Speech Emotion Recognition using Deep learning and Convolutional Recurrent Neural Network

Komal Melavanki
ECE
02FE21BEC042

Praveen Magadum
ECE
02FE21BEC064

Laxmi Kammar
ECE
02FE21BEC044

Chidambar Patil
ECE
02FE21BEC026

Guide : Prof. D.A.Torse

January 8, 2024

1 Abstract

Emotion recognition from speech signals is a challenging yet crucial task with diverse applications in human-computer interaction, sentiment analysis, and mental health monitoring. This project presents a novel approach to Speech Emotion Recognition (SER) by leveraging the capabilities of Deep Learning and a hybrid Convolutional Recurrent Neural Network (CRNN). The proposed system employs a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture both local and temporal features inherent in speech signals. The CNN component extracts relevant spectral and temporal patterns from spectrograms, while the RNN component captures long-term dependencies in the sequential nature of speech.The dataset used for training and evaluation encompasses a diverse set of emotional expressions, enabling the model to generalize well across various emotional states. The deep neural network architecture is fine-tuned through rigorous experimentation to optimize its performance in recognizing emotions accurately.

Keywords:
Convulational Neural network, Deep learning , Recurrent neural network, MPL classifiers.

2 Introduction

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion play in shaping human social interaction. Emotional displays convey considerable information about the mental state of an individual.The problem of speech emotion recognition can be solved by analysing one or more of these features. Choosing to follow the lexical features would require a transcript of the speech which would further require an additional step of text extraction from speech if one wants to predict emotions from real-time audio. Similarly, going forward with analysing visual features would require the excess to the video of the conversations which might not be feasible in every case while the analysis on the acoustic features can be done in real-time while the conversation is taking place as we'd just needthe audio data for accomplishing our task. Hence, we choose to analyse the acoustic features in this work.

Furthermore, the representation of emotions can be done in two ways:

1. **Discrete Classification**:
   Classifying emotions in discrete labels like anger, happiness, boredom etc.
2. **Dimensional Representation**: Representing emotions with dimensions such as Valence (on a negative to positive scale), Activation or Energy (on a low to high scale) and Dominance (on an active to passive scale) Both these approacheshave their pros and cons.

The dimensional approach is more elaborate and gives more context to predic tion but it is harder to implement and there is a lack of annotated audio data in a dimensional format. The discrete classification is more straightforward and easier to implement but it lacks the context of the prediction that dimensional representation provides. We have used the discrete classification approach in the current study for lack of dimensionally annotated data in the public domain.

3 Emotion and classification

This section is concerned with defining the term emotion, presenting its differentmodels. Also for recognizing emotions, there are several techniques and inputs that can be used. A brief description of all of the techniques is presented here.

1. **Definition :**

   A definition is both important and difficult because the everyday word "emotion" is a notoriously fluid term in meaning. Emotion is one of the most difficult concepts to define in psychology. In fact, there are differ- ent definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure.

Scientific discourse has drifted to other meanings and there is no consensus on a definition. Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes.

2. Categorization of emotions :

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous).

In the first approach, emotionsare described with a discrete number of classes. Many theorists have conducted studies to determine which emotions are basic. A most popular example is Ekman who proposed a list of six basic emotions , which are anger, disgust, fear, happiness, sadness, and surprise.

.

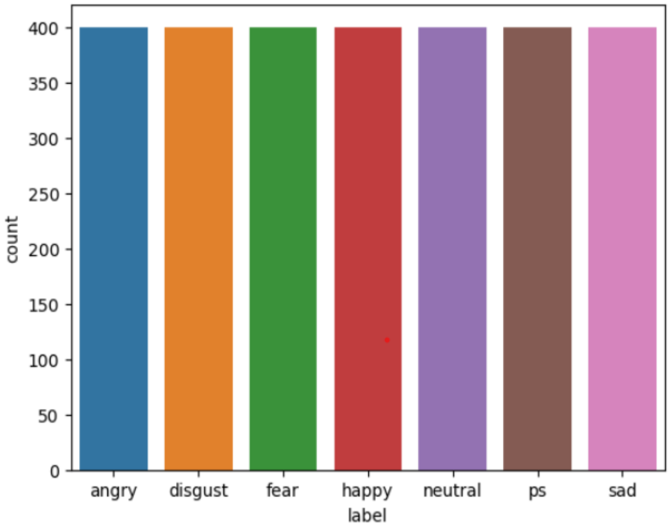


Figure 1 : Distributions of emotion with respect to gender

The categorical approach is commonly used in SER [30]. It characterizes emotions used in everyday emotion words such as joy and anger. In this work, a set of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) plus neutral, correspond- ing to the six emotions of Ekman’s model, were used for the recognitionof emotion from speech using the categorical approach.

4 Speech emotion recognition (SER) system

Block Diagram Our SER system consists of four main steps. First is the voice sample collection. The second features vector that is formed by extracting the features. As the next step, we tried to determine which features are most relevant to differentiate each emotion. These features are introduced to machine learning classifier for recognition. This process is described in Figure 1.

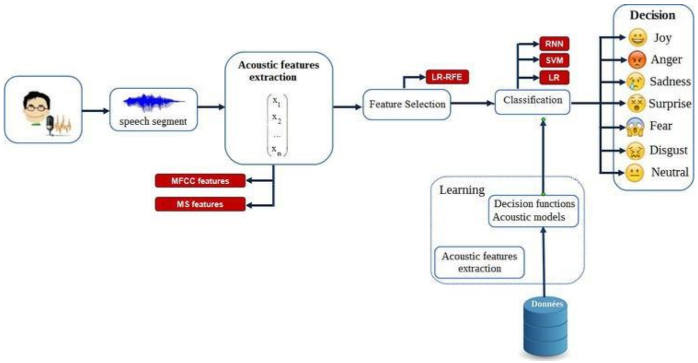


Figure 2: Block diagram of proposed system

5 Data Sources

The data used in this project was combined from five different data sources as mentioned below:

1. TESS (Toronto Emotional Speech Set): 2 female speakers (young and old), 2800 audio files, random words were spoken in 7 different emotions.

2. SAVEE (Surrey Audio-Visual Expressed Emotion): 4 male speakers, 480 audio files, same sentences were spoken in 7 different emotions.

2. RAVDESS: 2452 audio files, with 12 male speakers and 12 Female speak- ers, the lexical features (vocabulary) of the utterances are kept constantby speaking only 2 statements of equal lengths in 8 different emotions by all speakers.

6 Proposed Block Diagram

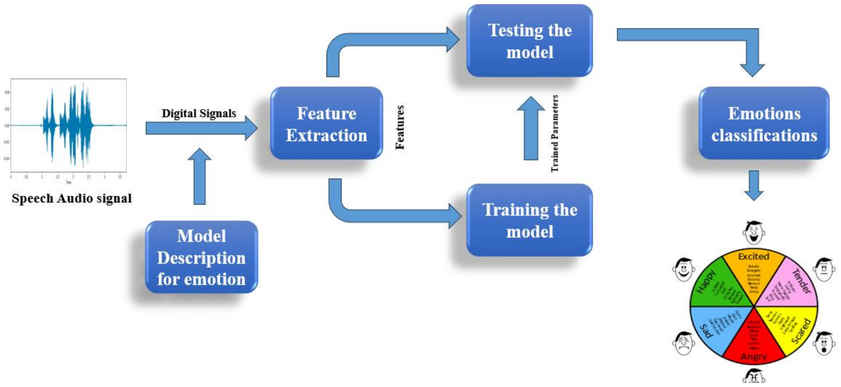
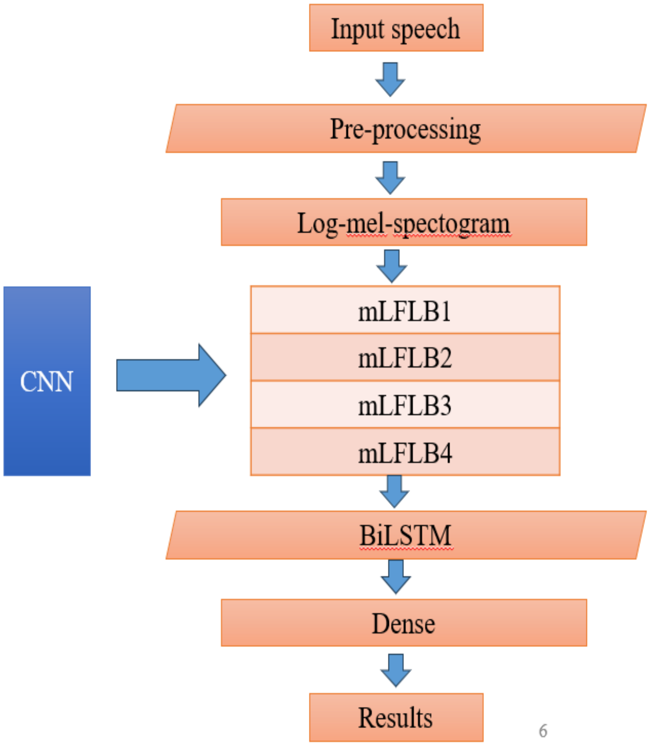


Figure 3 : Block diagram

7 Methodology



8 Features Used

From the Audio data we have extracted three key features which have been used in this study, namely,

- MFCC was by far the most researched about and utilized features in research papers and open source projects.

- Mel spectrogram plots amplitude on frequency vs time graph on a “Mel” scale. As the project is on emotion recognition, a purely subjective item, we found it better to plot the amplitude on Mel scale as Mel scale changesthe recorded frequency to “perceived frequency”.

- Researchers have also used Chroma in their projects as per literatures, thus we also tried basic modeling with only MFCC and Mel and with all MFCC, Mel, Chroma. The model with all of the features gave slightlybetter results, hence we chose to keep all three features

9 MFCC (Mel Frequency Cepstral Coefficients)

In the conventional analysis of time signals, any periodic component up as sharp peaks in the corresponding frequency spectrum

Any feature is obtained by applying Fourier Transform on a spectrogram. The special characteristic of MFCC is that it is taken ona Mel scale which is a scale that relates the perceived frequency of a tone tothe actual measured frequency.

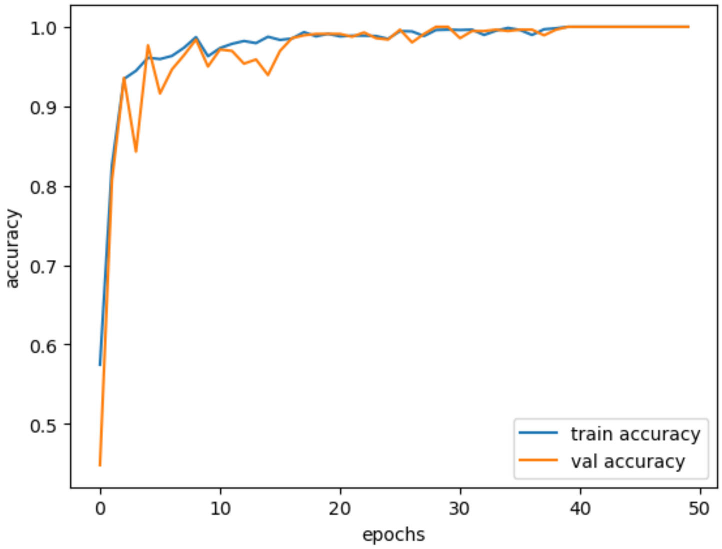
- It scales the frequency in order to match more closely what the human ear can hear. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.

- Mel Spectrogram: A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectro- gram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.

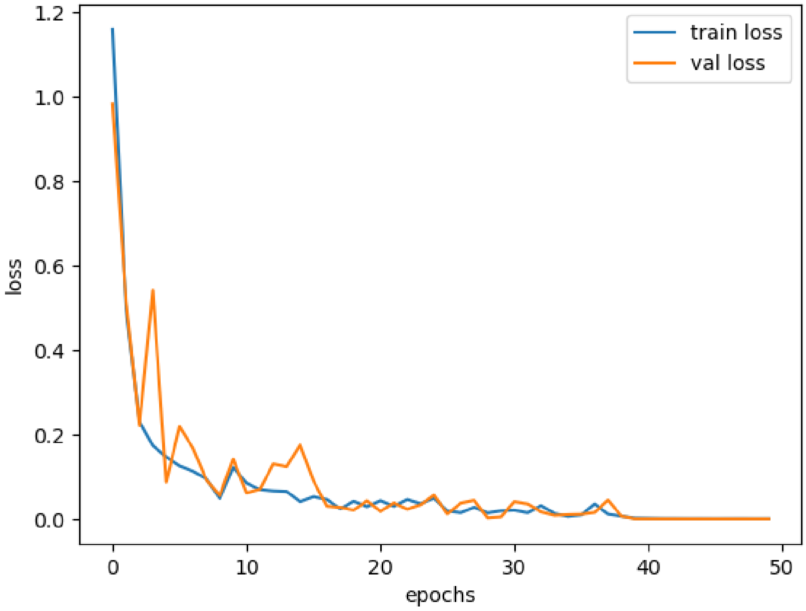
- Chroma: A Chroma vector is typically a 12-element feature vector indi- cating how much energy of each pitch class is present in the signal in a standard chromatic scale.

10 Results and Analysis

1) Train Accuracy and validation accuracy



2) Train loss and Validation Loss



11 Conclusion

In this current study, we presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions. Thus, two types of features (MFCC and MS) were extracted from two different acted databases (Berlin and Spanish databases), and a combination of these features was presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from these features. SER reported the best recognition rate of 94on the Spanish database using RNN classifier without speaker normalization (SN) and with feature selection (FS). For Berlin database, all of the classifiers achieve an accuracy of 83 percentile when a speaker normalization (SN) and a feature selection (FS) are applied to the features. From this result, we can see that RNN often perform better with more data and it suffers from the problem of very long training times. Therefore, we concluded that the SVM and MLR models have a good potential for practical usage for limited data in comparison with RNN .

Enhancement of the robustness of emotion recognition system is still possible by combining databases and by fusion of classifiers. The effect of training multiple emotion detectors can be investigated by fusing these into a single detection system. We aim also to use other feature selection methods because the quality of the feature selection affects the emotion recognition rate: a good emotion feature selection method can select features reflecting emotion state quickly. The overall aim of our work is to develop a system that will be used in a pedagogical interaction in classrooms, in order to help the teacher to orchestrate his class. For achieving this goal, we aim to test the system proposed in this work.

References :

[1] Berlin Database of Emotional Speech. http://www.elra.info/en/catalogues/catalogue-language-resources/

[2] https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/

[3] https://github.com/marcogdepinto/emotion-classification-from-audio-files?fbclid=IwAR2T4hhtWWfKdU4FwLS8LOAnF5sBwnmfc6PQHTGidzL aLl1uUVOvicx7TVw