International Institute of Information Technology, Hyderabad

Syllabification for Speech Emotion Recognition

Arshini Govindu
EC5.412 Speech Analysis and Linguistics

Overview

Syllabification And SER

- Syllabification
  - Preprocessing
  - Feature extraction
  - Voicing
  - VOP Detection
  - Syllable Boundary Detection
  - Removal Spurious Boundaries
- Speech Emotion Recognition
  -

Introduction

Syllabification is useful for a variety of reasons and so is speech emotion recognition

Problem

Speech emotion recognition can be done using a variety of features such as mfccs and spectrogram.

Syllabification can be done by using VOP detection plots and energy contour and speech/non-speech detection regions.

Syllabification

Methodology

- Normalise the speech signal.
- STE with windowing.
- Speech/Non-Speech detection using the majority of STE, the most dominant frequency and voicing information.
- VOP detection using HE of LP residual.
- FOGD operator on smoothed LP residual signal.
- If VOP is more than one in the speech region, it is an indicator of the presence of syllable boundary.
- Valley points in STE in that region are taken as the syllable boundary. Valley threshold is set to detect valley points
- Spurious boundaries are removed if there are more than one detected boundaries within a specific time frame.

Steps

First, frame the signal into windows of 20ms and 10ms using a hamming window.
Then, calculate the short-time energies for each frame and plot them.

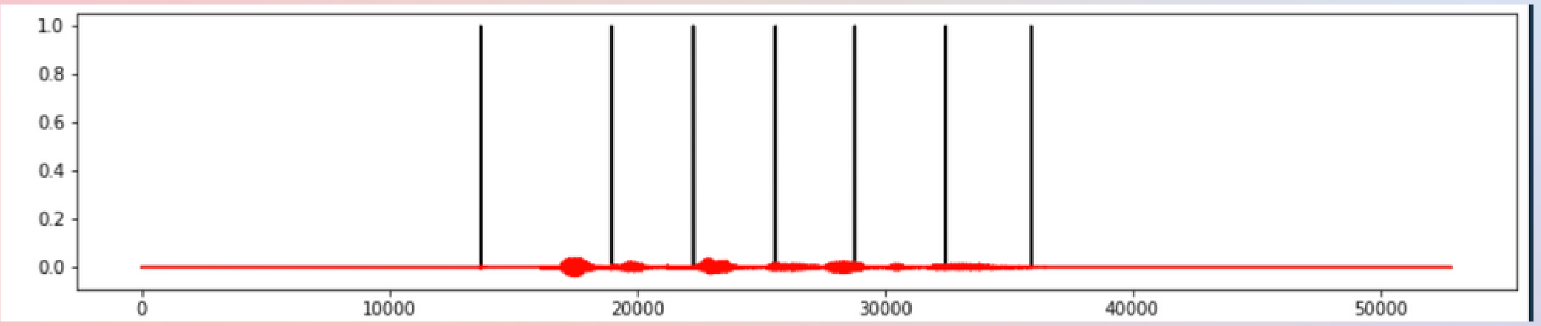
Write a subtopic or any related idea.

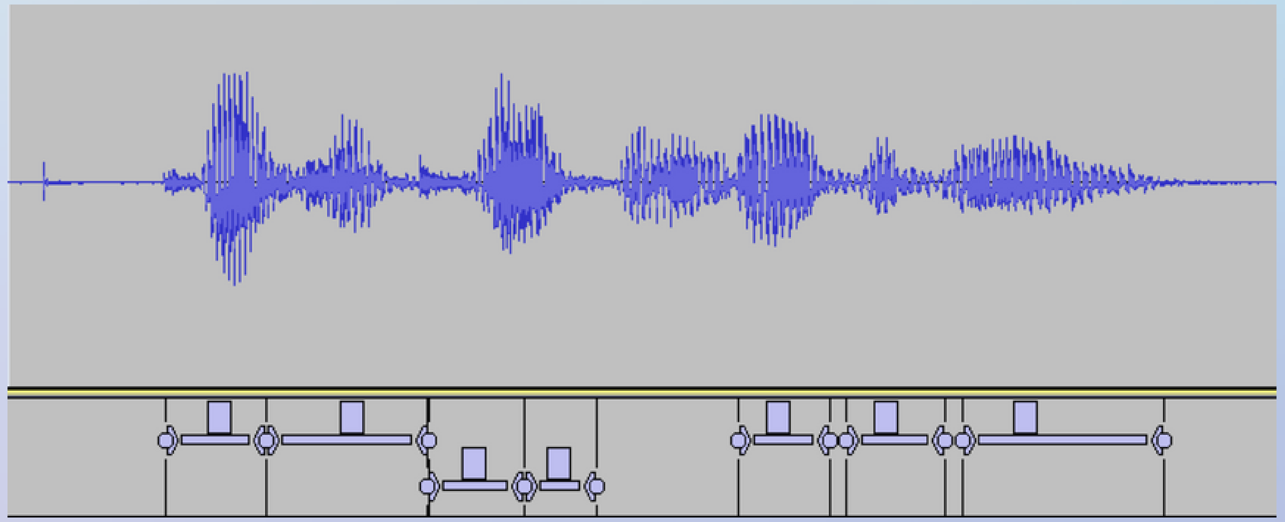
Find the magic and fun in presenting with Canva Presentations by pressing C for confetti, D for a drumroll, and O for bubbles.

- Speech/non-speech threshold for STE is chosen as ratio of standard deviation to mean of STE.
- The peaks in the VOP evidence plot represent the locations of the VOPs and are automatically located by finding the maximum value between two succes- sive positive to negative zero crossing with some threshold.
- Then the valley point in STE contour located between two consecutive VOPs is identified as the syllable boundary
- A syllable boundary at time ti is said to be spurious if there exists another syllable boundary at time tj such that $Etj < Eti$ and $| ti - tj | < 40ms$.

Code

The code takes an audio file as input and then gives the syllable boundaries on the waveform as output. The syllable boundaries detected by the code are almost similar to the manually-marked syllabel boundaries although it misses a few low energy syllable boundaries.





Speech Emotion Recognition

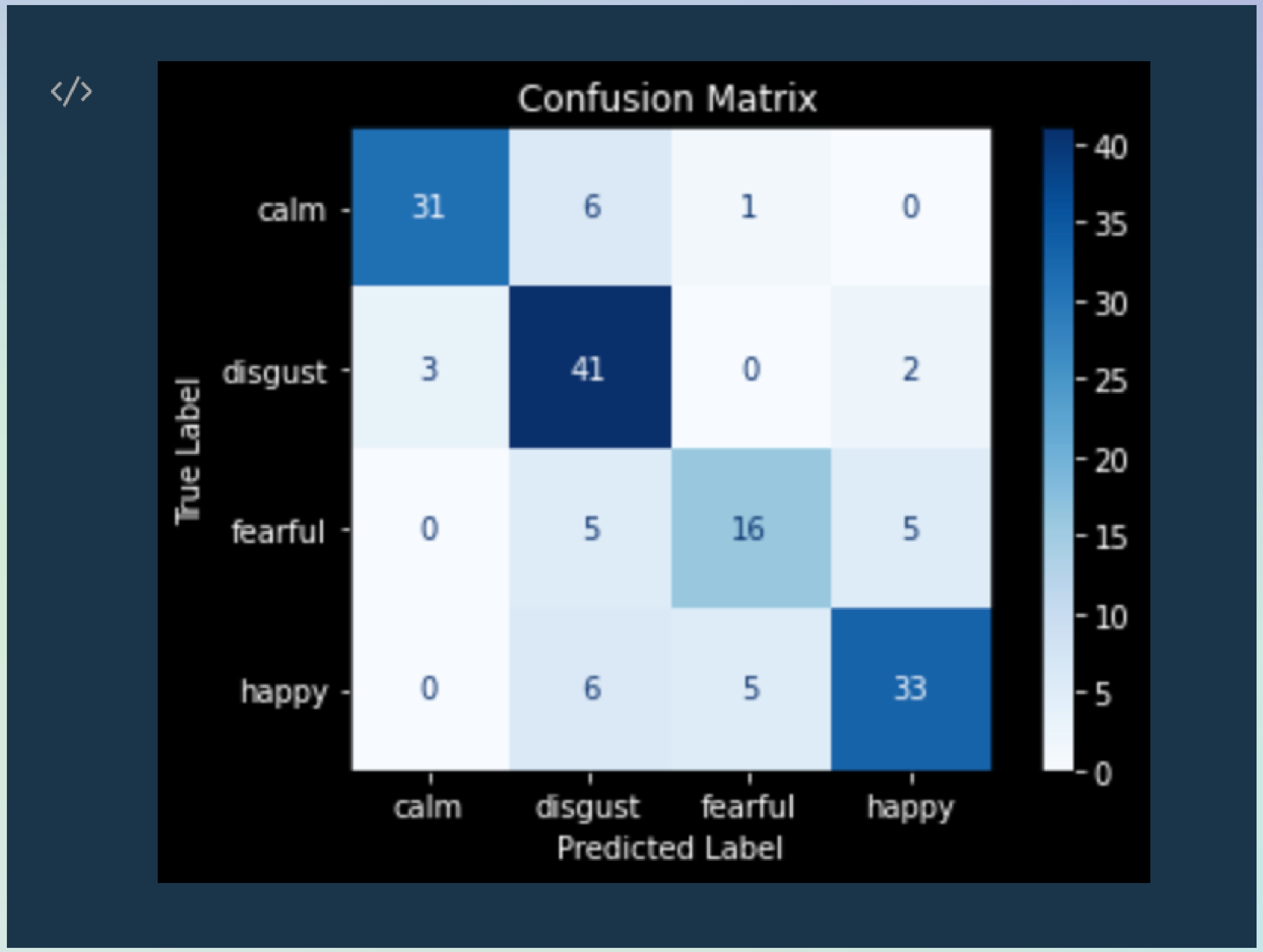
Overview

We use a Multi-Layer Perceptron to train this model. This is a feedforward ANN model. We used the RAVDESS dataset which has 60 recordings each by 24 participants rated on their emotions. The dataset contains 8 emotins but we use only 4 emotions to improve accuracy.

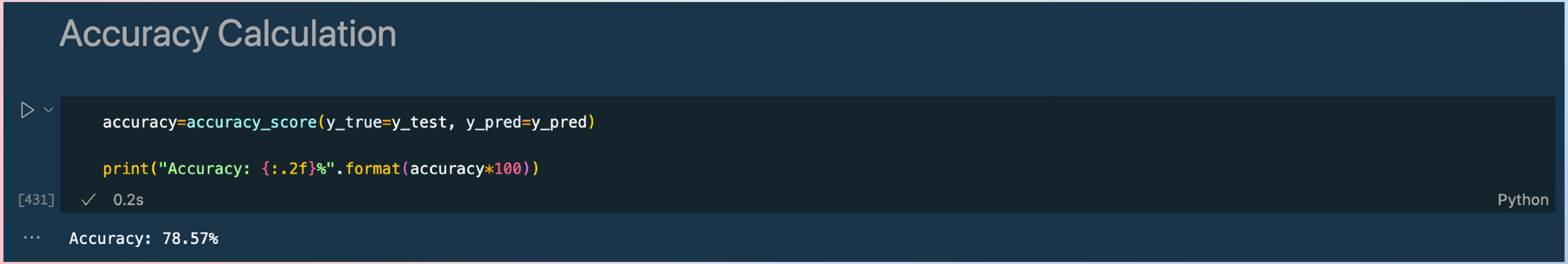
Methodology

- First, we access an audio file and display it.
- Next, we write a function to extract features: mfccs, pitch classes and spectrogram
- Then we define a dictionary of the four emotions we consider.
- The data is then split into test and train after selecting only the audio files which are from these 4 emotions.
- The model is trained with a 80:20 split
- Then the model predicts the emotions for the test set.
- We check the accuracies for the predicted emotions matching the y_test set.

Results



	calm	disgust	fearful	happy
calm	31	6	1	0
disgust	3	41	0	2
fearful	0	5	16	5
happy	0	6	5	33



Accuracy Calculation

```python
accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

print("Accuracy: {:.2f}%".format(accuracy*100))
```

```
Accuracy: 78.57%
```