Speech and Text Emotion Recognition using Machine Learning

AKHILA R[1],BHAVANA R[2], JEEVITHA N[3], MOUNIKA M[4],prof USMAN AIJAZ.N [5],Prof.DIVYA JYOTHI[6]

[1,2,3,4] Final Year Student, [5]Assistant Professor, [6]Assistant Professor Department of Information Science and Engineering, HKBK College of Engineering, Nagawara, Bengaluru, India

Email Id: 180588.is@hkbk.edu.in, usman.is@hkbk.edu.in,

Abstract— ***Speech emotion recognition is a challenging task, an extensive reliance as been placed on models that use audio features in building well-performing classifiers. In this project, we propose a model that utilizes the text data and audio signals to obtain a better understanding of speech data. As emotional dialogue is composed of sound and spoken content, our model encodes the information from audio and text sequences using neural networks and then combines the information from these sources to predict the emotion class. This architecture analyses speech data from the signal level to the language level, and it does utilize the information within the data more comprehensively then models that focus on audio features. Extensive experiments are conducted to investigate the efficiency and properties of the proposed model. Our proposed model outperforms previous state ofthe art methods in assigning data to one of the three emotion categories (i.e. angry ,happy, sad)when the model is applied to the data set.***

Keywords: *speech emotion recognition, deep learning, natural language processing.*

Recently, machine learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, speech recognition, text to speech generation and other machine learning related areas . similarly , substantial improvements in performance have been obtained when machine learning algorithms have been applied to statistical speech processing. In developing emotionally aware intelligence, the very first step is building robustt emotion classifiers that display good performance regardless of the application; this outcome is considered to be one of the fundamental research goals in affective computing. In particular, the speech emotion recognition task is one of the most important problems in the field of paralinguistics. This field has recently broadened its applications, as it is crucial factor in optimal human computer interactions, including dialog systems.

I. INTRODUCTION

Recently, machine learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, speech recognition, text to speech generation, and other machine learning related areas. Similarly, substantial improvements in performance have been obtained when machine learning algorithms have been applied to statistical speech processing. In developing emotionally aware intelligence, the very first step is building robust emotion classifiers that display good performance regardless of the application; this outcome is considered to be one of the fundamental research goals in effective computing. In particular, the speech emotion recognition task is one of the most important problems in the field of paralinguistic. This field has recently broadened its applications, as it is the crucial factor in optimal human computer interactions, including dialog systems.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model. Amongst the numerous models used for categorization of these emotions, a discrete emotional approach is considered as one of the fundamental approaches. It uses various emotions such as anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness another important model that is used is a three-dimensional continuous space with parameters such as arousal, valence, and potency.

The idea is to use an emotion mining from text classifier to predict the emotion or emotions expressed in the source utterance, then decide based on the detected emotions, which emotion e is expressed in the response. The response is evaluated using the same emotion classifier and is declared successful if e is predicted from the response. The emotion tagger we use is based on the work in but uses a deep learning model and trains on 9 emotions: anger, disgust, fear, guilt, joy, love, sadness, surprise, and thankfulness. These are based on the six basic emotions from Ekman's model to which we added guilt, love and thankfulness in the context of an open ended conversational agent that we aim to be emotionally intelligent for companionship to elderly users. a scientific perspective, recognition of emotions is nothing more than a mapping from a feature space to emotion descriptors or labels space. For the mapping between the two spaces, different machine learning algorithms have been used [30]. In general, theories to perform the mapping have solid analytical foundations and are well defined and validated. Hardly, however, is the same true for feature and emotion spaces. In other words, it is a challenging issue to determine which features to use and how to describe emotions.

The In Social emotions in nature and artifact: emotions in human and human-computer int reaction, S. Marsella J. Gratch, Ed. Oxford University Press, New York, problem of emotion recognsition from speech critically depends on these two factors, meaning that high and robust emotion recognition performance can be achieved only with the accurate selection of features and emotional labels. In this chapter, selection of correct features and emotional labels will be discussed in the view of building EASER systems.

The idea of recognizing emotions in speech has been of interest for many years. A quick search will produce many scientific publications on the topic. It is out of the scope of this chapter to present a detailed review of the existing literature. Instead, our focus is on selected aspects that are crucial for a robust speech emotion recognition system.

II. IMPLEMENTATION

2.1.1 ALGORITHMS :

Convolutional Neural Network Algorithm(CNN):

The gathered data should be processed and training using CNN algorithm in order to finish the training process soon. This involves 3 layers:

Convolutional layer:
Here, the extraction feature extraction will take place where only the useful features which are needed to the machine will be collected and unwanted features will be removed so that the training period will be finished soon.
Pooling layer:
In this, the size of the data or image will be reduced and it gives us a compressed document with important features which is needed for the machine.
Fully connected layer:
Here, the above data which we get from previous layer will befed to fully connected layer in a vector form. Then these compressed features will be split and get trained using CNN and will produce us the final output.

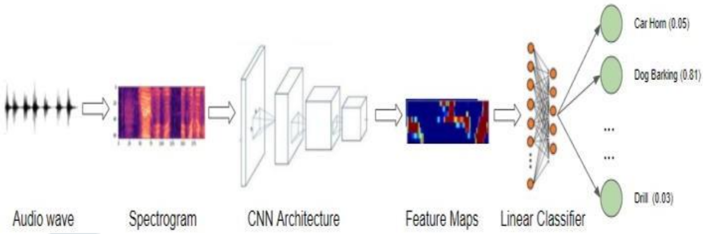


Fig : CNN algorithm

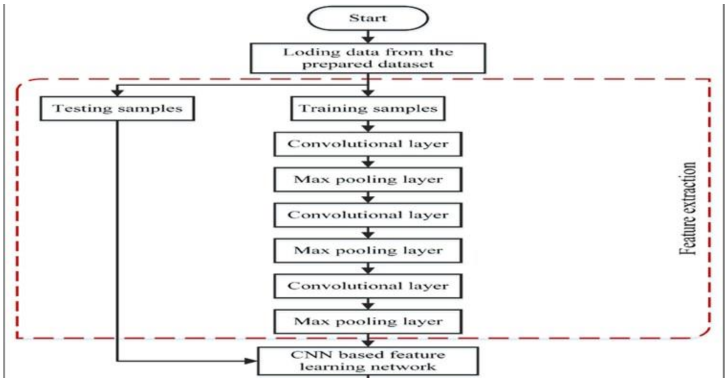


Fig: CNN flow in speech and emotion recognition

Multi-Layer Perceptron Algorithm:

An MLP is a supervised machine learning (ML) algorithm that belongs in the class of feedforward artificial neural networks.

STEP 1:Training data is propagated to the MLP through input layers. It passes through the hidden layers, if any forwarding outputs of activation functions to the next layer.Finally the output is generated at the output is generated at the output at the output layer by applying activation functions.

STEP 2:The predicted output will be compared with actual output and hence error will be calculated.

STEP 3:If error>0, apply backpropagation methodology to modify weights starting from output layer moving towards input layer.

STEP 4: Check accuracy score, if satisfied, stop, else goto step 1.

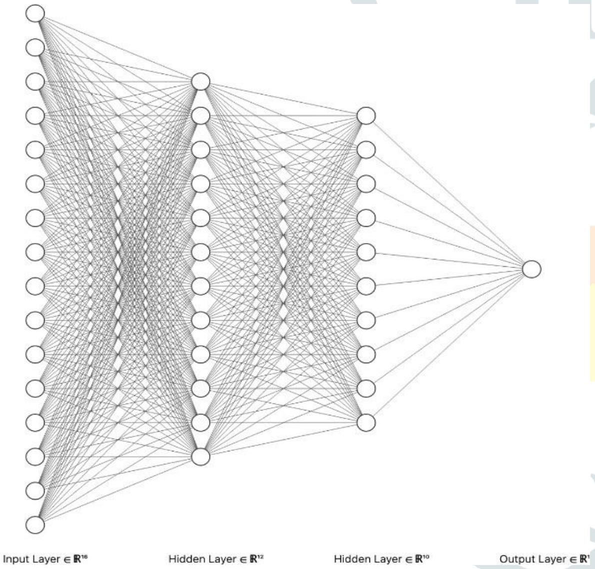


Fig: Multilayer perceptron tensor flow

2.1.2 Working Steps:

In this section we go through the flow of the working of the Speech and Text emotion recognition.

STEP 1:

First we need to train the data using CNN algorithm to the system.

STEP 2:

In the model, we need to record our voice/give input in textual format.

STEP 3:

Based on our voice intensity, the emotion of our speech will be recognized.

STEP 4:

And based on the textual features, emotion of the text will be recognized.

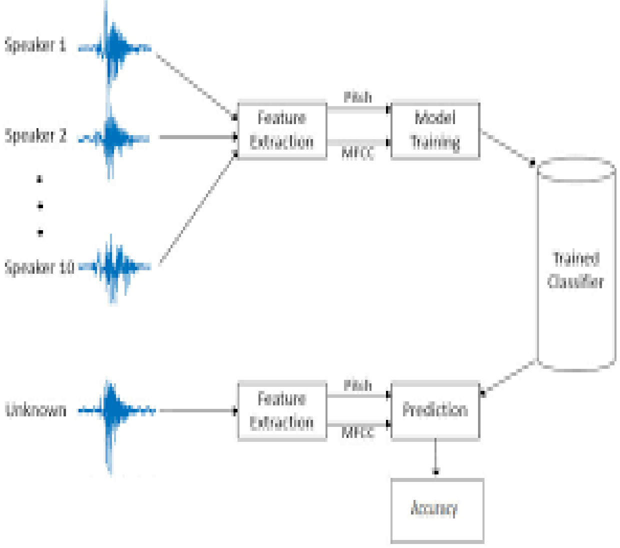


Fig:Work flow for the extraction and classification of emotion in speech.

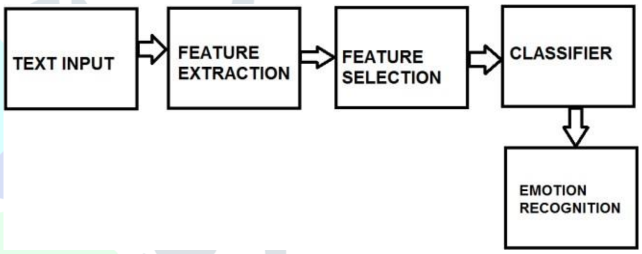


Fig:Work flow of the extraction and classification of emotion in text.

Ⅲ. SYSTEM ARCHITECTURE

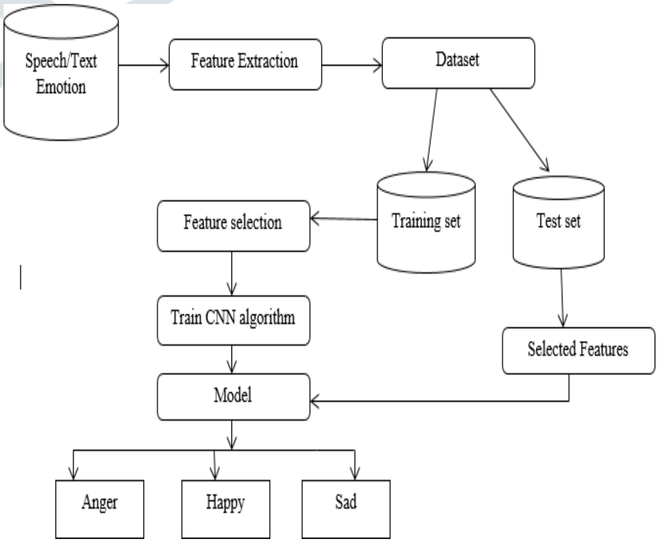


Fig: System Architecture

Initially, for the dataset we need to extract the features like voice intensity and keywords, and convert that into another dataset. Once we have extracted the important features (feature selection) for the machine to understand, we will train and test the dataset by using CNN. A model will be created to give the inputs for speech and text. Depending on the trained data, the model can predict whether the person is having which kind of emotion.

Problem Statement:

Existing System:

As emotional dialogue is composed of sound and spoken content, this model encodes the information from audio and text sequences using dual recurrent neural networks (RNNs) and then combines the information from these sources to predict the emotion class. This architecture analyzes speech data from the signal level to the language level, and it thus utilizes the information within the data more comprehensively than models that focus on audio features. Extensive experiments are conducted to investigate the efficacy and properties of the proposed model. The proposed model outperforms previous state-of-the-art methods in assigning data to one of four emotion categories (i.e., angry, happy, sad and neutral) when the model is applied to the IEMOCAP dataset, as reflected by accuracies ranging from 68.8% to 71.8%.

PROPOSED SYSTEM: The model consists of speech recognition where we need to record our voice. Based on the voice intensity and speech, it will predict the emotion. Another model consists of text based emotion recognition where the texts will be given as input and based on the keywords, it will predict our emotion. The dataset will be trained using CNN algorithm because of the high accuracy and the frontend will be developed using Flask framework.

CNN training: The gathered data should be processed and trained using CNN algorithm in order to finish the training process soon. This involves 3 layers: Convolution layer: Here, the feature extraction will take place where only the useful features which are needed to the machine will be collected and unwanted features will be removed so that training period will be finishedsoon. Pooling layer: In this, the size of the data or image will be reduced and give us a compressed document with important features which is needed for the machine.

Fully connected layer: Here, the above data which we get from the previous layer will be fed to fully connected layer in a vector form. Then these compressed features will be split and get trainedusing CNN and will produce us the final output. Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Data filtering is the process of choosing a smaller part of your data set and using that subset for viewing or analysis. Filtering is generally (but not always) temporary – the complete data set is kept, but only part of it is used for the calculation. Classification is a supervised learning problem: define a set of target classes and train a model to recognize. Based on the trained data, we can classify the results.

IV. TESTING AND OUTCOMES

Software testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect-free.It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements. Some prefer saying Software testing definition as a White Box and Black Box testing. In simple terms, Software Testing means the Verification of Application Under Test.

WHITE BOX TESTING:

White box testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It isused to test areas that cannot be reached from a blackbox level.

BLACK BOX TESTING:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested

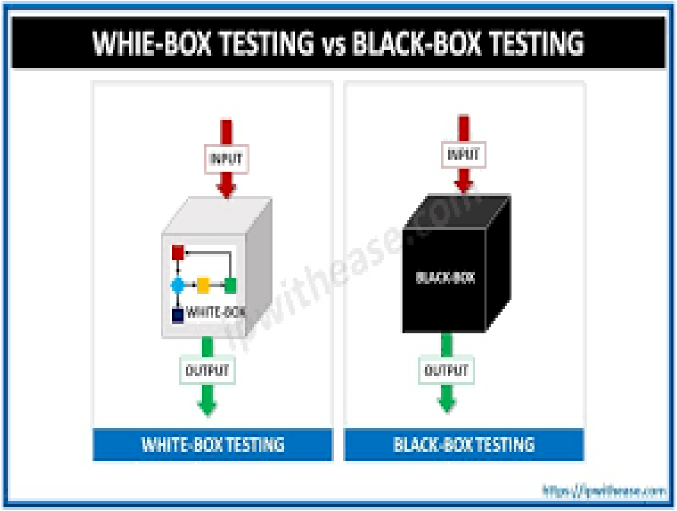
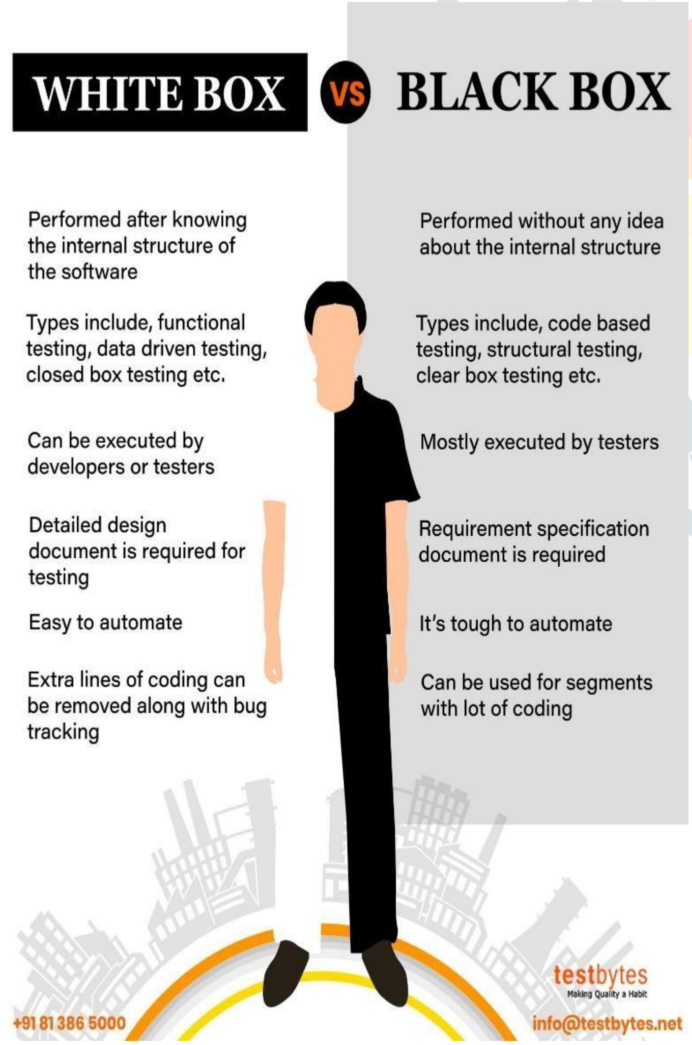


Fig: Black box and white box testing

Fig: Tabular differentiation between black andwhite



WHITE BOX	BLACK BOX
Performed after knowing the internal structure of the software	Performed without any idea about the internal structure
Types include, functional testing, data driven testing, closed box testing etc.	Types include, code based testing, structural testing, clear box testing etc.
Can be executed by developers or testers	Mostly executed by testers
Detailed design document is required for testing	Requirement specification document is required
Easy to automate	It's tough to automate
Extra lines of coding can be removed along with bug tracking	Can be used for segments with lot of coding

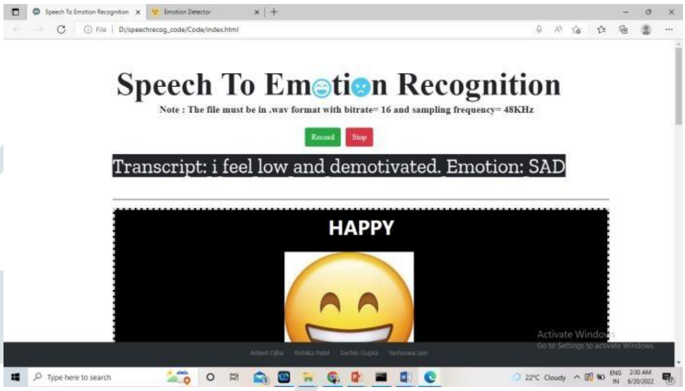
box testing.

Expected outcome of the project:

Once we record our voice in the website, it will check our voice intensity and predict what emotion the person is having. For text based, we provide the inputs in textual format, and based on the keywords in the text, emotion will be recognized.

Actual outcomes of the project:

SPEECH OUTCOMES**:**

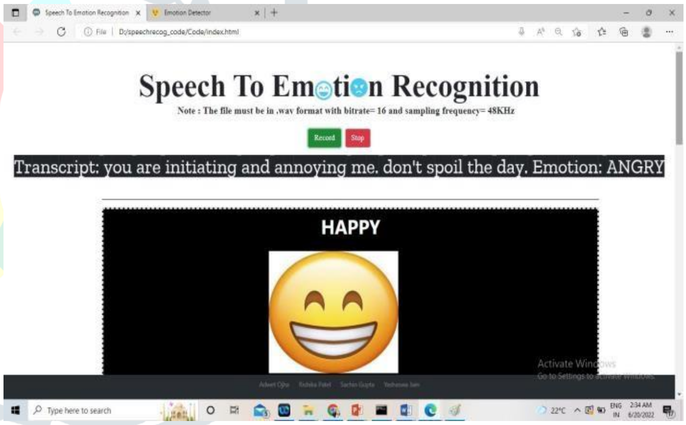


Speech To Emotion Recognition

Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: i feel low and demotivated. Emotion: SAD

Fig : SAD SER

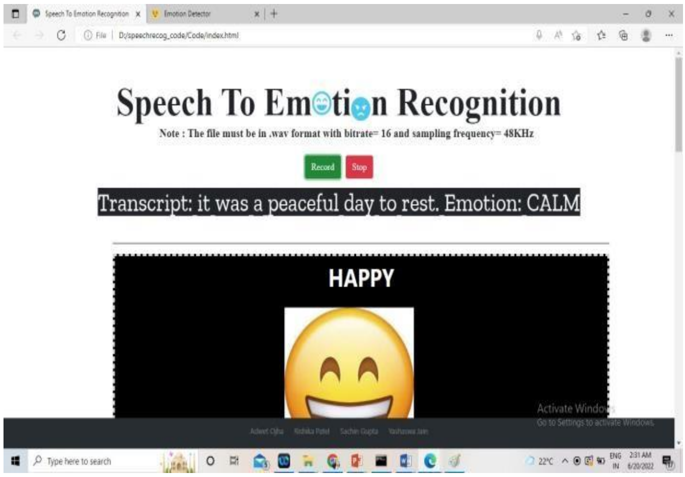


Speech To Emotion Recognition

Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: you are initiating and annoying me. don't spoil the day. Emotion: ANGRY

Fig: ANGRY SER

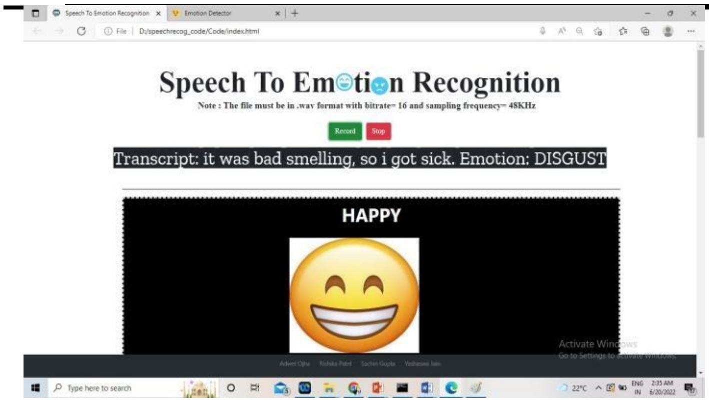


Speech To Emotion Recognition

Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: it was a peaceful day to rest. Emotion: CALM

Fig: CALM SER

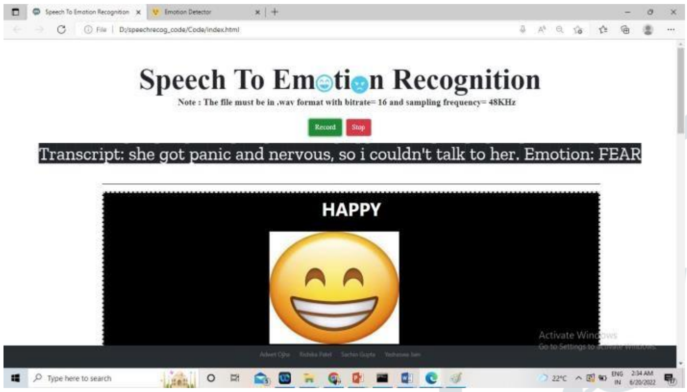

Speech To Emotion Recognition

Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: it was bad smelling, so i got sick. Emotion: DISGUST

HAPPY

Fig : DISGUST SER

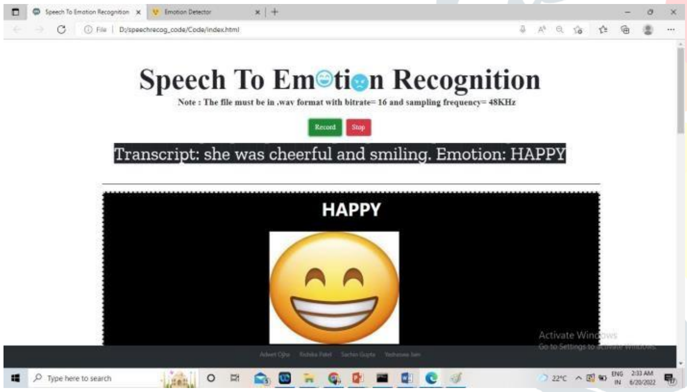

Speech To Emotion Recognition

Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: she got panic and nervous, so i couldn't talk to her. Emotion: FEAR

HAPPY

FIG: FEAR SER


Speech To Emotion Recognition

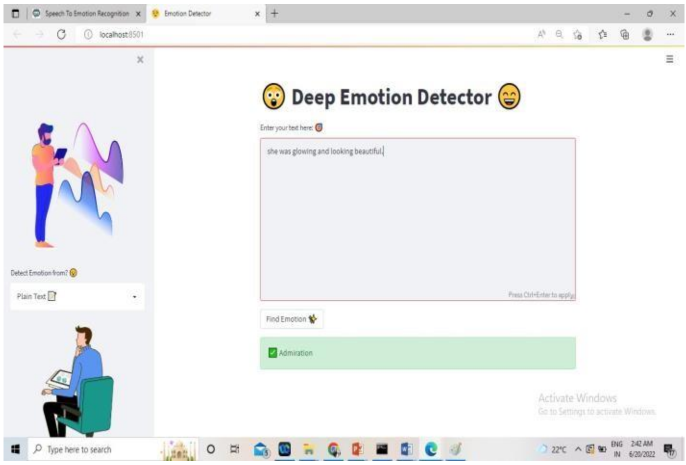
Note : The file must be in .wav format with bitrate= 16 and sampling frequency= 48KHz

Transcript: she was cheerful and smiling. Emotion: HAPPY

HAPPY

FIG: HAPPY SER

TEXT SNAPSHOTS:


Deep Emotion Detector

Enter your text here:

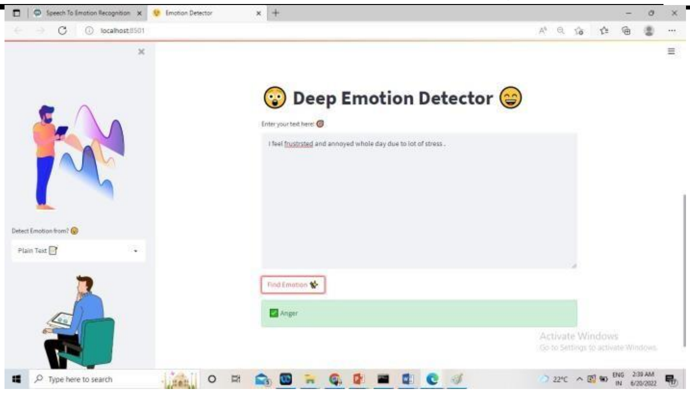
she was glowing and looking beautiful

Detect Emotion from?

Plain Text

Admiration

FIG: EMOTION ADMIRATION


Deep Emotion Detector

Enter your text here:

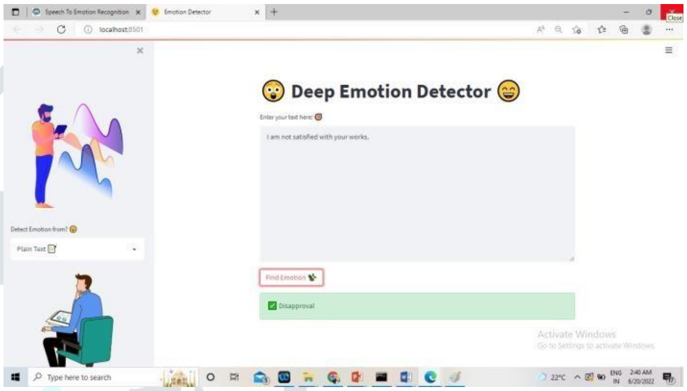
I feel frustrated and annoyed whole day due to lot of stress .

Detect Emotion from?

Plain Text

Anger

FIG: EMOTION ANGRY


Deep Emotion Detector

Enter your text here:

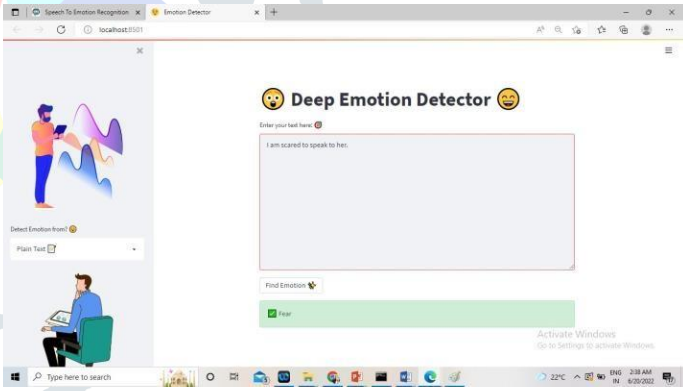
I am not satisfied with your works.

Detect Emotion from?

Plain Text

Disapproval

FIG:EMOTION DISSAPROVAL


Deep Emotion Detector

Enter your text here:

I am scared to speak to her.

Detect Emotion from?

Plain Text

Fear

FIG: EMOTION FEAR

Applications

- Lie Detectors
- Criminal Activities
- Investigations
- Nimhans

Ⅳ. CONCLUSION

In this paper we propose a novel multimodel dual recurrent encoded model that simultaneously utlilizes the text data , as well as the audio signals , to permit the better understanding of a machine . Our model detects the information from audio and text sequences using CNN and then combines the information from these sources using a feed forward neural model to predict the emotion class. Extensive experiments show that our proposed model outperforms other state of art methods in classifying the three emotion categories (angry, sad and happy).In particular ,it resolves the issue in which predictions frequently incorrectly yield the neutral class ,as occurs in previous models that focus on audio features.

Future Enhancement : The future work , we aim to extend the modalities to video inputs . Furthermore we plan to investigate the application of the attention mechanism to the data derived from multiple modalities. This approach seems likely to uncover enhanced learning schemes that will increase performance in both speech emotion recognition and other multi model classification tasks.

Ⅴ. REFERENCES

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” inAdvances in neural information processing systems, 2012, pp. 1097–1105.
2. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translationby jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.
3. Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai,Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang

Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in International Conference on Machine Learning, 2016, pp. 173–182.

4. Alex Graves, Santiago Fernandez, Faustino Gomez, and ´ Jurgen Schmidhuber, "Connectionist temporal classifi- ¨ cation: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 369–376.
5. Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, "Emotional chatting machine: Emotional conversationgeneration with internal and externalmemory," 2018.
6. Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri, "Automatic dialogue generation with expressed emotions," in Proceedings of the 2018
7. Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, 2018, vol. 2, pp 49–54.
8. Carlos Busso, Murtaza Bulut, and Shrikanth Narayanan, "Toward effective automatic recognition systems of emotionin speech," Social Emotions in Nature andArtifact, p. 110,2014.
9. Dong Yu and Li Deng, AUTOMATICSPEECHRECOGNITION., Springer,2016.
10. Google, "Cloud speech-to-text," http://cloud.google.com/speech-to-text/, 2018
11. Microsoft, "Microsoft speech api," http://docs.microsoft.com/en-us/azure/cognitiveservices/speech/home, 2018.907, 2018.
12. B. W. a. T. G. Lingli Yu, "A hierarchical support vector machine based on feature-driven method for speech emotion recognition," Artificial Immune Systems - ICARIS, pp. 901-907,2018.
13. http://pascal.kgw.tu-berlin.de/emodb/index-1280.html