International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Speech Emotion Recognition Using Deep Learning Techniques

Mr. Ajay Kumar Bansal[1], Ms. Shivangi[2], Asgar Aizaz Peerzada[3]

Author Affiliations

Dept. of Computer Applications, Lovely Professional University, Phagwara, Punjab, India

Abstract: With the increase in the enhancement of materialistic things in the world has turned the way of living of people extremely. Due to which there is tremendous increase in stress in day-to-day life. Such immense effect on the emotional state of people, has led to increase the importance of sentimental analysis. If emotional state of the person is not neutral, then one is unable to make any sort of decision related to either of profession or life. Emotional state of a person has direct relation with respect to any sort of action a person is going take in his living. So, this creates a sentimental analysis an essential activity that needed to be performed, whether it comes to the context of student siting in the class in order learn something new or any employee who is sitting in the office and needs to take important business decision in profession. In context with this Speech Emotion Recognition (SER) has made things easier to first analyze the sentimental state of a person and then try to turn one into neutral state to make one's living efficient. The proposed model is first going to check the current emotional state with the help of speech which carries various information like tone, pitch and speaking rate. Deep Learning technique named as LSTM (Long-Term Short-Term Memory) which is a part of RNN (Recurrent Neural Network) that is able to handle long-term dependencies in sequential data and application of sentimental analysis and speech emotion recognition, which was found out be most effective in such analysis through this paper.

Keywords: Speech emotion Recognition (SER), LSTM, RNN

Introduction

Emotions play a vital role in detecting the current mind state of a person. The only way to make communication better is the way how a person is expressing oneself and that is the only possible through sign language i.e., sound signals. Neutral state of mind helps a person to take intelligent decisions, give feedback and can learn anything new. So, emotions play an important role in classroom for a learner to acquire something new. This increases the need for SER (Speech Emotion Recognition) an important aspect now-a-day. SER could efficiently be carried by sound signals rather than any other biological signals. Three key issues need to be addressed for successful SER system, namely (1) choice of a good emotional speech database (2) extracting effective features (3) designing reliable classifiers using Deep learning algorithms. So, in the proposed model we are going to do sentimental analysis with the help of Deep learning algorithms that are LSTM with RNN, which is a type of recurrent neural network architecture that is used to handle the problem of vanishing gradients in traditional RNN.

Related Work

Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar did the review on Speech emotion Recognition using Deep Learning Techniques in which they discussed various classifiers such as K-Nearest Neighbor (KNN), Principal Component Analysis (PCA), MFCC and emotional trees for emotion recognition. Comparative study between the traditional and latest techniques for speech emotion recognition has been done out of which, they put to conclusion that DBM , DBN , RNN, CNN and AE have been much in the subject of research in recent years.

Apoorva Ganapathy , proposed with model of speech emotion recognition using deep learning methods , in which she did assessment of various approaches for sentimental analysis out of which combination of LSTM and CNN was found out to be dominant feature used for the proposed model.

R. Anusha, P. Subhashini, Darelli Jyothi, Potturi Harshitha, Janumpally Sushma, Namsamgari Mukesh in 2021, proposed with the same model that is speech emotion recognition but with different classifiers such as support vector machine, Recurrent neural network, K-nearest neighbor, Hidden Markov Model with the scope of obtaining relation between emotional state of a person and accordingly how one's behaved. They used RAVDESS dataset as a data for the system. The dataset was divided into two parts as 80% of training dataset and 20% of testing data. The model was created by training data inputs to the classifier and getting the output with 80% of the accuracy.

Kotikalapudi Vamsi Krishna, Navuluri Sainath ,A. Mary Posonia in 2022, proposed the model of Speech Emotion recognition with scope of detecting emotional state of a person on the basis on high frequency pitch and low frequency pitch using various machine learning classifiers as Support Vector Machine (SVM) , Multi-layer perception and audio feature (MFCC) , MEL, Chroma, Tonnetz were used . The model was trained to recognize the following emotions (sad, calm, neutral, surprised, angry, fearful, disgust), which resulted into 86.5% of the accuracy and testing it with the input audio they got the same.

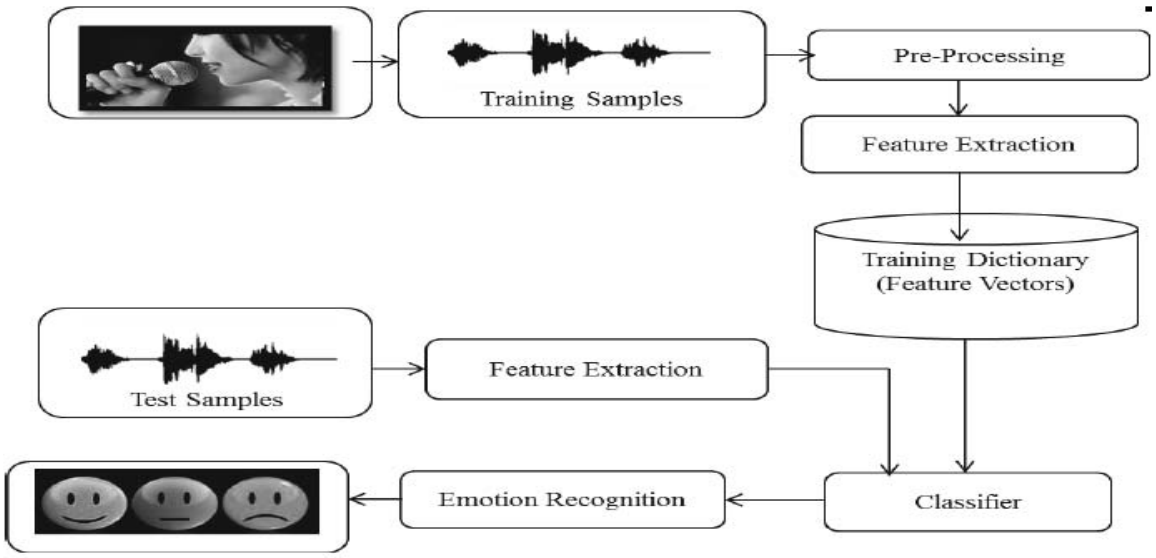
Comparative Analysis of Related Work

Author	Year	Techniques	Dataset	Remarks
1. Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar	2019	DBN, DBM, CNN, RNN, MFCC, AE	Berlin Database, Mandarin Database	Comparative study between various traditional and latest techniques for finding best deep learning algorithm
Apoorva Ganapathy	2016	LSTM, CNN, SVM, HMM	EMO-DB	Comparative study among various techniques and LSTM found to be optimum with combination of CNN
R. Anusha, P. Subhashini, Darelli Jyothi	2021	SVM, K-nearest neighbour. Hidden Markov Model	Kaggle	Aim with sentimental analysis and obtained results of about 80% of the accuracy
Potturi Harshitha, Janumpally Sushma, Namsamgari Mukesh	2021	SVM, RNN (Recurrent Neural Network)	Kaggle	Aim in obtaining connection with emotional state of a person with its behaviour
Kotikalapudi Vamsi Krishna , Navuluri Sainath , A. Mary Posonia	2022	SVM, MFCC, MEL,Chroma,Tonnetz	Kaggle	Detecting the emotional state of a person with respect to high and low frequency of pitch

Proposed Work

Speech is an important modality for sentimental analysis, as it acts as relevant communicational channel enriched with emotions. The voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker . Speech Emotion Recognition is abbreviated as SER, as it creates a natural Human Computer Interaction. This proposed model is using deep learning algorithms such as LSTM with the combination of RNN strategies. The initial step include is to accumulate data, which is really important and this methodology resolves various issues related to data quality. The next includes to distribute the dataset into two sets for training and testing set. The third step involve to train the model and extract the output regarding various emotions such as sad, happy, disgust, fear, angry . Toronto -Emotional -speech -set -tess is used as dataset for the proposed model .

System Architecture



the above figure explains the process of emotion detection using deep learning

Methodology

Google Co-lab

Google Colab, also known as Google Colaboratory, is a cloud-based platform provided by google that allows users to write, run, and share Python code online. It offers a Jupyter notebook interface that enables interactive coding, data analysis and machine learning tasks. Freely accessible is the main feature with attached pre-installed libraries and wide range of applications, made this Colab to be used in easier way.

Dataset

Toronto emotional speech set (TESS)

Kaggle Machine Learning Repository for Toronto emotional speech set (TESS). There are a set of 200 target words were spoken in the carrier phrase "say the word by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant, surprise, sadness, and neutral). There are 2800 data points (audio files) in total.

The dataset is organized such that each of the two female actor and their emotions are contain within its own folder and within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

Following is the process to extract features and training model

Step -1: Importing the required libraries

import the modules

In [2]:
```
import pandas as pd
import numpy as np
import os
import seaborn as sns
import matplotlib.pyplot as plt
import librosa
import librosa.display
from IPython.display import Audio
import warnings
warnings.filterwarnings('ignore')
```

Step-2

Load the dataset

load the dataset

In [3]:
```
paths = []
labels = []
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        paths.append(os.path.join(dirname, filename))
        label = filename.split('_')[-1]
        label = label.split('.')[0]
        labels.append(label.lower())
print('Dataset is Loaded')
```

```
Dataset is Loaded
```

Step-3

Training the model

```
from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout

model = Sequential([
    LSTM(256, return_sequences=False, input_shape=(40,1)),
    Dropout(0.2),
    Dense(128, activation='relu'),
    Dropout(0.2),
    Dense(64, activation='relu'),
    Dropout(0.2),
    Dense(7, activation='softmax')
])

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy
model.summary()
```

Step-4

Extracting features

In [16]:

```
emotion = 'fear'
path = np.array(df['speech'][df['label']==emotion])[0]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

In [17]:
```
emotion = 'angry'
path = np.array(df['speech'][df['label']==emotion])[1]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

[19]:
```
emotion = 'disgust'
path = np.array(df['speech'][df['label']==emotion])[0]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

[20]:
```
emotion = 'neutral'
path = np.array(df['speech'][df['label']==emotion])[0]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

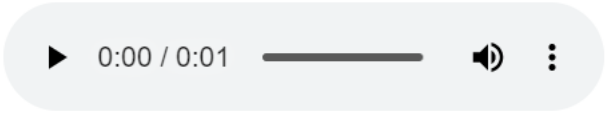
[21]:
```
emotion = 'sad'
path = np.array(df['speech'][df['label']==emotion])[0]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

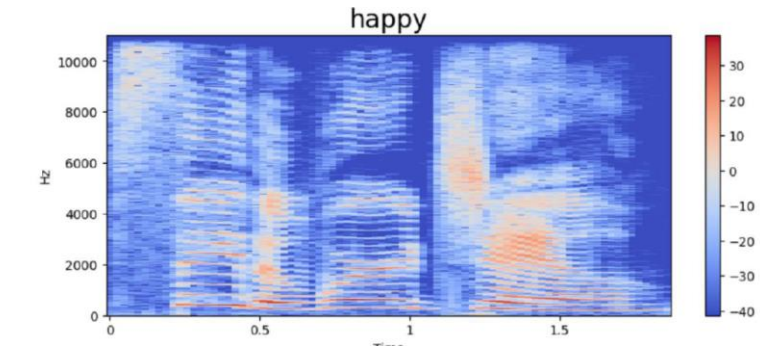
[22]:
```
emotion = 'happy'
path = np.array(df['speech'][df['label']==emotion])[0]
data, sampling_rate = librosa.load(path)
waveplot(data, sampling_rate, emotion)
spectogram(data, sampling_rate, emotion)
Audio(path)
```

Step-4 testing samples

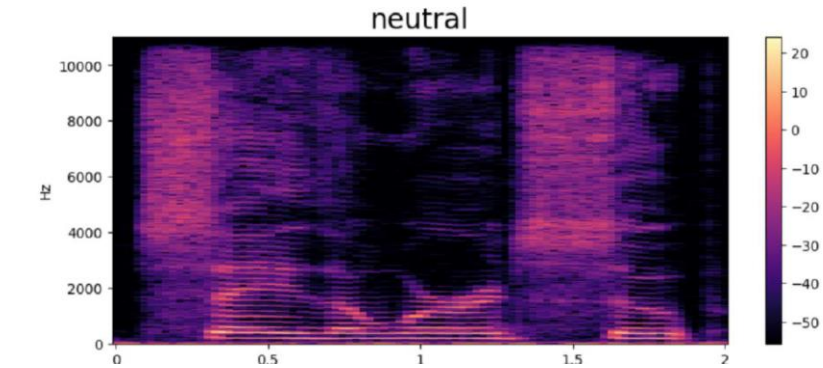
Out[16]:

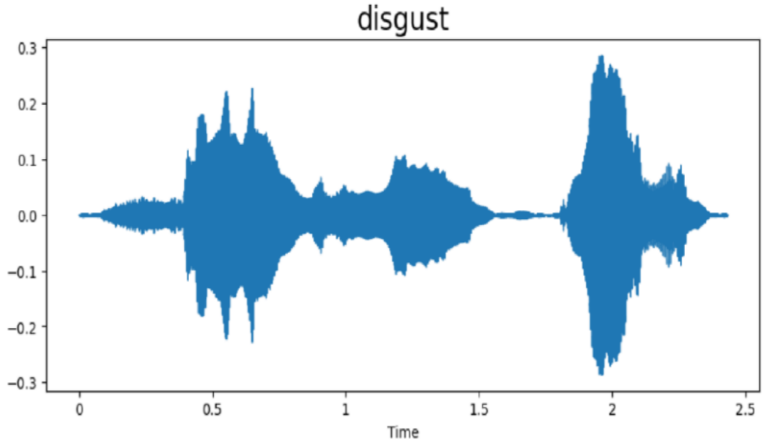
Out[17]:

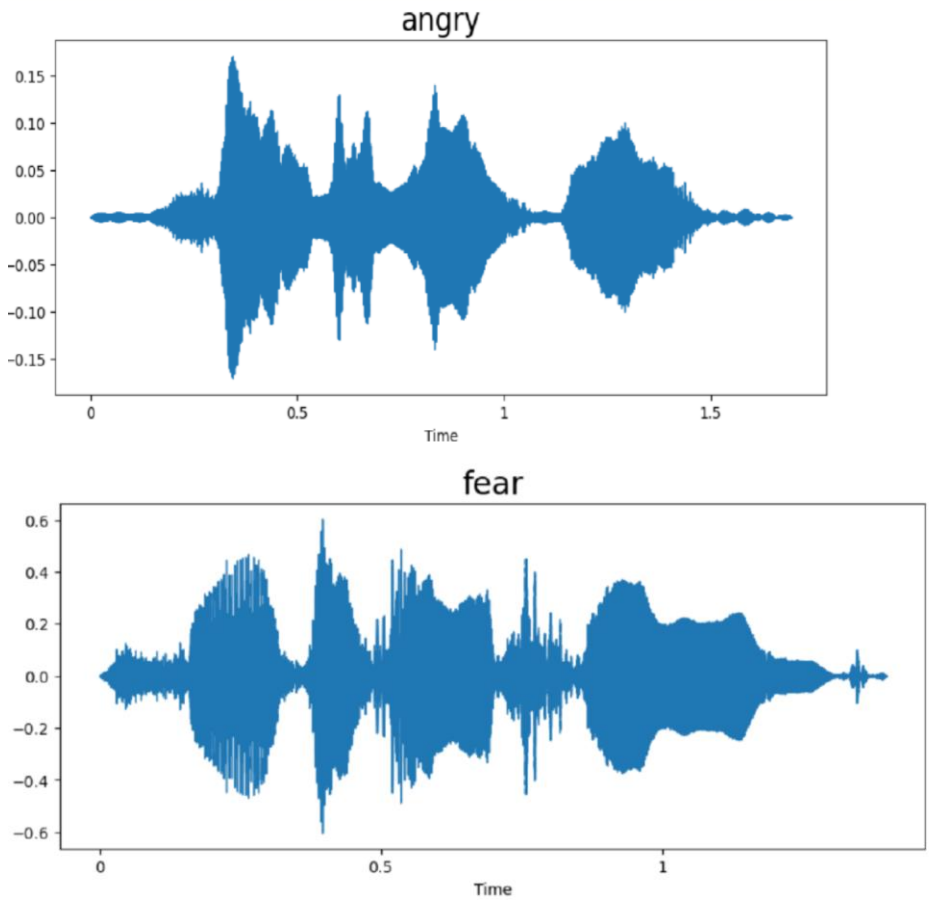



happy
Hz
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
30
20
10
0
−10
−20
−30
−40

]: ▶ 0:00 / 0:02

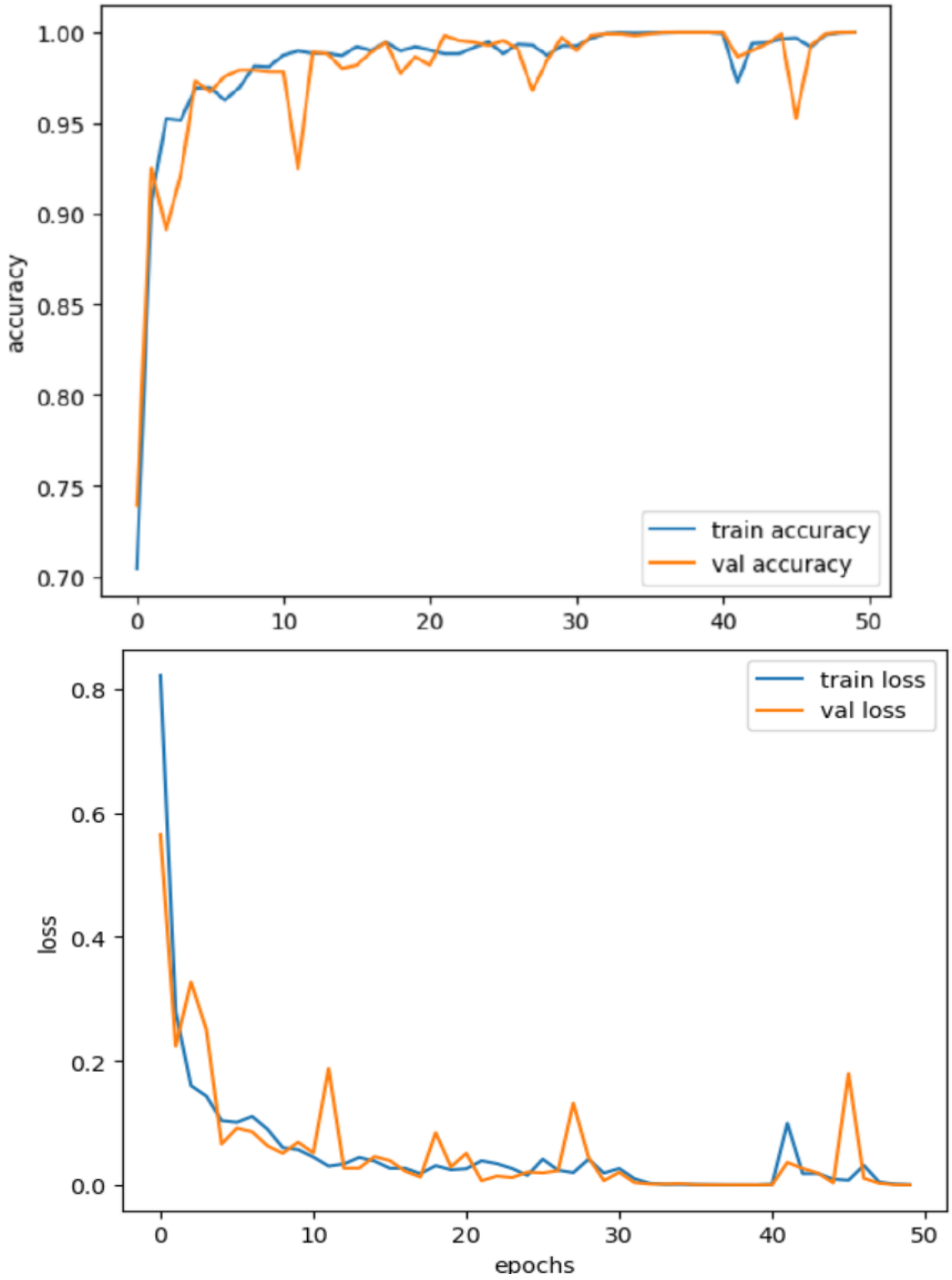

neutral
Hz
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
20
10
0
−10
−20
−30
−40
−50


disgust
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
0.5
1
1.5
2
2.5
Time



Conclusion and Results

The given plot is obtained while checking the accuracy during training the model, to check whether the LSTM is overfitting or underfitting. As, the model starts with increase and then continues to plateaus suggest that it is overfitting



The project followed several steps, including loading the dataset, visualizing the audio signals, extracting MFCC features, splitting the data into training and validation sets and building a deep learning model using LSTM layers. The model was trained for 50 epochs, and the training and validation loss and accuracy were plotted using graphs

By obtaining 72% accuracy as a result , we found that LSTM was appropriate model to do SER(Speech emotion recognition) analysis and effective enough to cope up with the real life problem , which is applicable in educational institutions as well as in any professional environment where emotional state of a person plays significant role in learning something new and understanding the surroundings

References

1. Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariq Ullah Jan, Mohammad Haseeb Zafar and Thamer Alhussain “Speech Emotion Recognition Using Deep learning Techniques: A Review”,IEEE Access Volume 7,2019
2. R.Anusha, P.Subhashini, Darelli Jyothi, Potturi Harshitha , Janumpally Sushma, Namsamgari Mukesh, “Speech Emotion Recognition using Machine Learning “ (ICOEI) IEEE Xplore Part Number :CFP21J32-Art
3. Kotikalapudi Vamsi Krishna , Navuluri Sainath , A. Mary Posonia “Speech Emotion recognition using Machine Learning “ 6th ICCCMC 2022 , IEEE Xplore Part Number:CFP22K25-ART
4. Li,X.,Li,X.,Li,L.& Li,M.(2020).Speech Emotion Recognition using deep learning : A Review.IEEE Access , 8 129925-129940
5. Abdulla,W.H , & Darweesh , F.(2020) , Speech Emotion Recognition Using Deep Learning Algorithms :A Review . International Journal of Computer Applications, 179(37)