Emotion Recognition Using Excitation Features

Team 1
Ayush Singhania (20171031)
Swayatta Daw (2020702016)

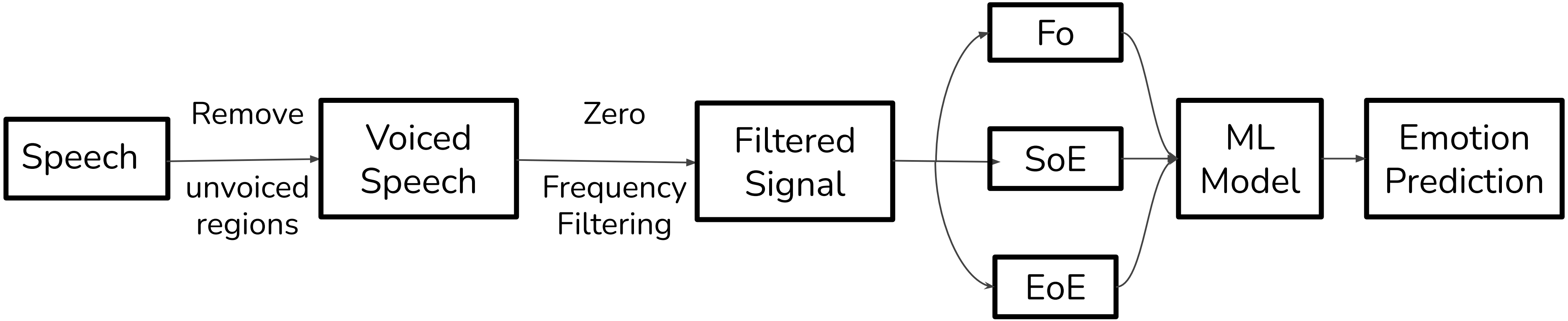
Aim

- To develop an emotion recognition model capable of classifying speech based on different emotional states of the speaker.

- 4 different emotions considered: Happy, Angry, Sad and Neutral.

- Speech features:
  - Vocal Tract
  - Excitation (Used in this project)
  - Prosody

Algo


Speech
Remove
unvoiced
regions
Voiced
Speech
Zero
Frequency
Filtering
Filtered
Signal
Fo
SoE
EoE
ML
Model
Emotion
Prediction

Removing Unvoiced Speech

- Energy Thresholding is used
- Threshold = max_energy / 20
- All speech frames with energy more than the threshold are considered

Zero Frequency Filtering

- Effects of excitation features is present at all frequencies.
- ZFF removes the effects of the vocal tract related features as they are present beyond a certain frequency threshold.
- Steps:
    - **Pre-emphasis :** x[n] = s[n] - s[n-1]
    - **Apply ZFF twice**
    - **Trend removal by subtracting moving average from each sample:**

$$y[n] = y2[n] - (1/(2N+1)) \sum_{m=-N}^{m=+N} y2[n+m]$$

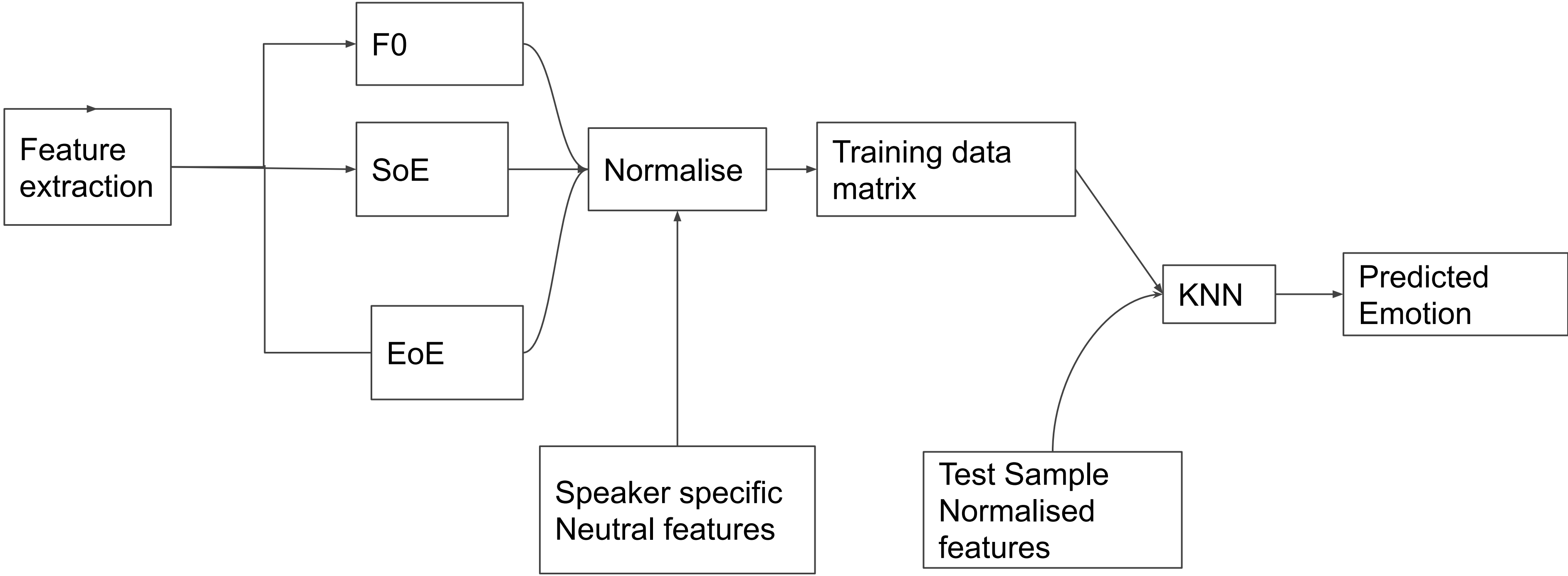
Feature Extraction

1. **Instantaneous Frequency (Fo)** : Inverse of difference in duration between two successive GCIs

2. **Strength of Excitation (SoE)** : Slope of the ZFF signal at each GCI

3. **Energy of Excitation (EoE)** : Energy of the Hilbert Envelope of the LP residual of the ZFF signal over 2ms duration around each GCI.

ML algo


Feature extraction
F0
SoE
EoE
Normalise
Training data matrix
KNN
Predicted Emotion
Speaker specific Neutral features
Test Sample Normalised features

Creating the data matrix :

1. **Normalising :**
   a) We first obtain the speaker specific neutral characteristics , the mean and the standard deviation for all the training samples.
   b) We normalise the feature with respect to the mean and std dev of that speaker.

$$N_{R_{F_0}} = \frac{R_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}} \qquad N_{E_{F_0}} = \frac{E_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}}$$

2. The audio samples are stacked row wise, with their respective class labels. The columns form the normalised features of each audio sample. Each row is zero-padded and the final data matrix is created.

3. Each test sample is normalised with respect to their neutral speaker features.

4. **KNN :-**

a) The squared sum of the euclidean distance is calculated for each test sample from all training samples :

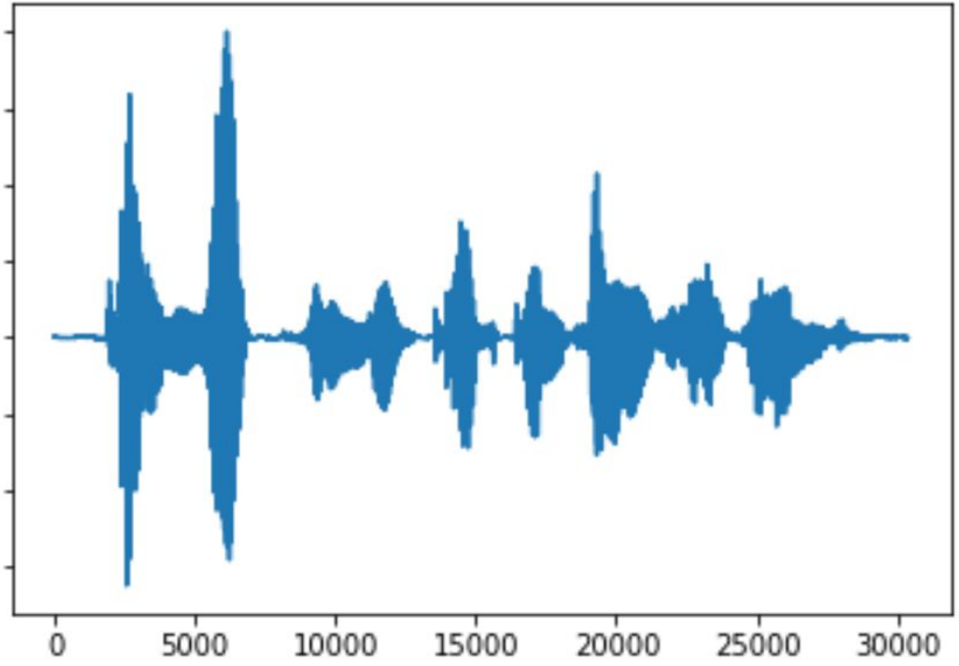
$$d = \sqrt{d1^{2} + d2^{2} + d3^{2}}$$

b) We find the K closest neighbours. Take K = 8 for best results.

c) **Prediction :**

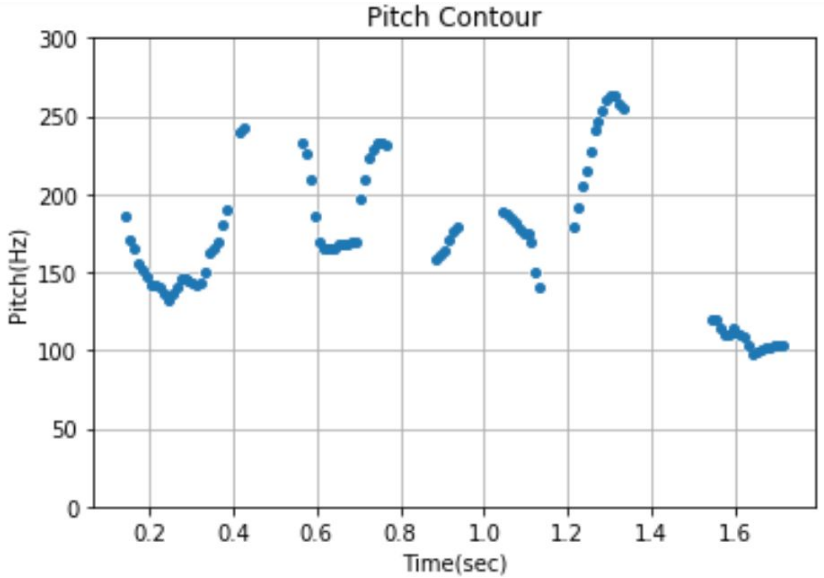
We find the class label with the maximum occurance. The test sample is classified based on that.

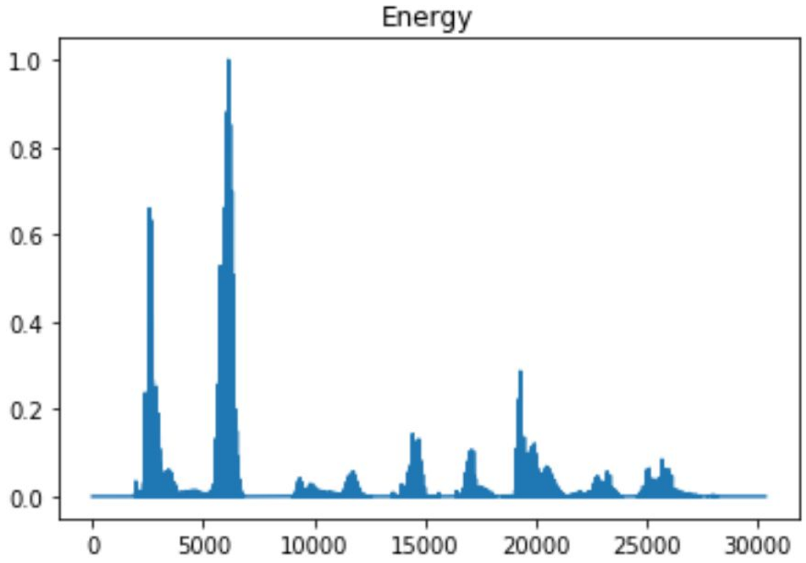
Results



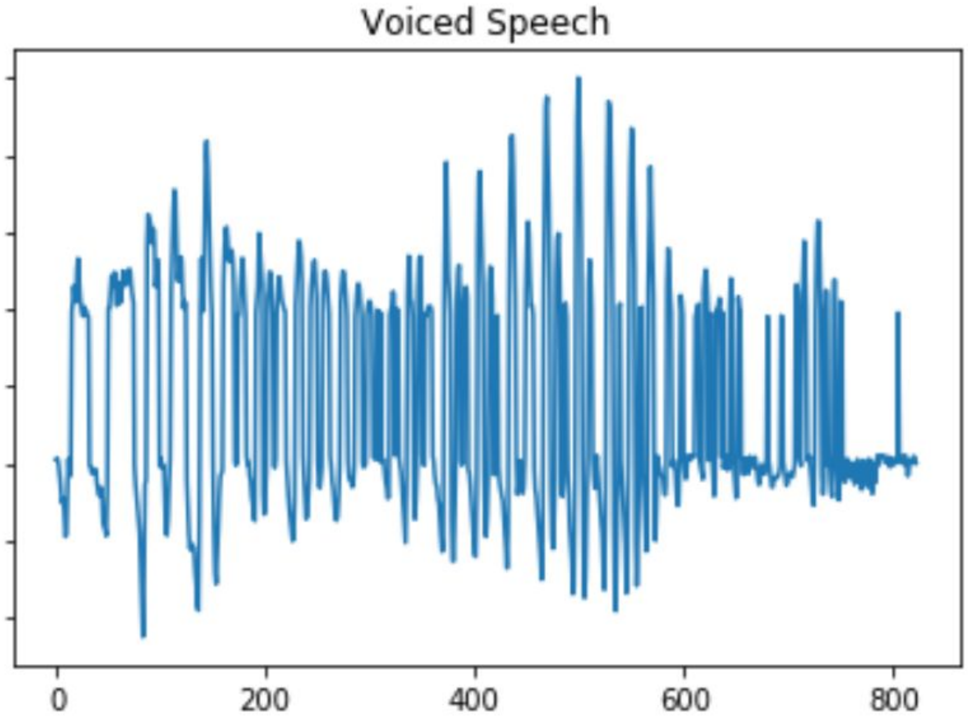
Input Speech

Results

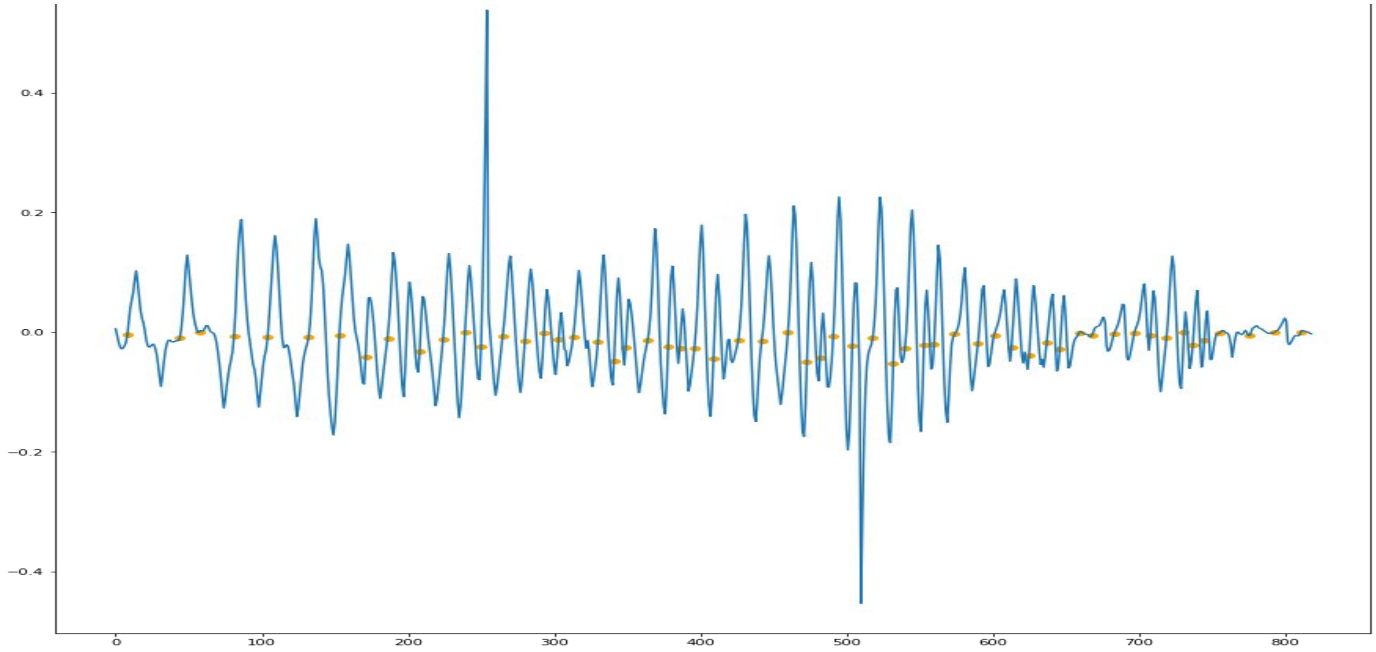

Pitch Contour
Pitch(Hz)
Time(sec)
300
250
200
150
100
50
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


Energy
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000

Results

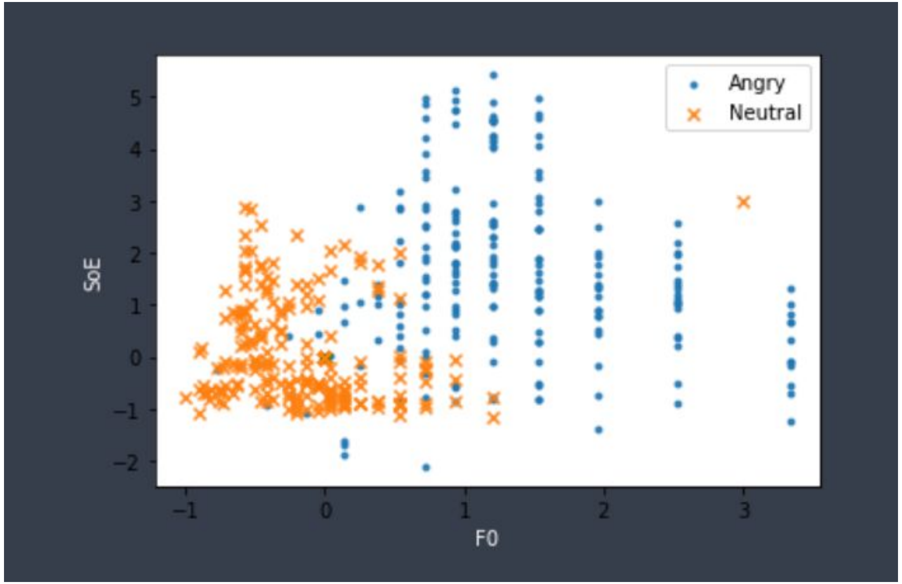


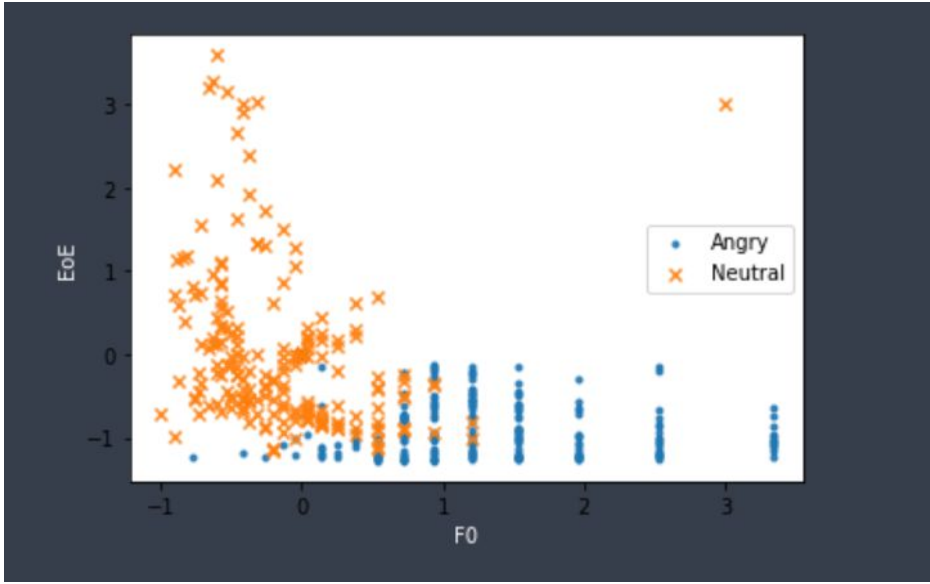
Results



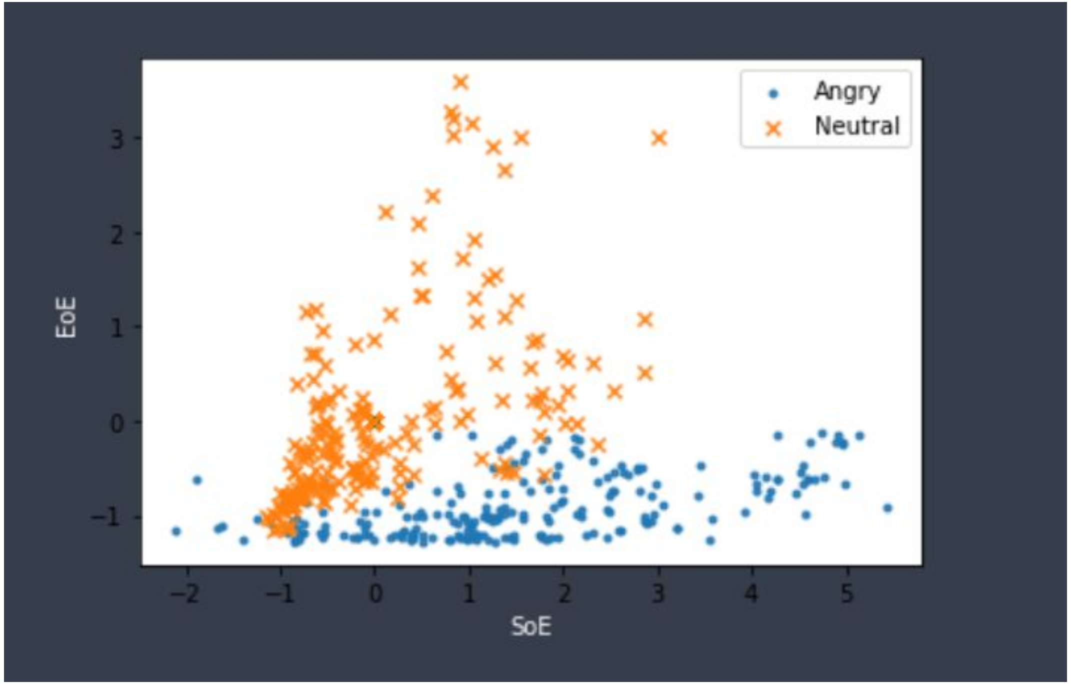
ZFF signal and GCI

Results

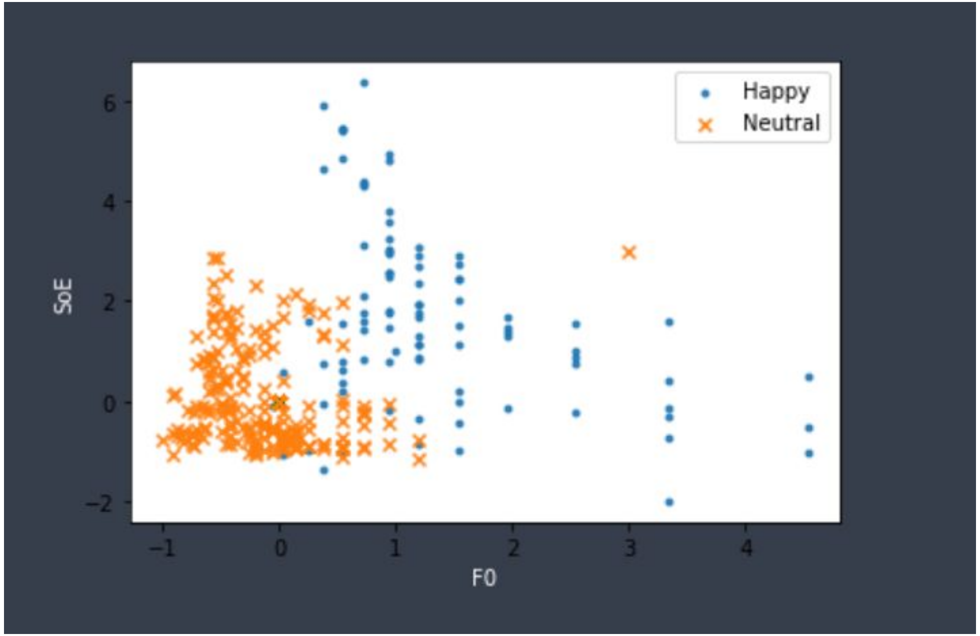

Angry
Neutral
SoE
F0
5
4
3
2
1
0
−1
−2
−1
0
1
2
3

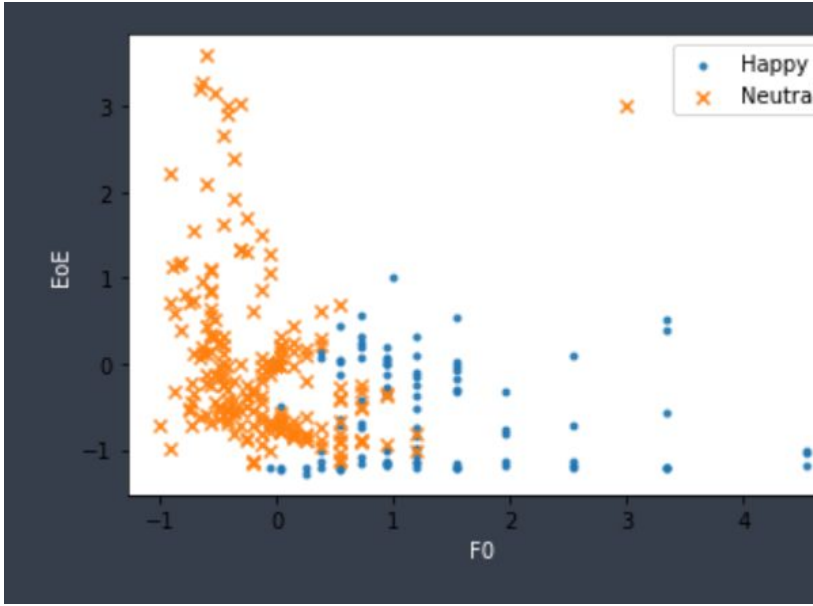

Angry
Neutral
EoE
F0
3
2
1
0
−1
−1
0
1
2
3

Results

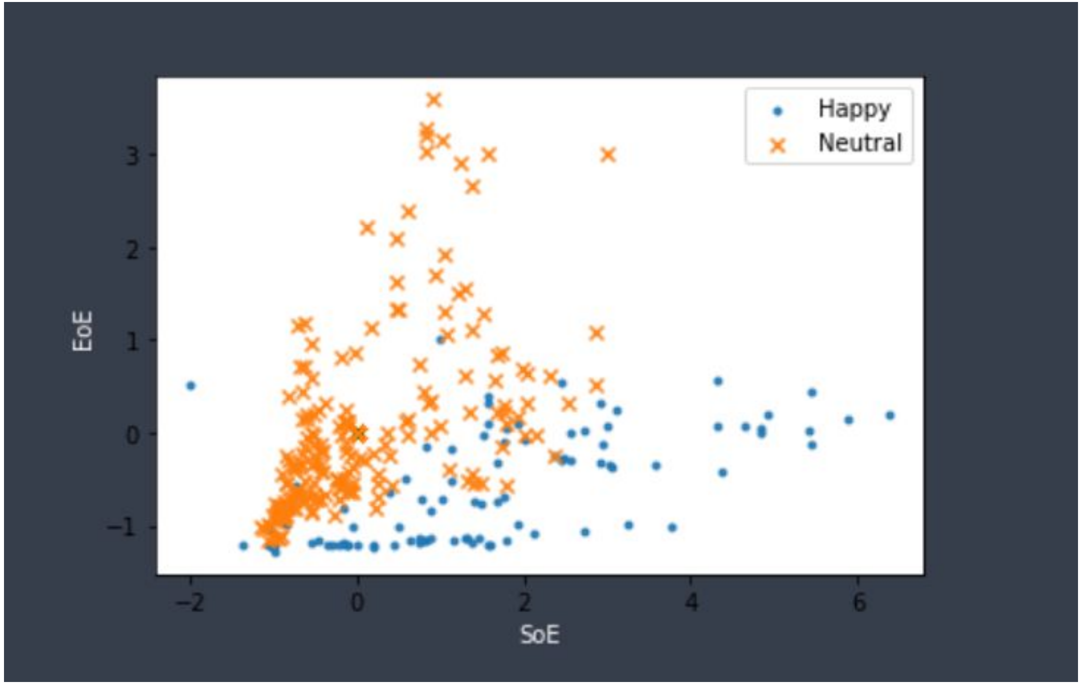


Results





Results



Final Results

Accuracy :

The Accuracy obtained based on this method is 78.33 %

Confusion Matrix :

	Angry	Happy	Sad	Neutral
Angry	0.67	0.27	0.0	0.07
Happy	0.07	0.8	0.0	0.13
Sad	0.0	0.0	1.0	0.0
Neutral	0.0	0.0	0.33	0.67

References

1. Analysis of Excitation Source Features of Speech for Emotion Recognition
   Sudarsana Reddy Kadiri, P. Gangamohan, Suryakanth V Gangashetty and B. Yegnanarayana

2. Database of German emotional speech
   Felix Burkhardt, Walter Sendelmeier