Zephyrus

DSN2099

Team

Shubham Tejani
20BAI10152

Parv Paliwal
20BAI10228

Bhavyangana Kanthed
20BAI10380

Nayan Kumar
20BAI10386

M D Shah Fahad
Guide

Dr S Sountharrajan
Programme Chair

Introduction

Emotions can be considered as the first natural communication strategy and can be really complex to understand.

Speech Emotion Recognition (SER) is the task of recognizing emotions from speech signals while a person talks. This is very important in advancing human–computer interaction.

Existing Work and Limitations

1

01. In 2018, Norman Swain reviewed studies between 2000 and 2017 on SER systems based on three perspectives: database, feature extraction, and classifiers.

02. The research has an extensive section on databases and feature extraction;

03. however, only traditional machine learning methods have been considered as classifying tool, and the authors are regretting neural networks and deep learning approaches.

2

01. In 2019 , RA Khalil reviewed discrete approaches in SER using deep learning.

02. Several deep learning approaches, including deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoder, have been mentioned along with some of their limitations and strengths in the study.

03. However,the research is not addressing the accessible approaches to overcome weaknesses.

Proposed Work

The classification model of emotion recognition here proposed is based on a deep learning strategy based on convolutional neural networks (CNN), Support Vector Machine (SVM) classifier, MLP Classifier.

MFCC

01. MFCC is a different interpretation of the Mel-frequency cepstrum (MFC)

02. The MFC coefficients have mainly been used as the consequence of their capability to represent the amplitude spectrum of the sound wave in a compact vectorial form.

Wave Reconstruction

01. The audio file is divided into frames, usually using a fixed window size, to obtain statistically stationary waves

02. The amplitude spectrum is normalized with a reduction of the "Mel" frequency scale.

MFCC sequence

01. For each audio file,40 features have been extracted.

02. The feature has been generated by converting each audio file to a floating-point time series.

03. Then, an MFCC sequence has been created from the time series.

SER Model

Helps recognize emotion through speech. For eg,

- **Fear, anger, or joy - loud and fast & high pitched speech.**
- **Sadness or tiredness generates slow and low-pitched speech.**

SER for enhancing Human Computer Interaction. For project,

- **Using Machine Learning and Deep Learning**
- **And Ravdess data set to train the model for Speech Emotion Recognition**

Methodology

The key idea is considering the MFCC commonly referred to as the “spectrum of a spectrum”, as the only feature to train the model.

Dataset

- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset
- Toronto emotional speech set (TESS) dataset

CNN

01. The deep neural network(CNN) designed for the classification task can work on vectors of 40 features for each audio file provided as input

MLP

01. A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN).

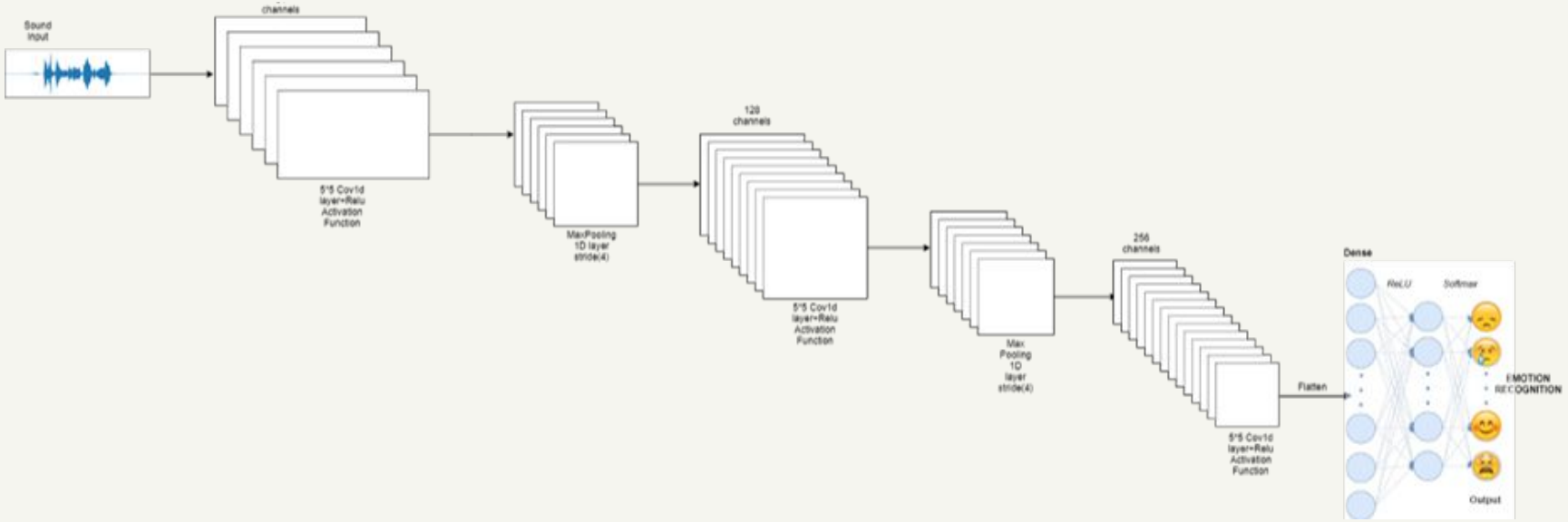
02. MLP utilizes a supervised learning technique called backpropagation for training.

SVM

01. “Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges.

Methodology

An architecture based on deep neural networks for the classification of emotions using audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto emotional speech set (TESS)



Novelty

- Approaches based on deep neural networks are an excellent basis for solving the task.
- Model is general enough to work in a real application context correctly.
- Increased accuracy - Combining the two datasets - RAVDESS & TESS.
- Removed audio collected from video files from RAVDESS dataset as it caused overfitting.
- Scope for improvisation upon more data addition.
- Model may be enhanced further based on real time data.
- May incorporate other learning models to study mixed emotion.

Real Time Usage

Enhances HC Interaction

Interactive movies,
Story telling, etc.

Quality Enhancement

Of service of calls at the call centers.

Psychological Treatments

Identifying patients under depression, reducing suicides.

Surveillance Systems

Useful in case of modern surveillance systems.

Dynamic Web Page

Update web page based on emotions recognized.

E-com & Online Market

Website responses based on speech and emotions.

Requirements

Hardware Requirements

1. PC with 250 GB or more Storage Space
2. PC with 4 GB RAM.
3. PC with i3 and Above.

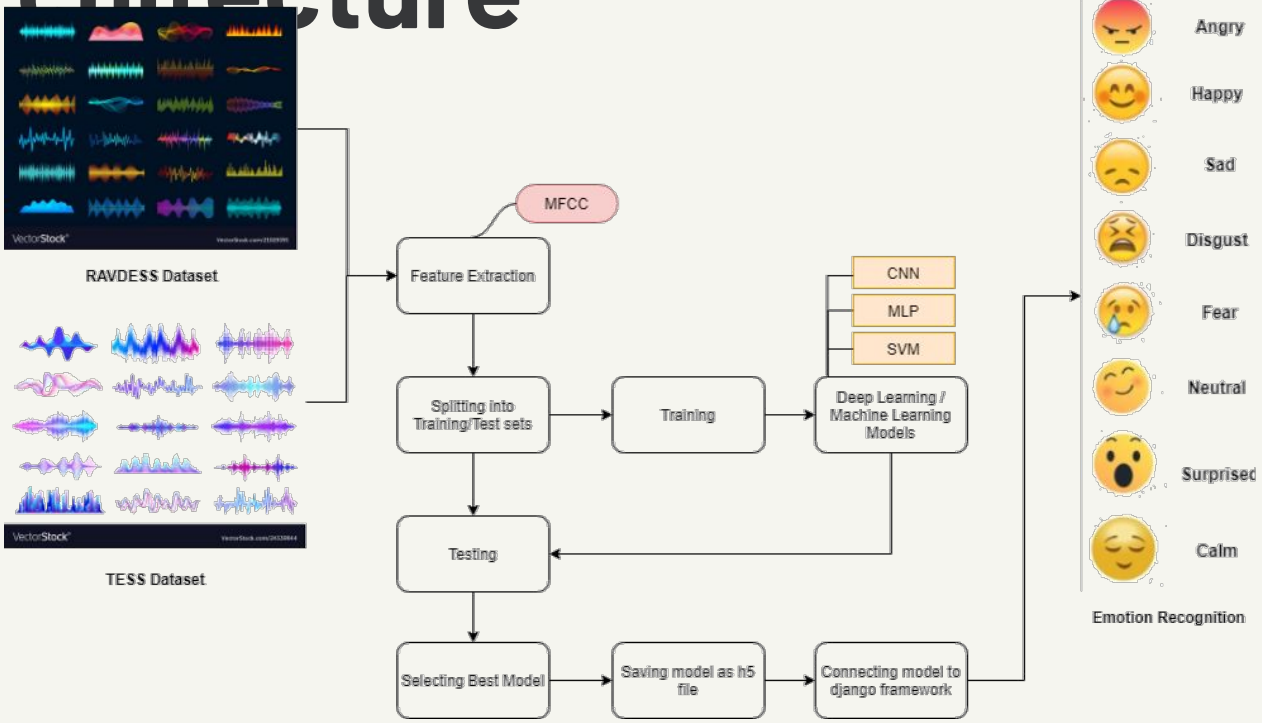
Software Requirements

1. librosa
2. soundfile
3. numpy
4. os
5. glob
6. pickle
7. sklearn
8. keras

System Requirements

1. Operating system - Windows 7 and above.
2. Language - Python
3. IDE - Google Colab
4. Browser - Google Chrome

Overall System Architecture



DATASET USED

The dataset is built using 5252 samples from:

- Ryerson Audio-Visual Database of Emotional Speech and Song **(RAVDESS)** dataset.
- Toronto emotional speech set **(TESS)** dataset.

Limitation

Lack of data

Noisy/Non-Noisy environment

Continuity of the speech

Different cultures and language

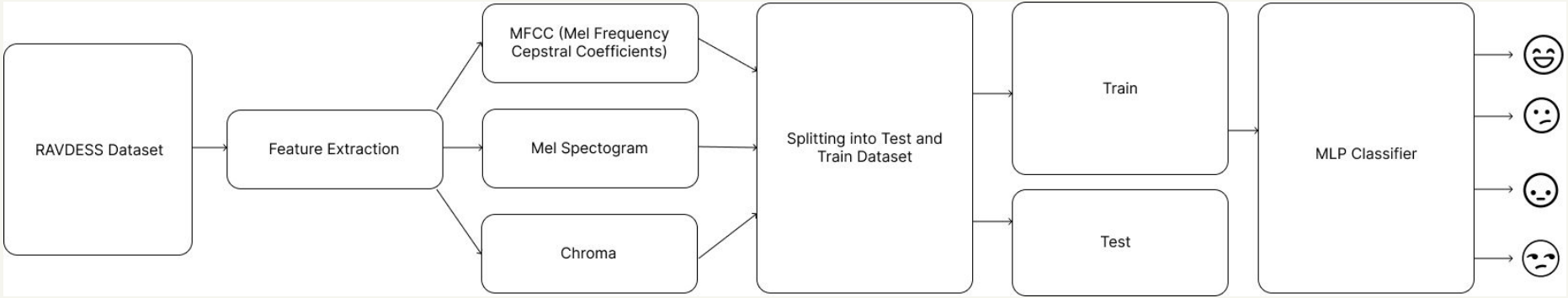
Module Description

- We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non linearly separated.

- Mel-frequency cepstrum coefficients (MFCC), chroma and mel features are extracted from the speech signals and used to train the MLP classifier.

- For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, NumPy and soundfile to analyze the speech modulations and recognize the emotion.

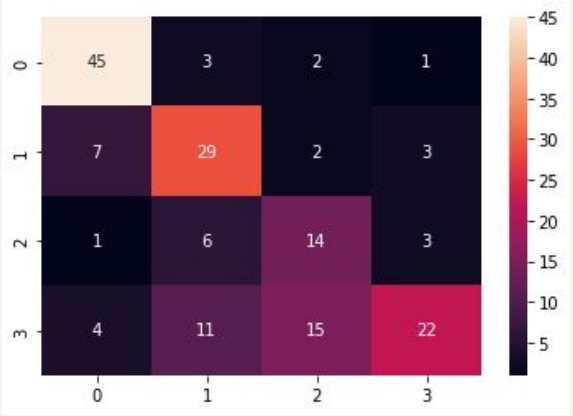
Module Workflow



Result and Discussion

- The results obtained from the evaluation phase show the effectiveness of the model compared to the baselines and the state of the art on the RAVDESS dataset.precision, recall and F1-score obtained for the emotional classes like angry, happy, neutral and sad. The accuracy of our SER system comes out to be 65% accurate.

- This project shows that MLPs are very powerful in classifying speech signals. have obtained higher accuracies as compared to other approaches for individual emotions. .The results obtained in this study demonstrate that speech recognition is feasible, and that MLPs can be used for any task concerning recognizing of speech and demonstrating the accuracy of each emotion present in the speech



	precision	recall	f1-score	support
angry	0.79	0.88	0.83	51
happy	0.59	0.71	0.64	41
neutral	0.42	0.58	0.49	24
sad	0.76	0.42	0.54	52
accuracy			0.65	168
macro avg	0.64	0.65	0.63	168
weighted avg	0.68	0.65	0.65	168

Conclusion

In this project, the features are categorized into two broad categories of acoustic and non-acoustic. Further, the acoustic features are categorized into the prosody, spectral, wavelet-based, voicequality, non-linear, and deep-learning-based features. The nonacoustic features are categorized into linguistic, discourse, face, gesture, and video. The strength of each type of acoustic feature is studied with respect to different emotions. This study helps to select the right features for SER. The scenarios are also presented where the non-acoustic features may be useful in combination with the acoustic features. The feature selection algorithms are discussed with respect to SER. The related discussion in this paper will help in selecting suitable feature selection algorithms.

The performance of a module is highly dependent on the quality of pre-processing. Mel Frequency Cepstrum Coefficients are very dependable. Every human emotion has been thoroughly studied, analyzed and the accuracy has been checked. The results obtained in this study demonstrate that speech recognition is feasible, and that MLPs can be used for any task concerning recognizing of speech and demonstrating the accuracy of each emotion present in the speech.

References

[1] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[2] RAVDESS

https://zenodo.org/record/1188976

[3] TESS
https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess

[4] LOGAN, B., ET AL. Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

Thank you!