Netaji Subhas university of Technology

Speech to emotion recognition and prioritization

Name :- Punit

Roll number :- 2019UCO1605

Section :- 2

Branch:- Computer engineering

Abstract :-

At peak times (like in festivals or some catastrophic emergency) the number of calls to emergency numbers increases manifold on emergency numbers . Prioritizing calls can be tough in these situations. We present a solution here to tackle this problem.

In general this system can prioritize calls in real time ,whether it is a call center that wants to tackle angry customers first or a mental health hospital that wants to prioritize sad patient calls first.

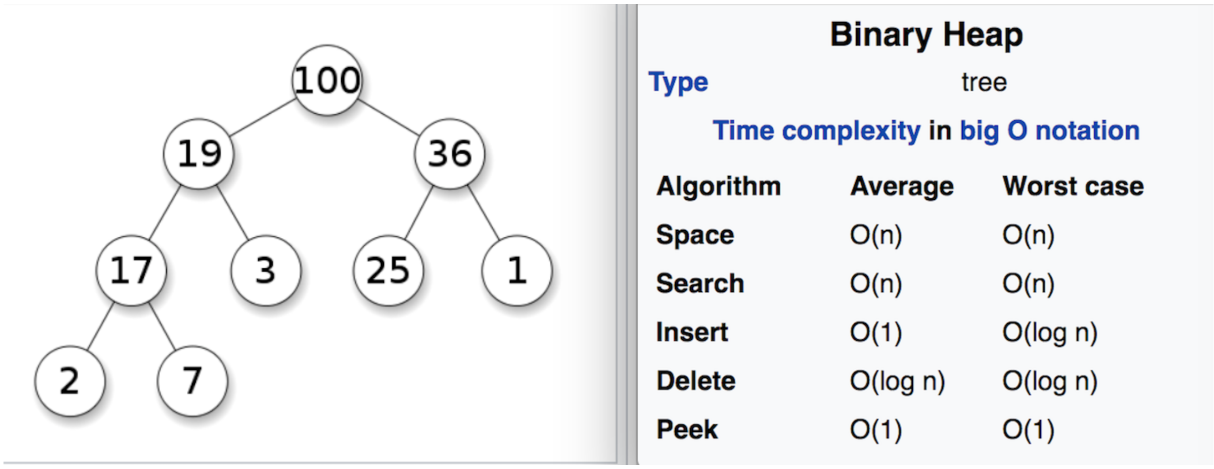
To address this we can use different machine learning algorithms. Using machine learning we can find many patterns like age, gender or emotion of the caller. Using this information we can classify emergency calls. In this report, we are finding patterns of emotion of different callers. We can prioritize these emotions using a binary heap as per requirement of emergency services like fearful emotion for emergency services like ambulance,police or fire brigade services and happy ,joyful for hotel booking, food ordering in cash on delivery mode.

Introduction :-

The aim of emotion recognition in speech is to automatically detect emotions from human audio.

Emotion recognition in spoken dialogues has been gaining increasing interest in recent years.Speech Emotion Recognition (SER) is a hot research topic in the Human Computer Interaction(HCI) field.

Another part of this project is to devise a system so that audio can be prioritized on the basis of emotions . Since this prioritization is to be done online , we decided to use a binary heap. Every query that comes can be inserted in the heap with time complexity of O(logn) .



Binary Heap

Type	tree	
Time complexity in big O notation		
Algorithm	**Average**	**Worst case**
Space	O(n)	O(n)
Search	O(n)	O(n)
Insert	O(1)	O(log n)
Delete	O(log n)	O(log n)
Peek	O(1)	O(1)

Time complexity analysis of binary heap

We first tried classification using logistic Regression but we were getting low accuracy with that. So we decided to use Multilayer Perceptron (MLP)
We tried different combinations of layers and units in them and then took 1 hidden layer with 300 units in it . Sampling rate was taken to be 16000 , So for a 1 second audio file it checks 16000 times for Mel-spectrogram, Mel-frequency cepstral coefficients (MFCCs) and chroma . After extracting features , we pass it through MLP and then insert into a binary heap . The order is predefined using python dictionary data structure and with the index of the file it is pushed into the heap data structure after making a pair.

Objective :-

1. Our Speech Recognition system can flag\report fake emergency calls and prioritize these emotions from most fearful to joyful emotions.
2. We will implement this with the help of Max Heap (Priority queue) .
3. As a person tries to make an emergency call there is a buffer time allotted in which we analyze their sound profiles with the help of machine learning algorithms like MLP Classifier to store them in Priority queues.
4. And then we address the important calls first (calls which have fear/angry emotions)

Flowchart :-

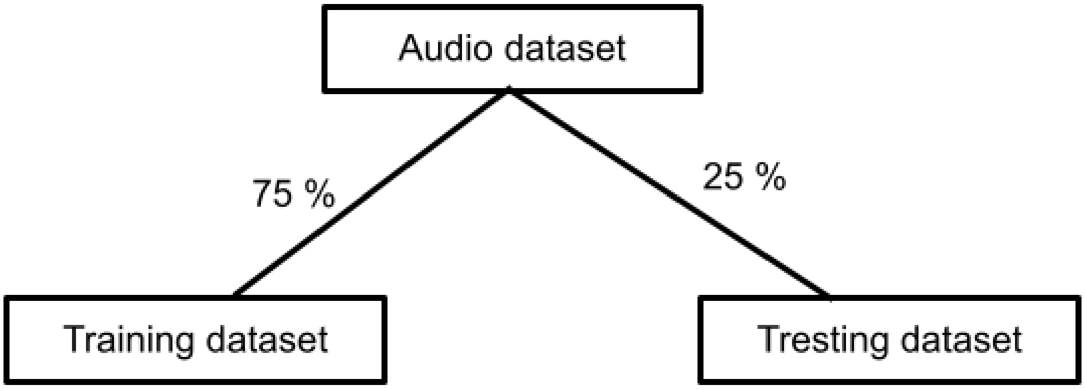


Figure 1

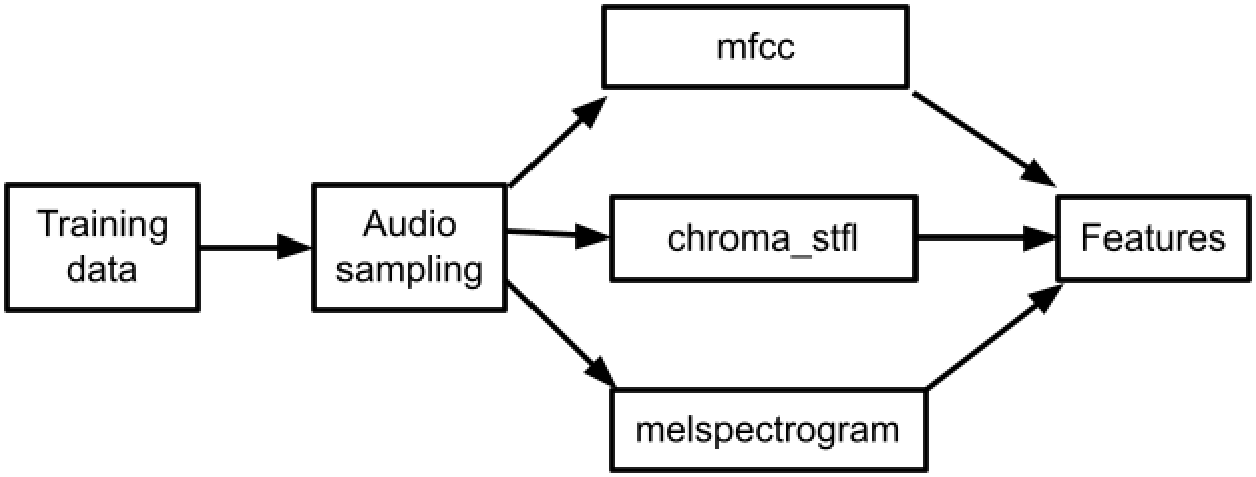
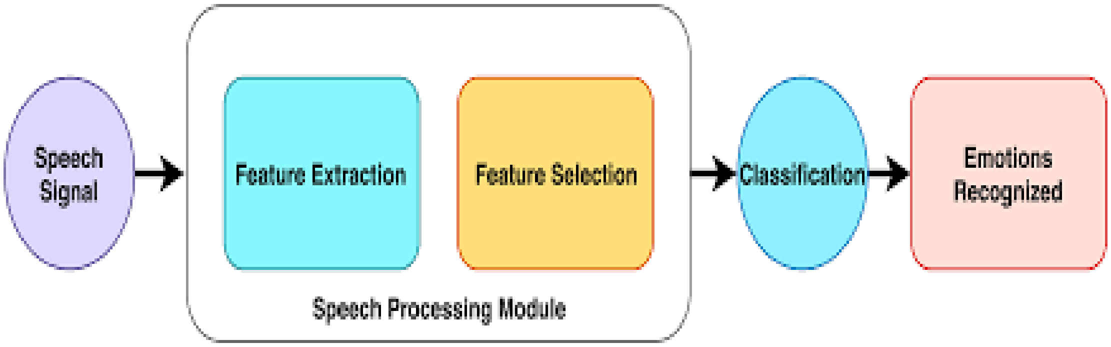


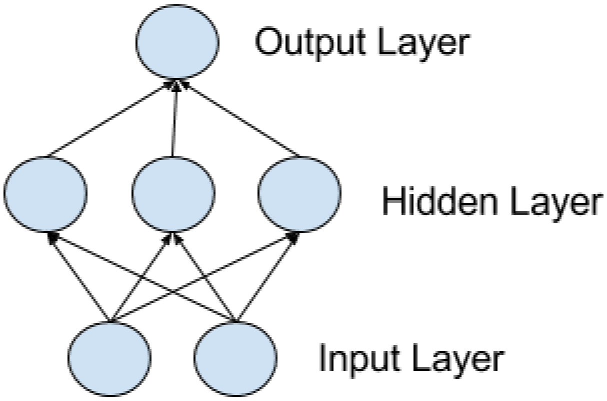
Figure 2



Methodology :-

Multi Layer Perceptron:-

Multi Layer Perceptron(MLP) is a class of feedforward artificial neural network(ANN). An MLP consists of at least 3 layers-input layer,hidden layer and output layer. MLP uses backpropagation for training.Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.In MLP neurons are arranged into network of neurons in which a row of neuron is known as layer

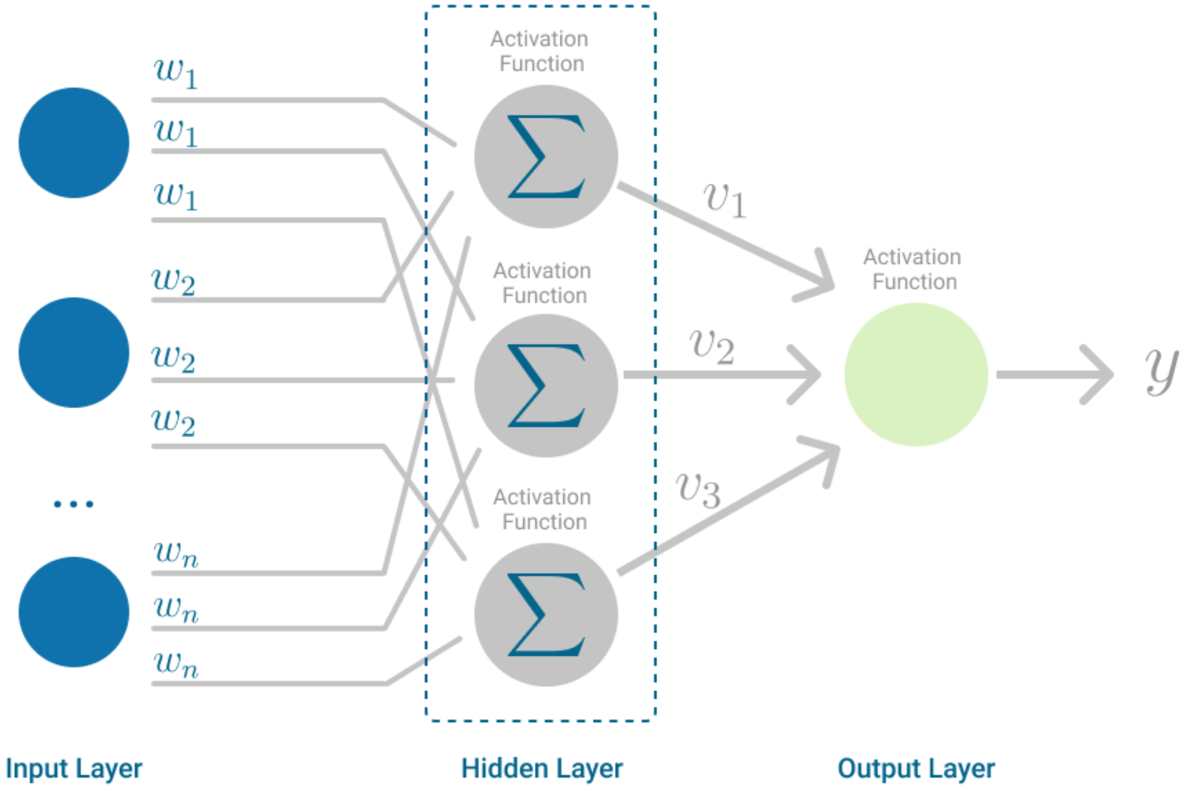


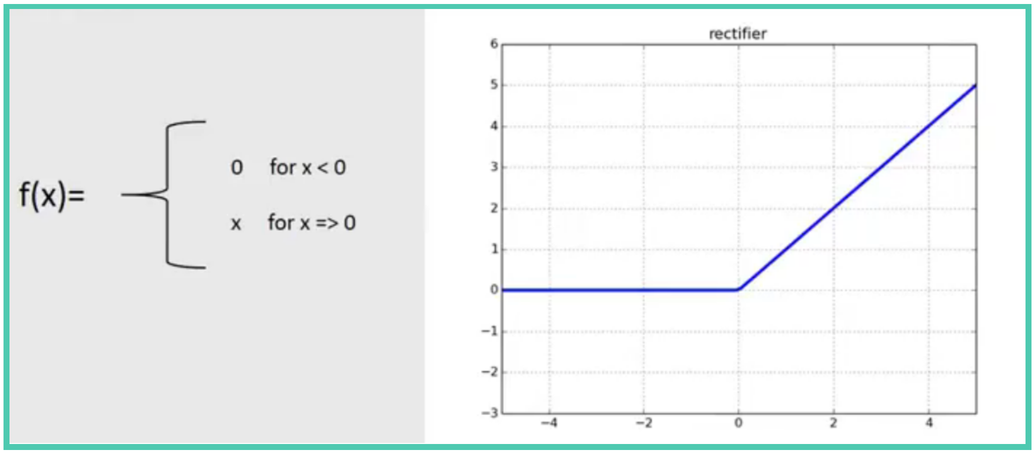
Input Layer:-

The lowermost layer that takes input from dataset is called input layer.It may contain one or more than one neuron.

Hidden Layer:- Layers after the input layer are called hidden layer because they are not directly exposed to input layer. A neural network may have more than one hidden layer

Output Layer:- The final layer that is responsible for outputting a value or vector of values that correspond to the format required for the problem.





ReLU function used as activation function in our Multilayer perceptron

In BackPropagation, after the weighted sums are forwarded through all layers, the gradient of mean squared error is computed across all input and output layers Then, to propagate it back, the weights of the first hidden layer are updated with the value of the gradient. That's how the weights are propagated back to the starting point in Back Propagation.

$$\underbrace{\Delta_w(t)}_{\text{Gradient Current Iteration}} = -\overset{\text{Bias}}{\varepsilon}\,\frac{\overset{\text{Error}}{dE}}{\underbrace{dw_{(t)}}_{\text{Weight vector}}} + \overset{\text{Learning Rate}}{\alpha}\underbrace{\Delta_{w(t-1)}}_{\text{Gradient Previous Iteration}}$$

Now with the help of MLP ,the speech to emotion detection system is implemented as a machine learning model.

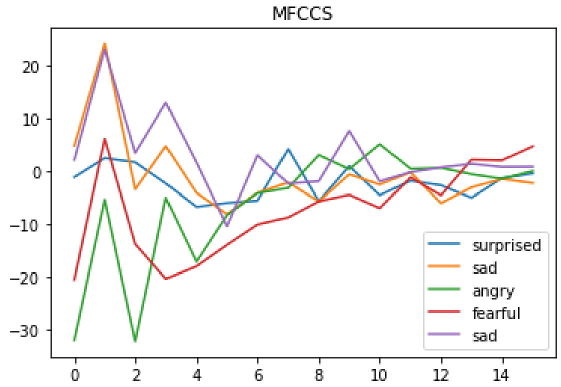
Algorithm :-

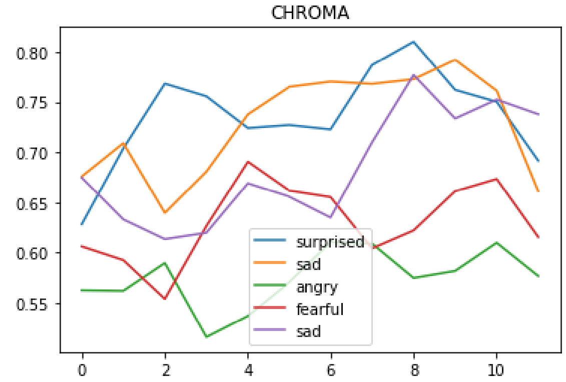
- Load data from the audio files.
- Do sampling using sound_file library.
- Extract features from sampled data.
    - Extract mfcc feature using librosa library.
    - Extract chroma_stfl feature using librosa library.
    - Extract mel spectrogram feature using librosa library.
    - Push the mean of each feature into a list.
    - Return the features list of the file.
- Append features of all files in X.
- Make a dictionary of emotion.
- Extract emotion from each file name and append it in Y.
- Split testing and training data in 25% and 75% respectively.
- Create machine learning model using MLPclassifier with few conditions:-
    - Alpha =0.01
    - Learning_rate = ‘adaptive’
    - batch_size=256
    - hidden_layer_sizes=300

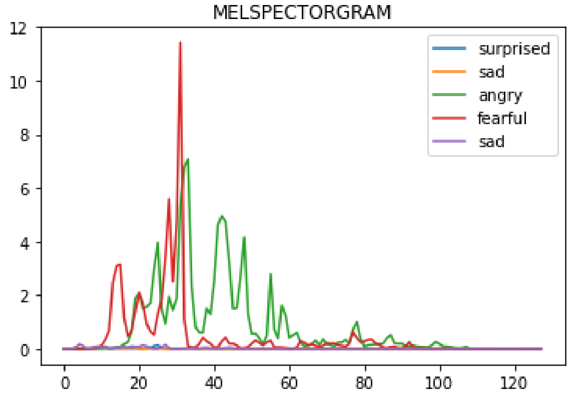
- Training model on training data.
- After training the model, find the accuracy of the model on testing data. So, after significant accuracy we can deploy our model for prioritizing calls.

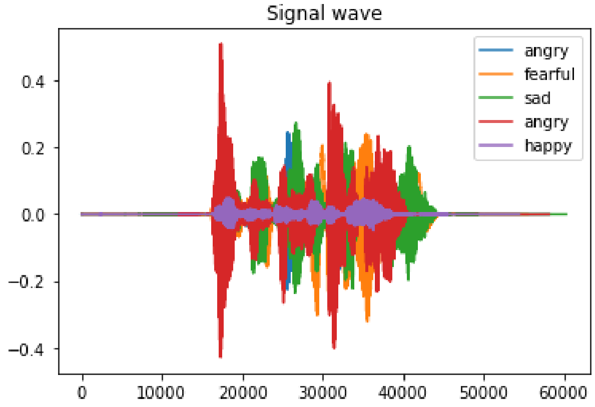
About features of audio file :-

In this project, we extracted the main 3 features from the audio file. Which are Mel-frequency cepstral coefficient, chroma and mel-spectrogram.







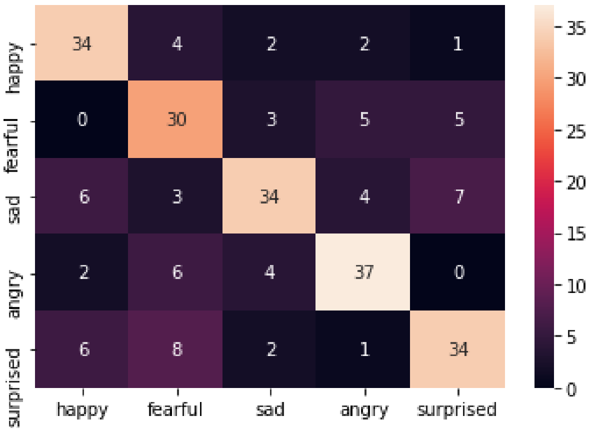


- Mel-frequency cepstral coefficients(MFCCs) give a list of boldness in sample data.With the help of mfccs we can predict most accurate audio. It returns a list of size 20. We removed the first 2 features of the list because of the redundant value of start.
- Chroma is used to map an audio sample on the fixed 12 notes of English music. Chroma returns a list of size 12.
- Melspectrogam uses mel scale to plot frequency vs time. It return a list 128 size. This list represents frequency at the nth time in the list.

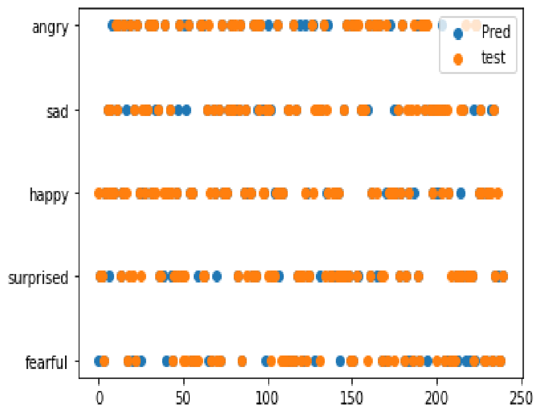
- Signal wave represent sample strength vs time(in millisecond).

Result :-

The accuracy of our model is close to 72.83 %.Through this model we can easily classify different sound signal into angry , fearful, joyful,surprised and neutral.From this accuracy result we can infer that out of 10 audio samples we can easily distinguish 7 and arrange them in priority queue out of which -Highest Priority is given to Fearful(Distress) emotion afterwards Surprised,Angry, Joyful and other emotions.



	happy	fearful	sad	angry	surprised
happy	34	4	2	2	1
fearful	0	30	3	5	5
sad	6	3	34	4	7
angry	2	6	4	37	0
surprised	6	8	2	1	34



Plot 1

Plot 2

Plot 1 is a confusion matrix representing correct and incorrect predictions or classification. With the help of confusion matrix we can find accuracy, recall, precision and F1 score.

Plot 2 is a scatter to show how well our data fitted on test data. We plotted the predicted result with blue ink.Then above which we plotted the actual result with orange ink. In graph places where blue colour is visible in data is error.

Dataset :-

We used Ryerson audio visual dataset from kaggle.That dataset contains 24 actors with each has 60 files of different emotions like happy, sad, fearful, anry, calm and surprise. Each audio fill name is formed with different information about the file.
Dataset link :-
https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

Other Related Work :-

1. http://ijasret.com/VolumeArticles/FullTextPDF/829_34.SPEECH_BASED_EMOTION_RECOGNITION.pdf
2. http://www.aagasc.edu.in/cs/msccs/4-MLP.pdf
3. https://www.researchgate.net/publication/320089581_Speech_emotion_recognition_with_deep_learning