भारतीय
प्रौद्योगिकी
संस्थान
काशी हिन्दू विश्वविद्यालय

IIT

INDIAN
INSTITUTE OF
TECHNOLOGY
BANARAS HINDU UNIVERSITY

UG PROJECT REPORT

•**PROJECT AIM:** DEVELOPMENT OF SPEECH EMOTION RECOGNITION MODEL USING CNN.

•**Submitted by:**

1. Gaurav Mahaur - 19095036
2. Saurabh Yadav - 19095087
3. Varsha Jangir - 19095104
4. Nishant K. Robin - 19095121

•**Under the guidance of:**

Dr. Satyabrata Jit
Department of Electronics Engineering

Index

-

ABSTRACT

Speech Emotion Recognition (SER) is a very interesting application of machine learning. As emotions play a vital role in communication, the detection and analysis of the same are of vital importance in today's digital world of remote communication. This report illustrates the various steps in designing and implementing SER system that processes and classifies speech signals to detect emotions embedded in them. Initially, the speech emotion signals are collected from a database such as the TESS database. After that, feature extraction is considered, and it is carried out by the Pitch and Energy, Mel-Frequency Cepstral Coefficients (MFCC). The mentioned feature extraction process is widely used for classifying the speech data and performs better in performance. The extracted features are used for the recognition purpose by the CNN network. In the proposed CNN network, either one or more pairs of convolutions, besides, max-pooling layers remain present. With the utilization of the CNN network, emotions are recognized through the input speech signal.

AIM

To develop a model that can detect different types of emotions with the help of voices. Basically, this model will be used to understand the emotional state of different states of people.

INTRODUCTION

Speech emotion recognition is the task of recognizing emotions from speech signals. Understanding one's feelings at the time of communication are constructive in comprehending the conversation and responding

appropriately. As a part of the current research area, This emotion detection model using machine learning techniques was developed. This automatic SER can be used to identify the emotions of different people.

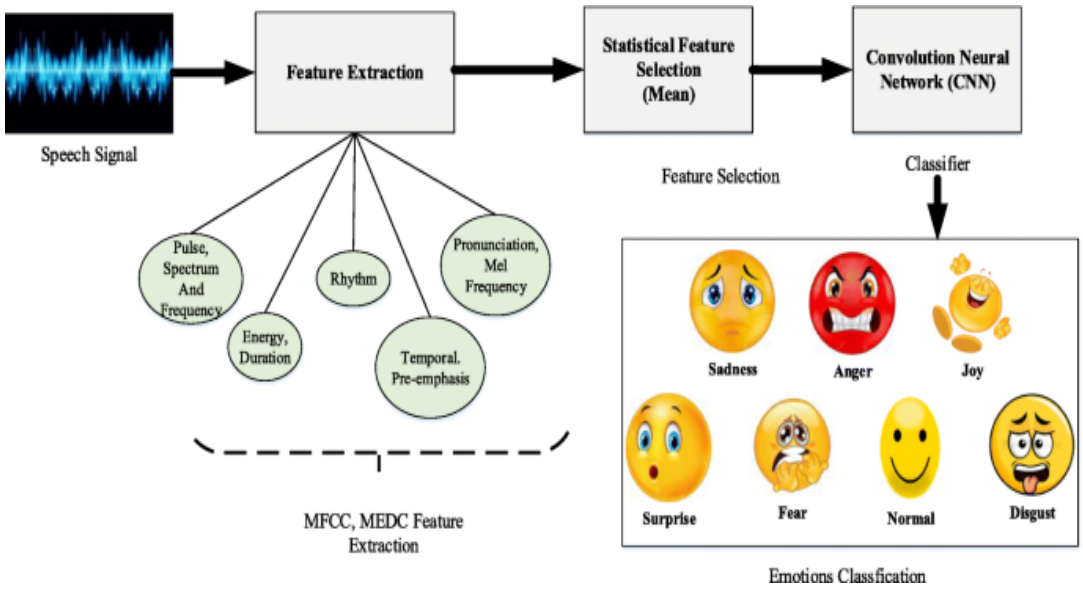
SER helps smart speakers and virtual assistants to understand their users better. The other application can help translate from one language to another, especially as other languages have different ways of projecting emotions through speech. SER is also beneficial in online interactive tutorials and courses. Understanding the student's emotional state will help the machine decide how to present the rest of the course contents. It can recognize the driver's state of mind and help prevent accidents and disasters. Another application is in therapy sessions; by employing SER, therapists will understand their patients' state and possibly underlying hidden emotions as well. The service industry and e-commerce can utilize speech emotion recognition in call centers to give early alerts to customer service and supervisors of the caller's state of mind.

These are some of the problems which can be solved by developing an automatic SER model with high accuracy.

Target Audience - students, therapists, service industry and e-commerce, etc.

METHODOLOGY

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better.



Flow of emotion classification model

A. Dataset used :

The dataset which is used in the training of the model is the ***Toronto Emotional Speech Set (TESS)*** which is one of the 4 key dataset. This dataset is an acted dataset primarily developed for analyzing the effect of age on the ability to recognize emotions. This dataset is all comprised of two female actors, about 60 and 20 years old. Each actor has simulated eight emotions for 200 neutral sentences. Emotions in this dataset are: angry, pleasantly surprised, disgusted, happy, sad, fearful, and neutral. In total 2800 audio files are comprised in this dataset.

Now the entire dataset is used to train the model on the different algorithms.

```
def load_data(test_size=0.2):
    x=[]
    y=[]
    for file in glob.glob(r"C:\Users\Saurabh\ml\btp\data\*"):
        basename = os.path.basename(file)
        label = file.split('_')[-1]
        label = label.split('.')[0]

        for f in glob.glob(file+"\*"):
            y.append(label.lower())
            x.append(get_features(f))
    return x,y
```

Function to load dataset

B. Data Pre-processing:

Data must be cleaned to perform any meaningful analysis. As a next step, the dataset thus collected had to be inspected for its quality. Some of the data quality issues addressed for this experimentation include:

1. Missing value analysis
2. Outlier identification
3. Null value handling
4. Invalid data
5. Duplicate data

C.NORMALIZATION AND STANDARDIZATION:

Different characteristics of the audio signal, represented by its features, are computed on different units or scales. Rescaling the values to a uniform range will ensure accurate calculations are made. Many algorithms use distance

metrics for their computation. Therefore it is necessary that all the values in the dataset are normalized.

Normalization alters all numeric values to lie in the range 0 to 1. For this purpose, all outliers in the data must be eliminated prior to normalizing the data. Standardization transforms the data to have a mean value of zero and a variance of one.

D. Feature Extraction:

From the Audio data, we have extracted three key features which have been used in this study, namely, MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram, and Chroma. The Python implementation of the ***Librosa*** package was used in their extraction.

In [3]:
```
def get_features(filename):
    y, sr = librosa.load(filename, duration=3, offset=0.5)
    mfcc = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T, axis=0)
    return mfcc
```

Function for extraction of features

Details about the features are mentioned below-

1.MFCC (Mel Frequency Cepstral Coefficients)- In the conventional analysis of time signals, any periodic component (for example, echoes) shows up as sharp peaks in the corresponding frequency spectrum (i.e. Fourier spectrum. Which is obtained by applying a Fourier transform on the time signal). Any cepstrum feature is obtained by applying Fourier Transform on a spectrogram. The special characteristic of MFCC is that it is taken on a Mel scale which is a scale that relates the perceived frequency of a tone to the 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.

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

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

.

CNN MODEL AND ARCHITECTURES

Convolutional neural networks (CNNs) or shift-invariant artificial neural networks (SIANNs) are particular types of neural networks that, in their hidden layer they have different filters or regions that respond to a specific feature of the input signal. CNN will operate by training and testing the modules. Input data are fed through the convolutional layer series with a link of connected layers. The classification process is done by the Softmax function with a probabilistic value between the range of 0 and 1. Since CNN is interlinked, hidden layers offer a much easier way to train and test the data. The backpropagation algorithm is used in CNN for computing the optimization of the parameters. The designed CNN has three main characteristics: location, weight distribution, and pooling. They each have the potential to improve speech recognition performance. The space in the conventional fly units allows for greater strength against non-white noise, where some straps are cleaner

than others. This is because useful features can be counted locally from cleaner parts of the spectrum, and only a limited number of features are affected by noise. This presents an excellent opportunity for the upper layers of the network to handle this noise because they can combine the calculated high-level features for each frequency band. This is clearer than dealing with input features in the lower layers like standard, fully connected neural networks. Also, the local network reduces the number of loads.

We have developed the CNN model with Keras and constructed it with 5 layers — 4 Conv1D layers followed by a Dense layer.

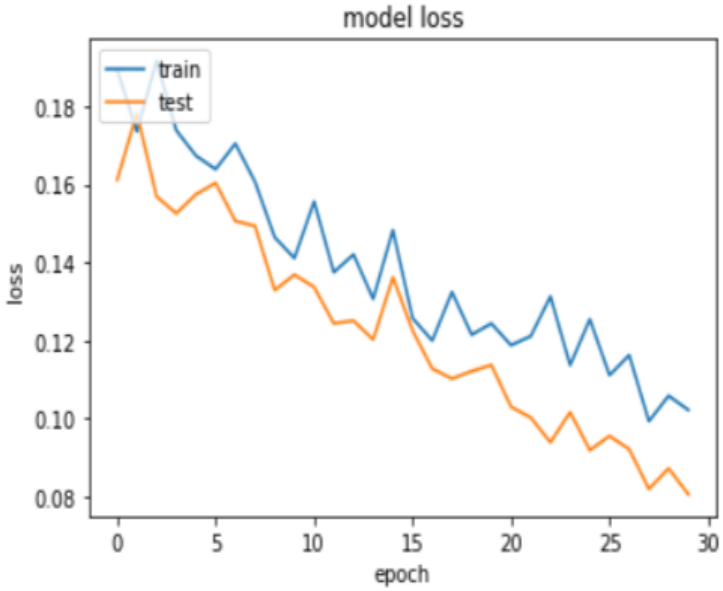
In [19]:
```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from tensorflow.keras.models import Sequential, load_model
from tensorflow.keras.regularizers import l2
from keras.layers import Conv1D, MaxPooling1D, AveragePooling1D
# from keras.utils import to_categorical
from keras.layers import Activation, Dense
from keras.layers import Input, Flatten, Dropout, Activation
model = Sequential()

model.add(Conv1D(256, 5,padding='same',
                 input_shape=(40,1)))
model.add(Dropout(0.2))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same'))
model.add(Dropout(0.2))
model.add(Activation('relu'))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(128, 5,padding='same',))
model.add(Dropout(0.2))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same',))
model.add(Dropout(0.2))
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(8))
model.add(Activation('softmax'))
opt = keras.optimizers.RMSprop(lr=0.00001, decay=1e-6)
model.summary()
```

Architecture of the CNN Model

EXPERIMENTAL RESULT

This report has been applied to different algorithms on the TESS Dataset to discover a better classification performance of the network. From the TESS dataset, here taken 2800 audio files which contain one of the 7 categories of emotion. 80 % of the audio files from these are used for training, and 20 % are used for testing the model. The network is trained for 30 epochs. The training accuracy of the proposed SER model is **96.64%** and the validation accuracy is **97.67%**.The precision of the model is **97.91%** and the recall value is **97.39%.** The comparison of the proposed model with the conventional models shows that the results of this model are good and promising to use in real-world applications.

These graph shows the relation between accuracy and loss with the number of epochs



CONCLUSION

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. In this report, the steps of building a speech emotion detection system were discussed in detail Initially, the limited number of publicly available speech databases made it challenging to implement a well-trained model. Next, several novel approaches to feature extraction had been proposed in the earlier works, and selecting the best approach included performing many experiments. At the end of the experimentation, it can be concluded that an integrated feature space will produce a better recognition rate when compared to a single feature.

FUTURE WORK

For future advancements, the proposed project can be further modeled in terms of efficiency, accuracy, and usability. Additional to the emotions, the model can be extended to recognize feelings such as depression and mood changes. Such systems can be used by therapists to monitor the mood swings of the patients. A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem than emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. A sentiment detection using vocabulary can be integrated with speech emotion detection to identify a possible sarcasm. Therefore, in the future, there would emerge many applications of speech-based emotion. recognition system.

REFERENCES

1) https://www.researchgate.net/publication/322873355_Speech_Emotion_Recognition_Methods_and_Cases_Study
2) https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess --The dataset used in this model
3) https://medium.com/@patrickbfuller/librosa-a-python-audio-library-60014eeaccfb - Librosa library article
4) https://ieeexplore.ieee.org/document/8308186 – CNN article