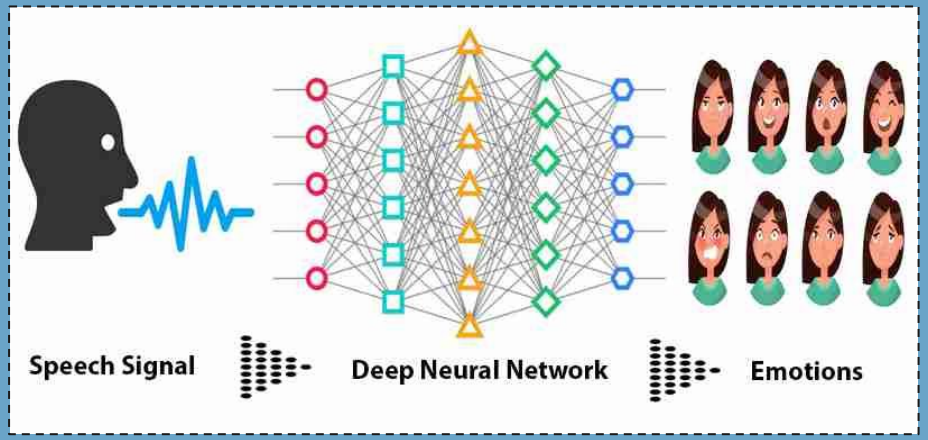
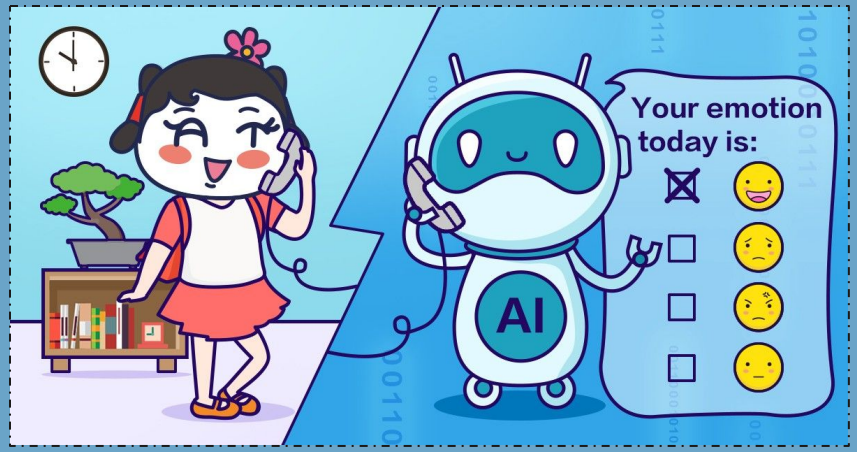
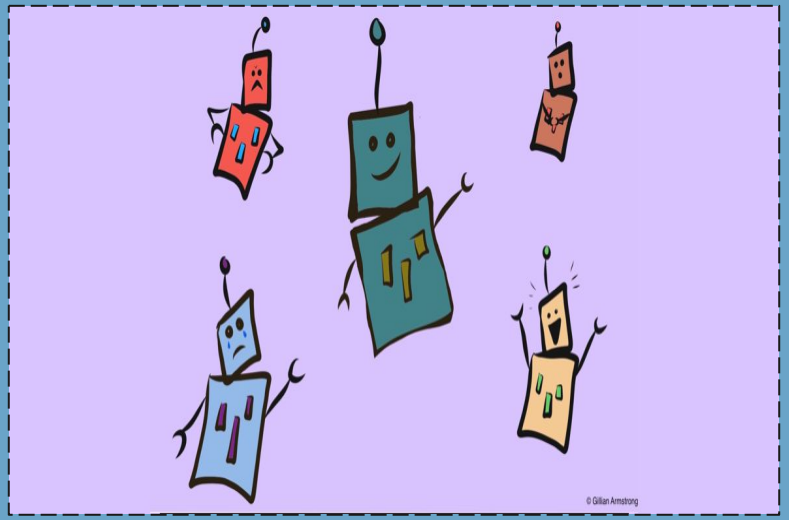
SPEECH EMOTION RECOGNITION USING DEEP LEARNING METHODS



Contents



Contents



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. This has opened up a new research field called automatic emotion recognition. Emotions can be expressed through various modalities such as speech,facial expressions and body language.

Motivation

The market for speech recognition devices is increasing day by day.

Speech and Voice Recognition Market Size to Reach USD 28,335.3 Mn by 2026, Advent of Machine-friendly Voice Recognition Format to Boost Growth, says Fortune Business Insights

Apple Siri Vs Google Assistant Which Is Better In Speech Recognition?

Hey, Alexa! Smart Classes With Amazon Alexa Are a Very Real Thing in Schools in Bastar

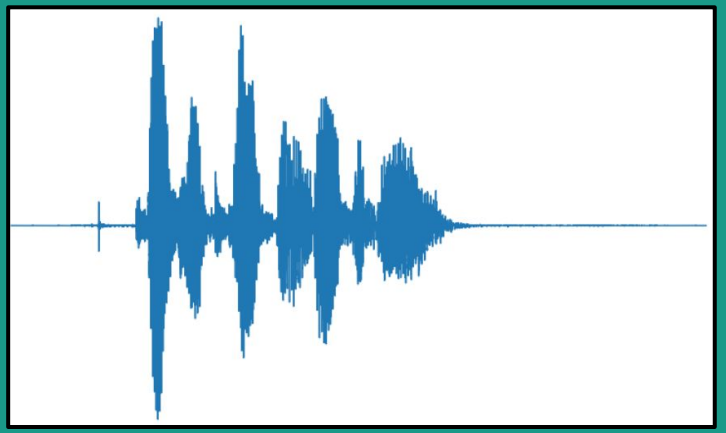
Xiaomi Integrates Voice Assistant into Child's Electric Scooter

BBC releases its own 'Beeb' voice assistant in beta

Apple and Google have trained their virtual assistants to rebut 'All lives matter'

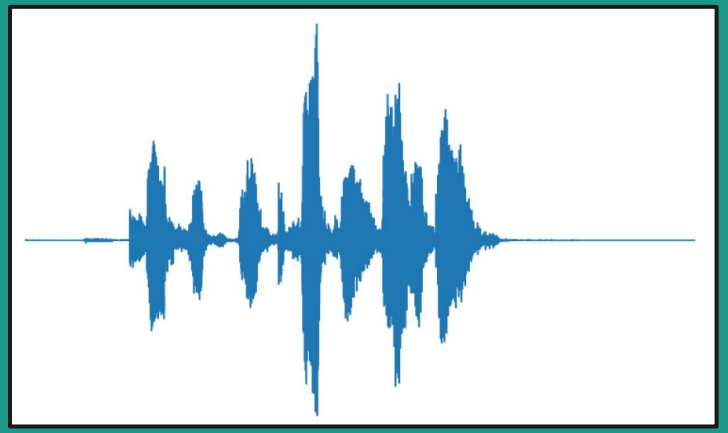
Walmart's Indian E-Commerce Subsidiary Flipkart Debuts Voice Assistant for Grocery Shopping in Hindi and English

NEUTRAL

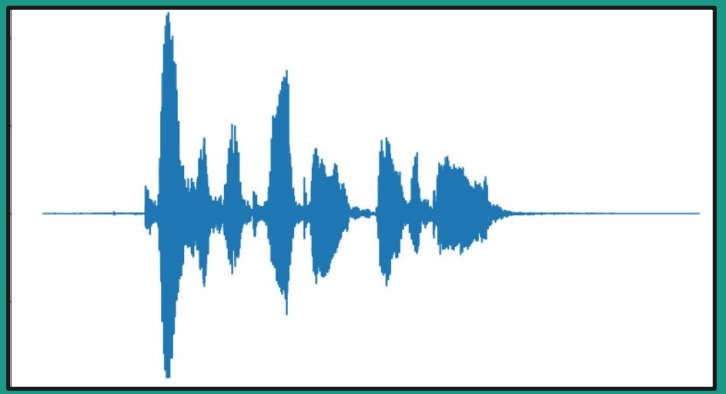


The phrase **“The kids are talking by the door”** has been spoken in 4 different emotions and we have plotted them.The change in emotion is evident in plots of the speech signal.

ANGRY

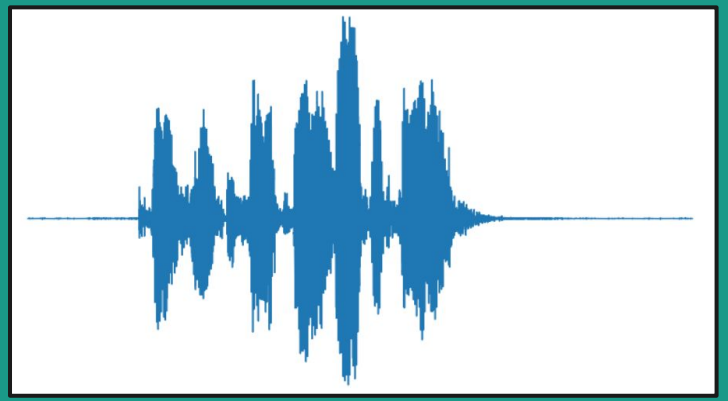


SAD

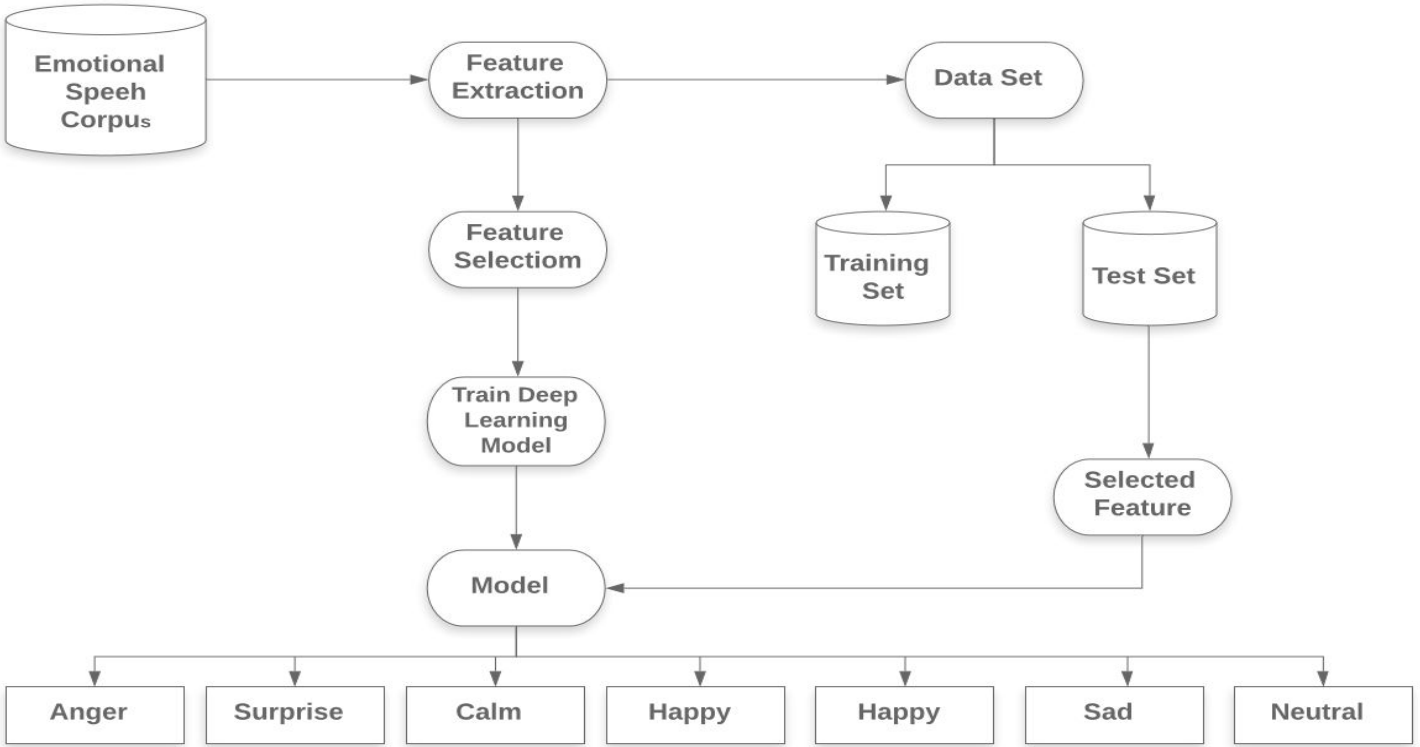


In speech,apart from text the emotional content of the speech attributes to the true meaning of the speech.When these voice assistants are armed with such information they can contribute to a more emotionally sensitive dialog.

HAPPY



Flowchart of our system



Issues addressed while designing SER systems

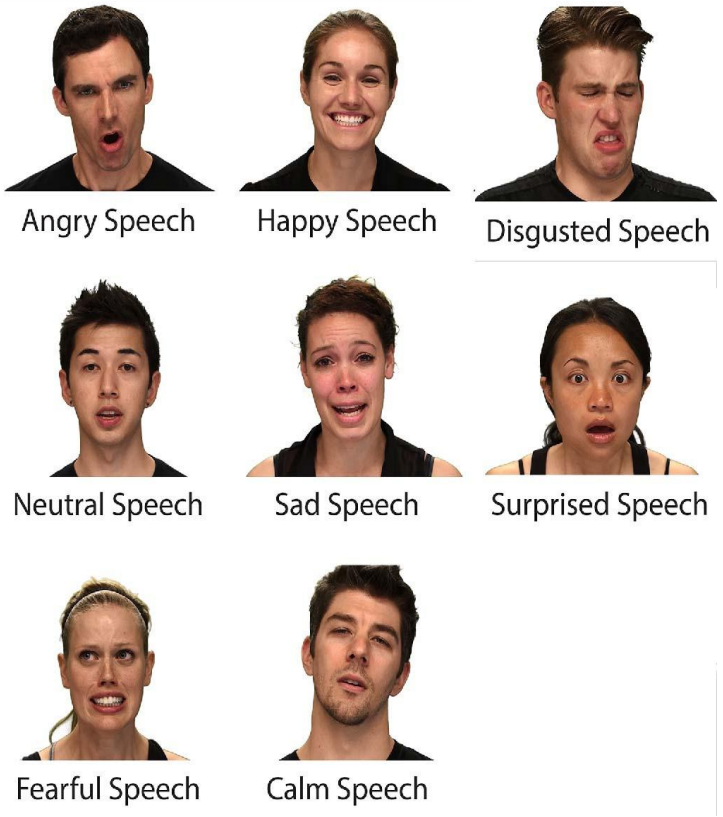
Three key issues need to be addressed for successful SER system:

1. Choice of a good emotional speech database.
2. Extracting effective features.
3. Designing reliable classifiers using machine learning algorithms.

Emotional speech database

For training speech emotion recognition system we require a speech corpus that capture the emotions in speech in a clear manner.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. This gender-balanced database “consists of 24 professional actors, each performing 104 unique vocalizations with emotions that include: happy, sad, angry, fearful, surprise, disgust, calm, and neutral” . Each actor enacted 2 statements for each emotion: “Kids are talking by the door” and “Dogs are sitting by the door.” These statements were also recorded in two different emotional intensities, normal and strong, for each emotion, except for neutral (normal only) . Actors repeated each vocalization twice . There are a total of 1440 speech utterances.



FEATURE EXTRACTION

Throughout machine learning, handcrafted features are used to simplify the task of working with complex data.The speech signal contains a large number of parameters that reflect the emotional characteristics.

- **Acoustic**

They describe the wave properties of a speech. It includes Fourier frequencies, energy-based features, Mel-Frequency Cepstral Coefficients (MFCC) and similar.

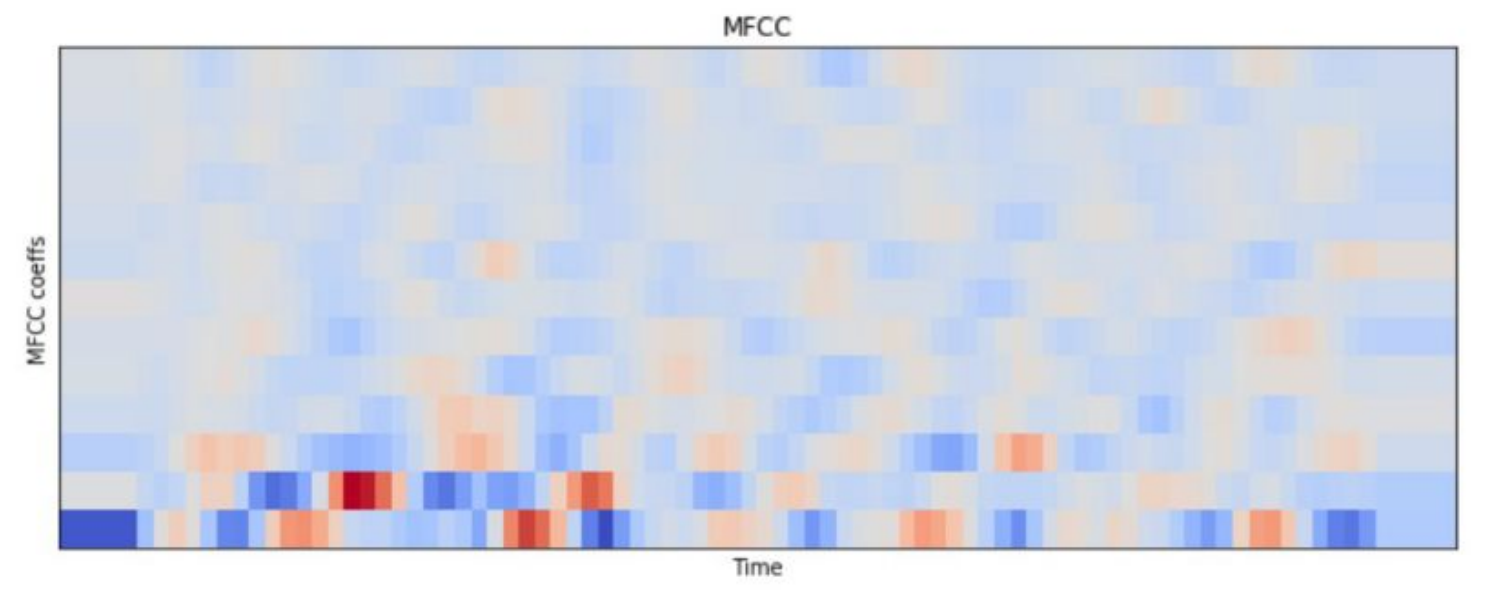
- **Prosodic**

This type of features measures peculiarities of speech like pauses between words, prosodies and loudness. These speech details depend on a speaker, and use of them in the speaker-free systems is debatable. Therefore they are not used in this work.

- **Linguistic**

These features are based on semantic information contained in speech. Exact transcriptions require a lot of assessor's work. In future it is possible to include speech recognition to the pipeline to use automatically recognized text. But for now authors do not use linguistic features.

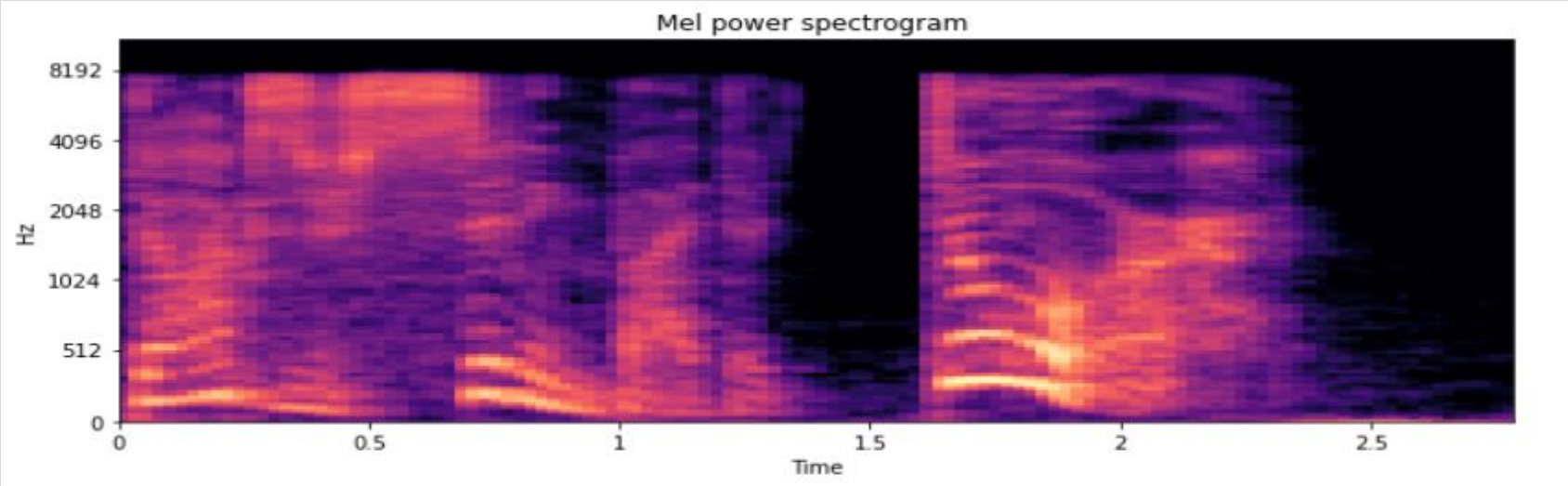
Feature Extraction



Mel Frequency Cepstral Coefficients

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum")

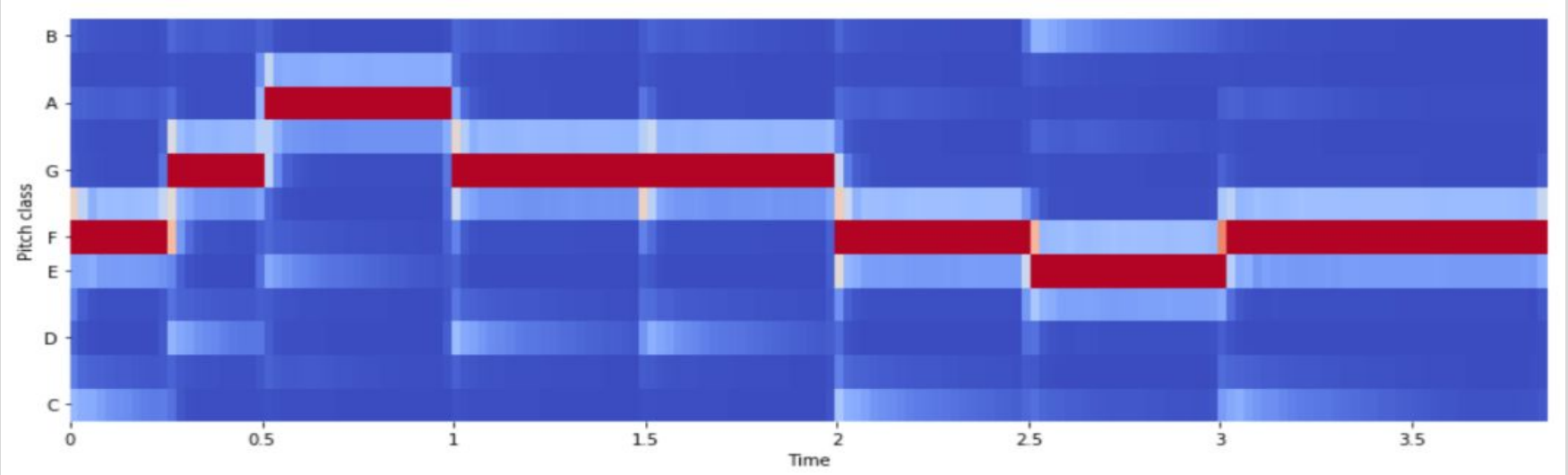
Feature Extraction



Mel Spectrogram

A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. In a spectrogram representation plot — one axis represents the time, the second axis represents frequencies and the colors represent magnitude (amplitude) of the observed frequency at a particular time.

Feature Extraction



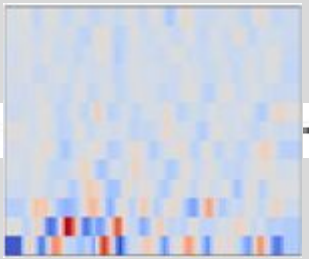
Chroma

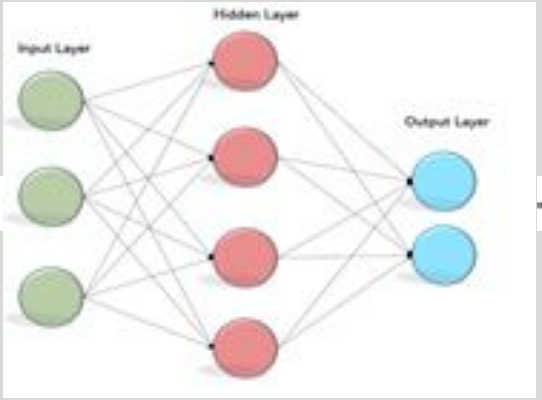
Chroma feature or chromagram closely relates to the twelve different pitch classes.The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. It is computed in two steps: • First, the spectrum is computed in the logarithmic scale, with selection of the 20 highest dB, and restriction to a certain frequency range that covers an integer number of octaves. • Then, the spectrum energy is redistributed along the different pitches (i.e., chromas).

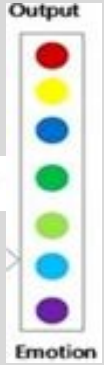
DEEP LEARNING METHODOLOGIES USED

1. Multi Layer Perceptron(MLP)
2. Convolutional Neural Network(CNN)
3. Convolutional Neural Network - Long Short Term Network(CNN-LSTM)

MULTI LAYER PERCEPTRON







Speech signal

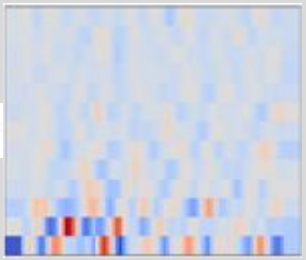
Feature Extraction

MLP Classifier

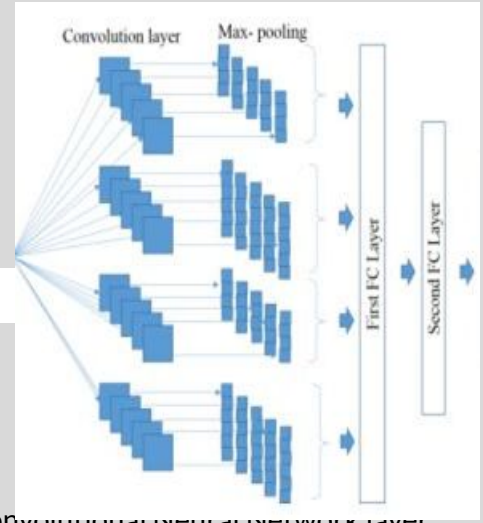
Emotion

CONVOLUTIONAL NEURAL NETWORK

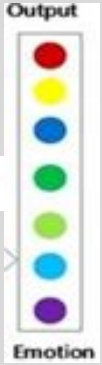
Speech signal



Feature Extraction

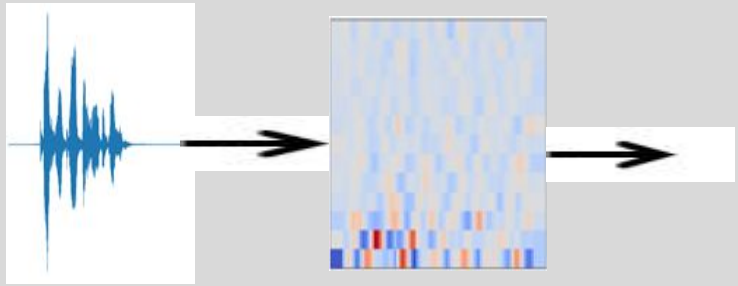


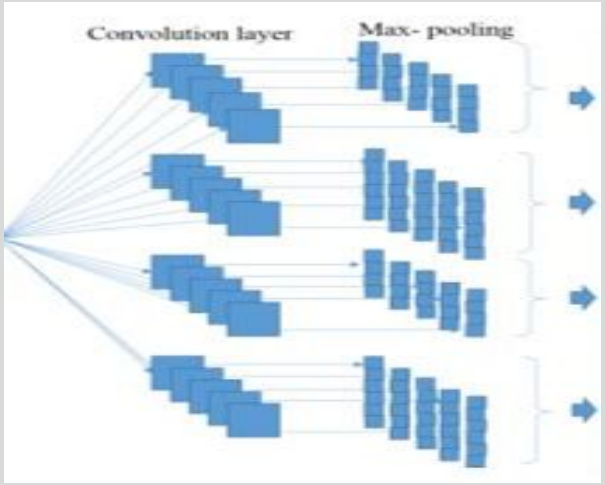
Convolutional Neural Network layer

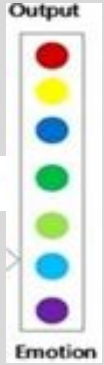


Emotion

CNN-LSTM







Speech signal

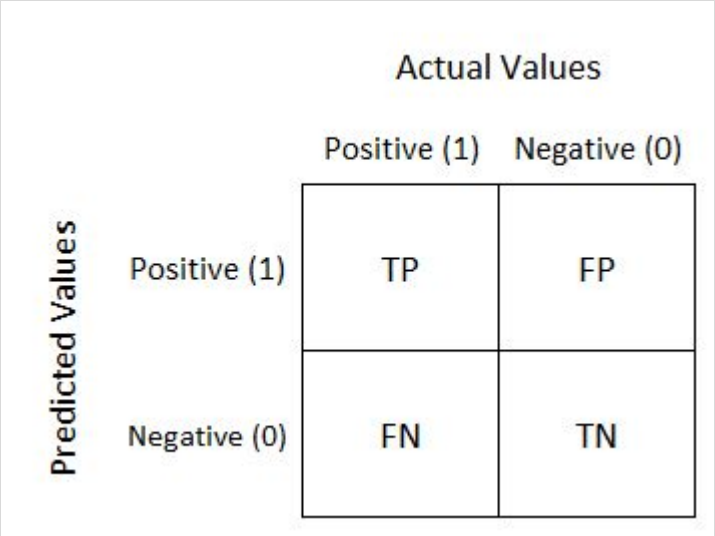
Feature Extraction

CNN Long Short-Term Memory Network

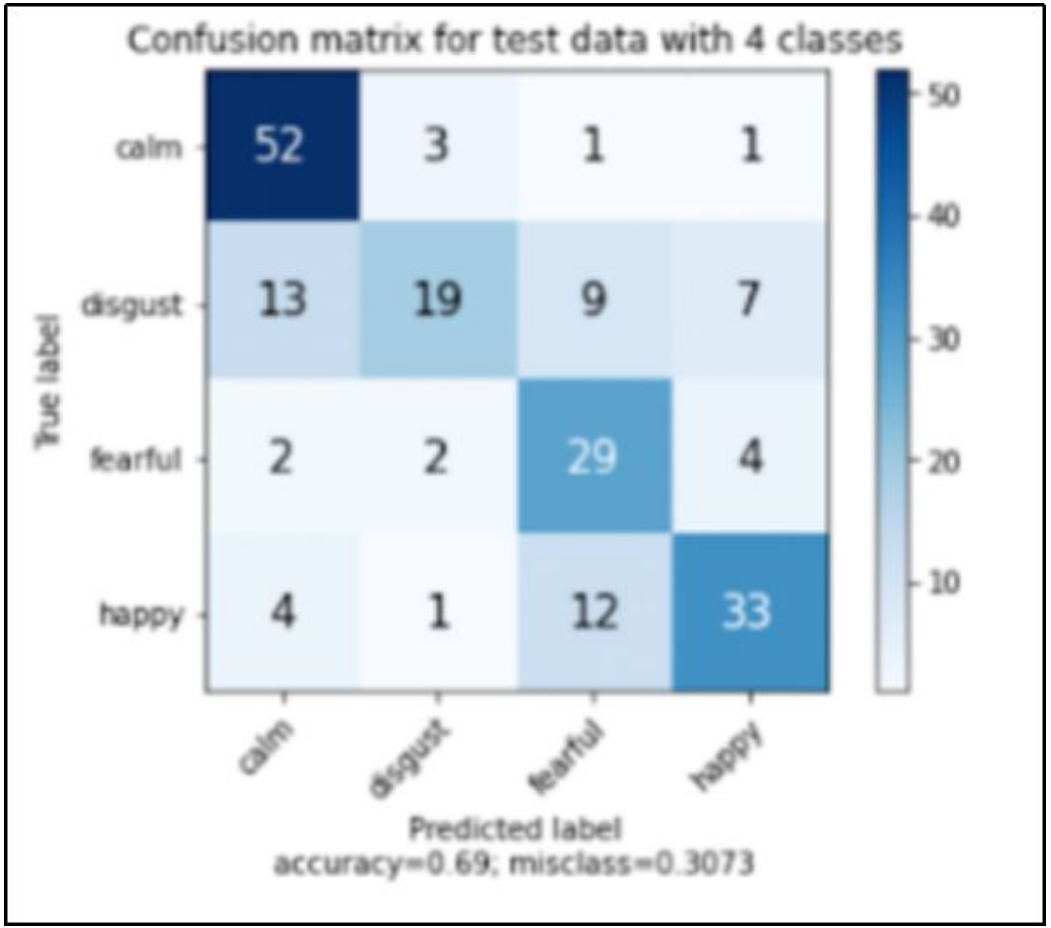
Emotion

RESULTS

To evaluate the performance of classification tasks in machine learning we use the confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

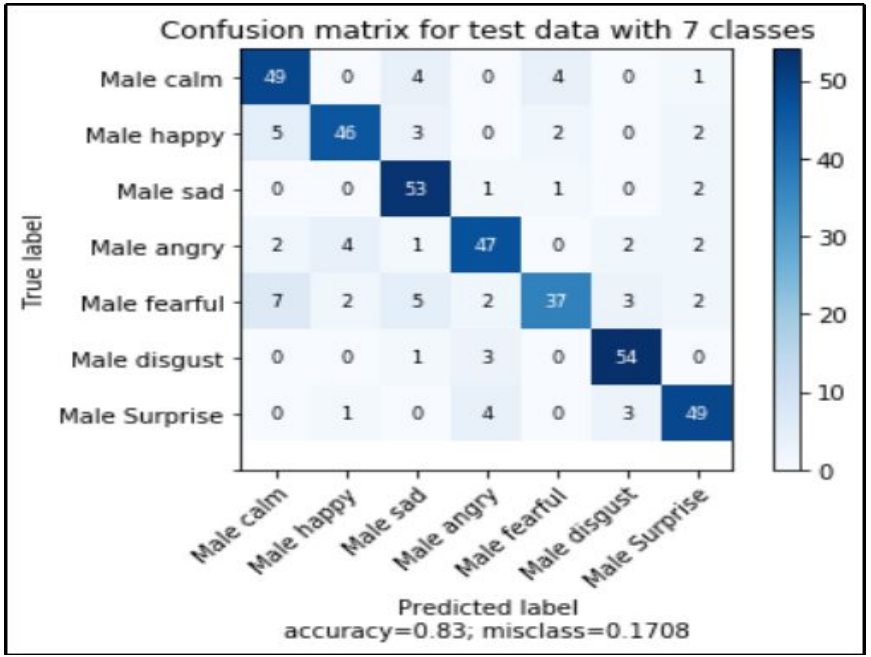


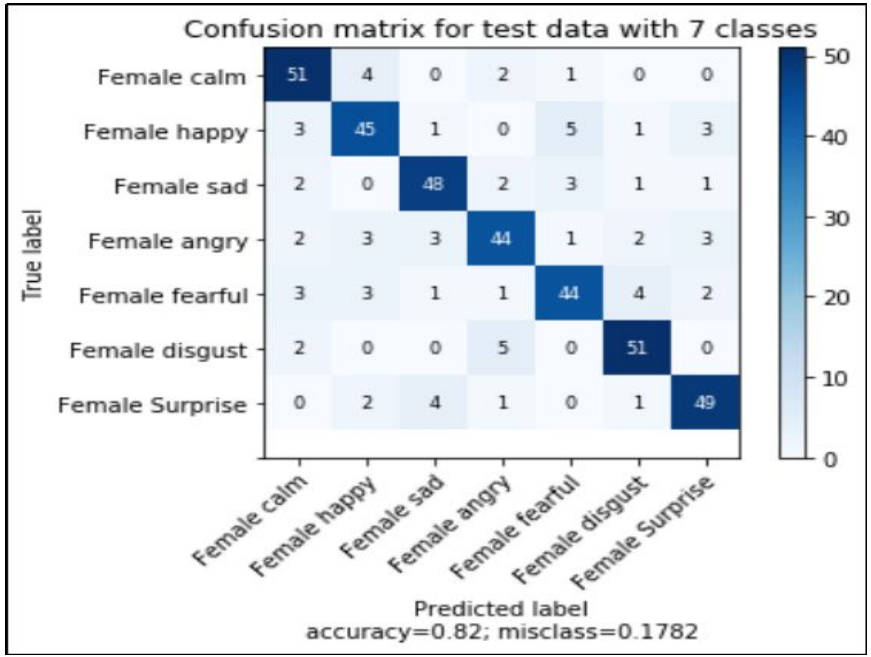
MULTI LAYER PERCEPTRON



The MLP classifier gives us an accuracy of 69% for the four basic emotions calm,disgust fearful and happy

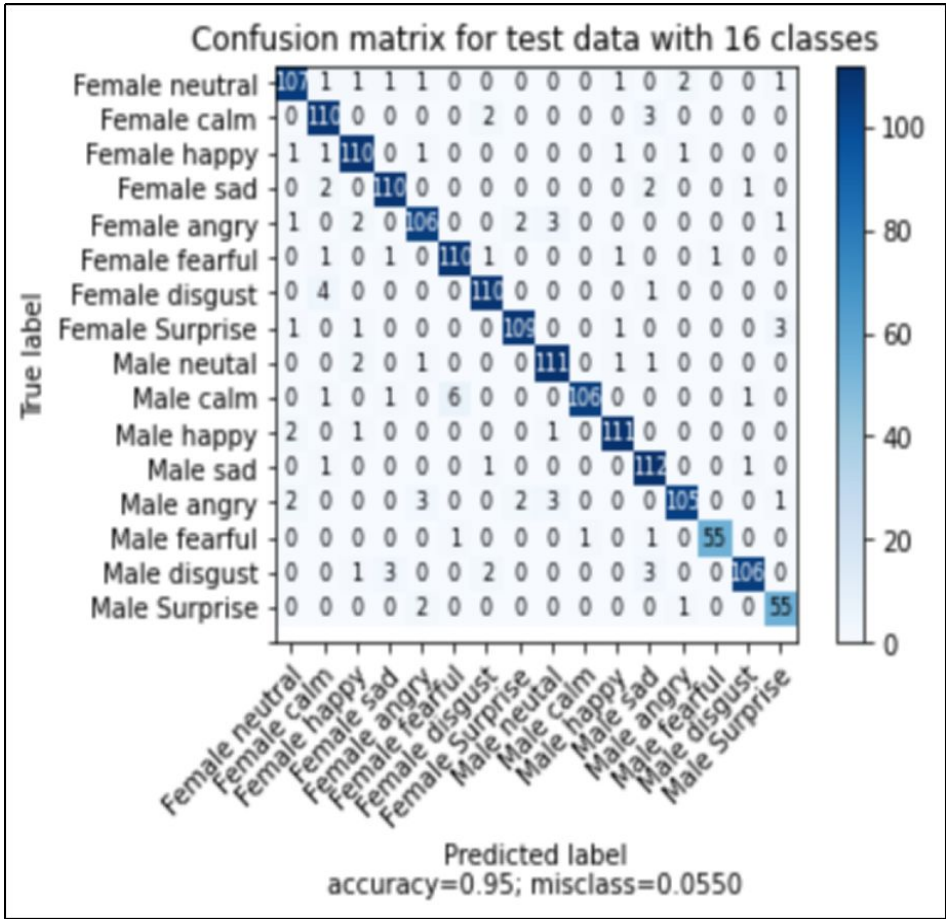
CONVOLUTIONAL NEURAL NETWORK





The CNN model gives us an accuracy of 83% for emotions by male speakers and an accuracy of 82% for female speakers.

CNN-LSTM



CNN-LSTM model recognises 16 classes of emotion and gender namely **Female neutral,Female calm,Female happy,Female sad,Female angry,Female fearful,Female disgust,Female Surprise,Male neutral,Male calm,Male happy,Male sad,Male angry,Male fearful,Male disgust,Male Surprise** with an accuracy of 95%

FUTURE SCOPE

Emotional AI is on the rise

Amazon, which operates the Alexa digital assistant in millions of people's homes, has filed patents for emotion-detecting technology that would recognise whether a user is happy, angry, sad, fearful or stressed. That could, say, help Alexa select what mood music to play or how to personalise a shopping offer.

Affectiva has developed an in-vehicle emotion recognition system, using cameras and microphones, to sense whether a driver is drowsy, distracted or angry and can respond by tugging the seatbelt or lowering the temperature.

Fujitsu, the Japanese IT conglomerate, is incorporating "line of sight" sensors in shop floor mannequins and sending push notifications to nearby sales staff suggesting how they can best personalise their service to customers.

THANKYOU