learning on Spectrograms. *J. Intell. Fuzzy Syst.* 2020, 39, 2791–2796. [CrossRef]
26. Xu, M.; Zhang, F.; Khan, S.U. Improve accuracy of speech emotion recognition with attention head fusion. In Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 1058–1064.
27. Scotti, V.; Galati, F.; Sbattella, L.; Tedesco, R. Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition. In Proceedings of the International Conference on Pattern Recognition, Talca, Chile, 7–19 April 2022; Springer: Cham, Switzerland; pp. 114–128.

AUTHOR BIOGRAPHY:



STUDENT AT INSTITUTE OF AERONAUTICAL ENGINEERING

NAME : PODUGU SAIKIRAN
ROLL NUMBER : 19951A04D6
DEPARTMENT : ELECTRONICS AND COMMUNICATION
EMAIL ID : podugusaikiran1@gmail.com
MOBILE NUMBER : 9505595861



PROFESSOR AT INSTITUTE OF AERONAUTICAL ENGINEERING COLLEGE

NAME : DR. C PADMANABHA REDDY
FACULTY ID : IARE10622
MAIL ID : v.padmanabhareddy@iare.ac.in



PROFESSOR AT INSTITUTE OF AERONAUTICAL ENGINEERING COLLEGE

NAME : DR. S CHINA VENKATESWARLU
FACULTY ID : IARE10624
MAIL ID : c.venkateswarlu@iare.ac.in