Time Frequency Analysis

Shashwat Shrivastava	20161181
Harsh Sharma	20171157

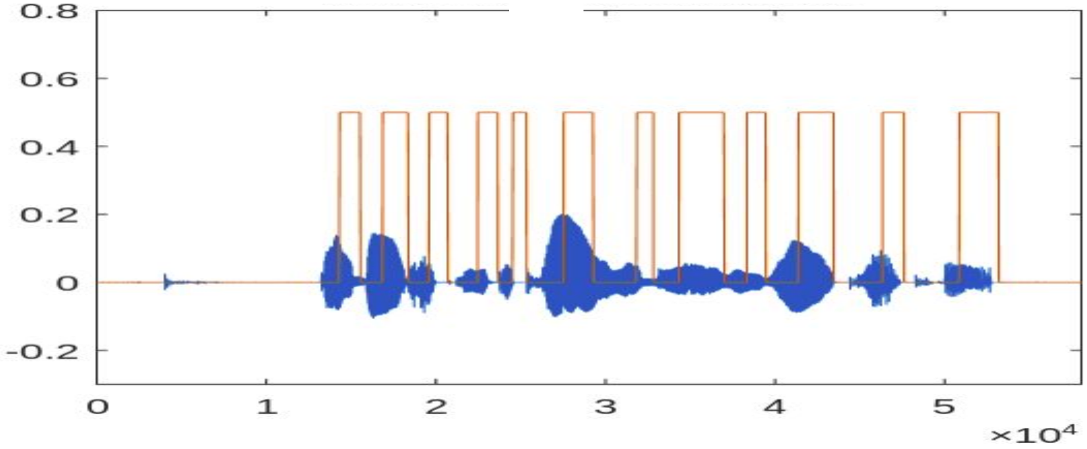
Problem Statement

Detection of Vowel Region using non-negative frequency weighted energy operator

AIM

Output detected vowel region in the speech signal.

“don’t ask me to carry an oily rag like that” 🔊



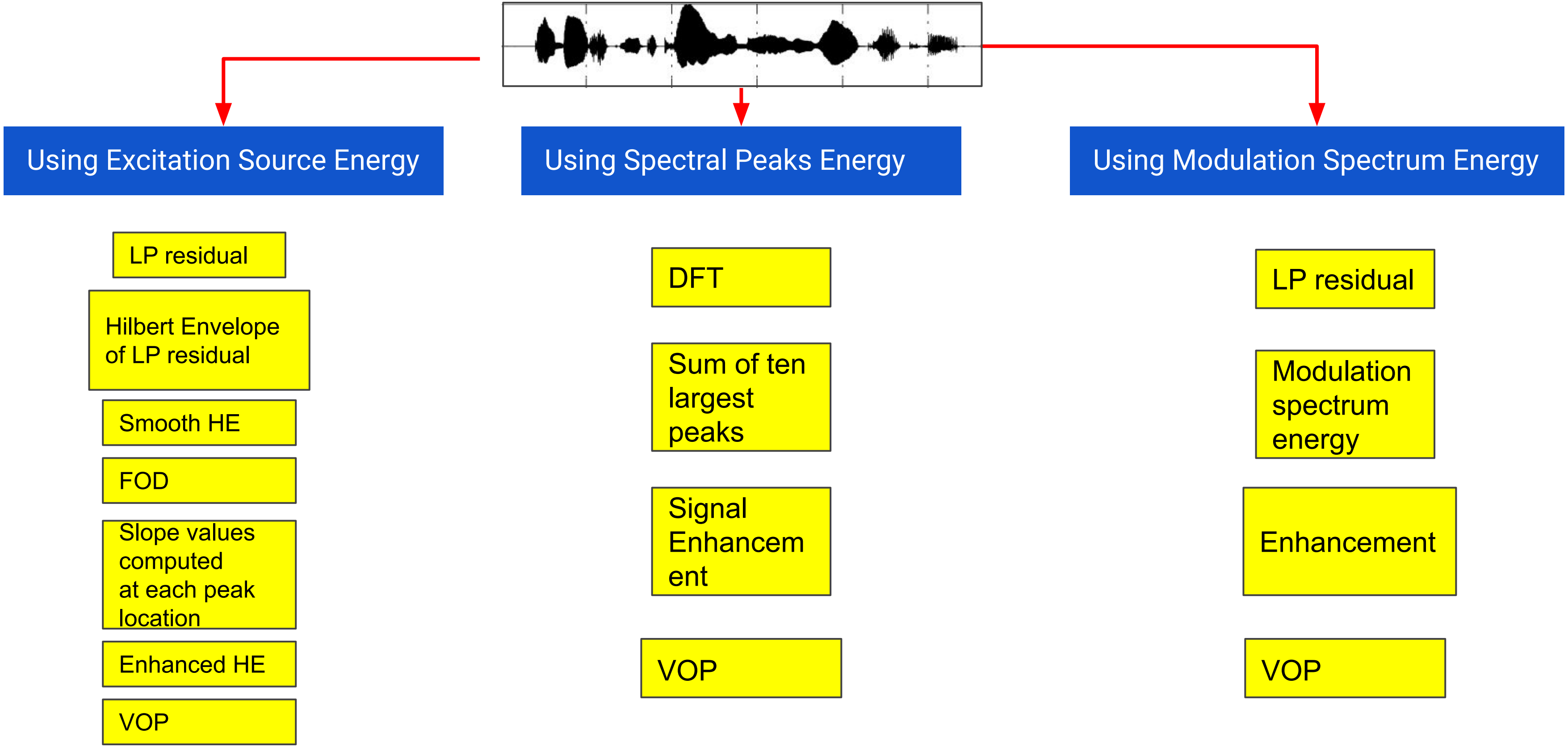
Some Important points on Vowel

- A speech signal is composed of voiced signal, unvoiced signal and noise. Vowels along with few consonants a considered as voiced while others as unvoiced.
- VOWELS are produced by keeping the vocal tract in an open position with minimum obstruction along the length and using glottal vibration as the excitation
- Voiced speech is produced by vibrations in vocal cords, this can have high energy compared to unvoiced speech which are low energy
- Therefore voiced signals tend to be louder like the vowels a, e, i, u, o. Unvoiced signals, on the other hand, tend to be more abrupt like the stop consonants p, t, k.
- Vowel Onset Point is the instant at which the beginning of a Vowel takes place and Vowel End Point is the instant at which the ending of Vowel take place during speech production.
- There are significant changes occurring in the energies of excitation source, spectral peaks, and modulation spectrum at the point when a vowel is pronounced in speech.
- So our project is to detect Vowel Region using non-negative frequency-weighted energy operator

Baseline Methods

1. **COMB**
2. **FGCI**

COMB: VOP Detection



Using Excitation Source Energy

LP residual

Hilbert Envelope of LP residual

Smooth HE

FOD

Slope values computed at each peak location

Enhanced HE

VOP

Using Spectral Peaks Energy

DFT

Sum of ten largest peaks

Signal Enhancement

VOP

Using Modulation Spectrum Energy

LP residual

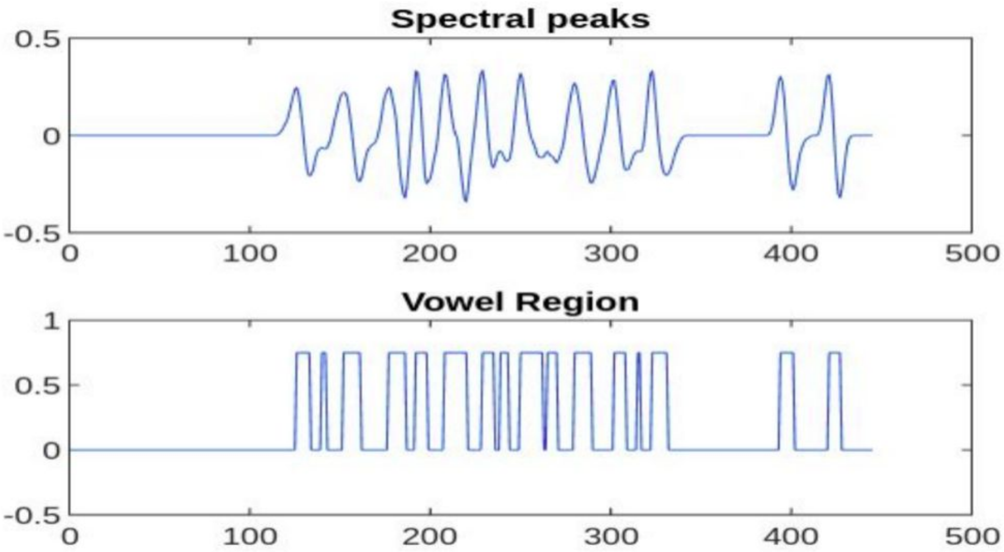
Modulation spectrum energy

Enhancement

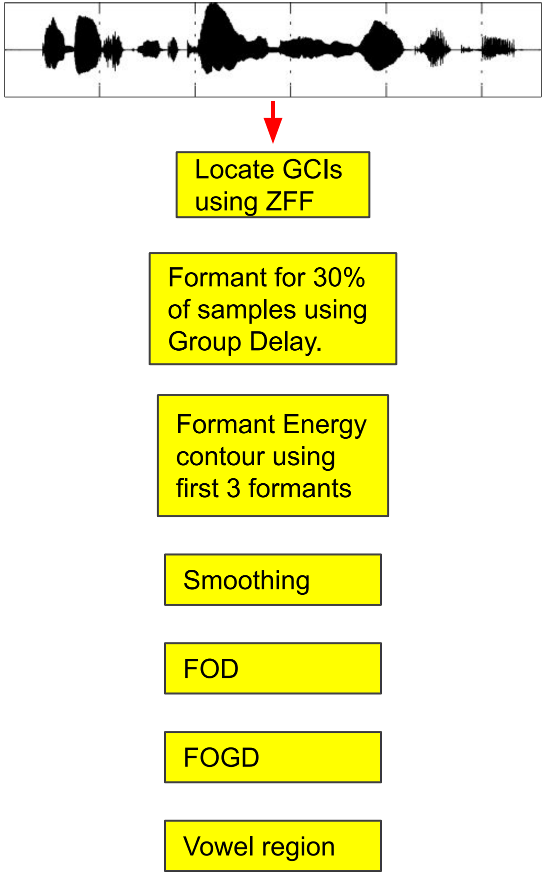
VOP

COMB: VOP Detection

Combine three VOP evidences from above to generate a final VOP evidence.



FGCI



Locate GCIs using ZFF

Formant for 30% of samples using Group Delay.

Formant Energy contour using first 3 formants

Smoothing

FOD

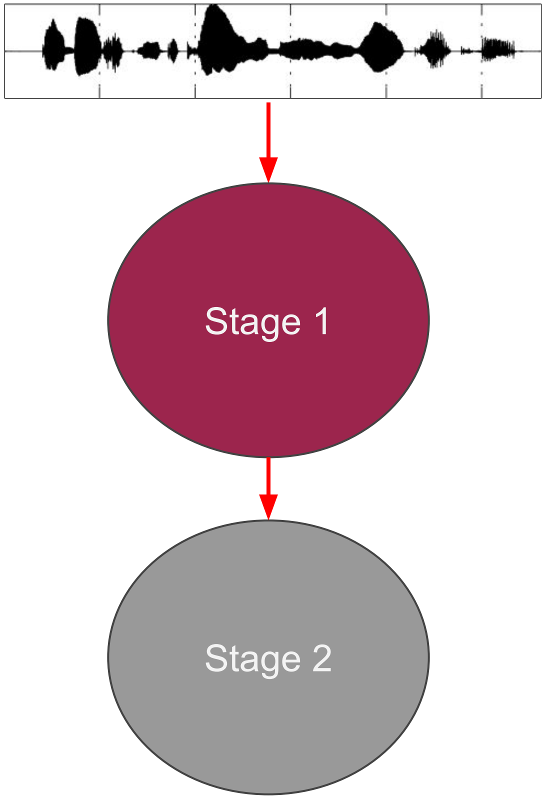
FOGD

Vowel region

Significance of using non-negative frequency-weighted energy operator

- The Motivation for the proposed vowel region detection is that, the levels of energy in speech signal is distributed across a range of frequencies and change with time.
- A signal processing tool to track the dynamic energy transitions can be used as cues to detect landmarks using frequency dependent non-negative energy operator.
- The non-negative, frequency-weighted energy operator serves as a tool, which produces instantaneous energy contour of the speech signal eliminating the block processing of the speech signal.
- It produces better time localization pertaining to the energy contour
- The sharp rise and fall of energies around GCIs can be visualized as VOPs and VEPs.

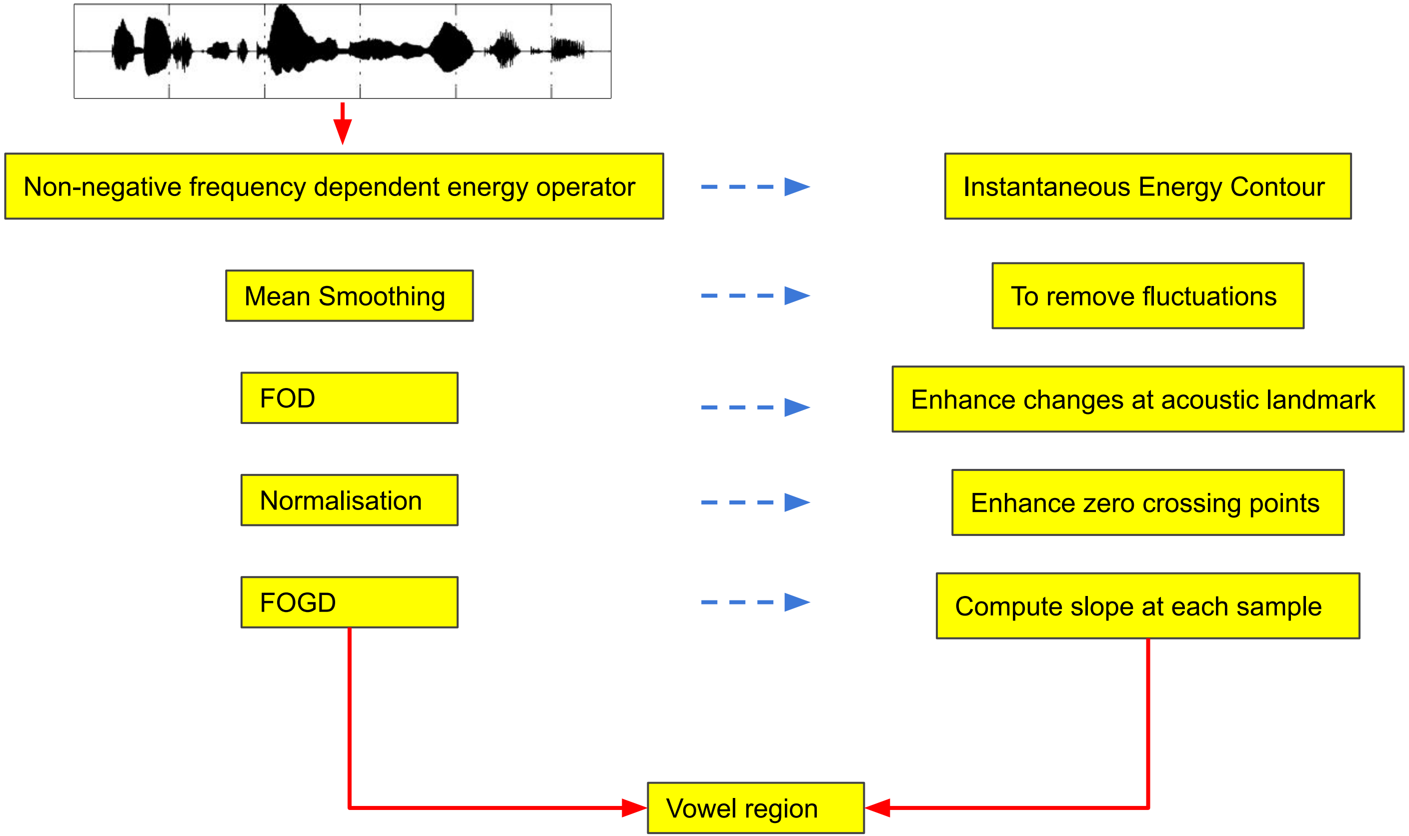
Proposed Method



Stage 1

Stage 2

Stage 1



Non-negative frequency dependent energy operator

Instantaneous Energy Contour

Mean Smoothing

To remove fluctuations

FOD

Enhance changes at acoustic landmark

Normalisation

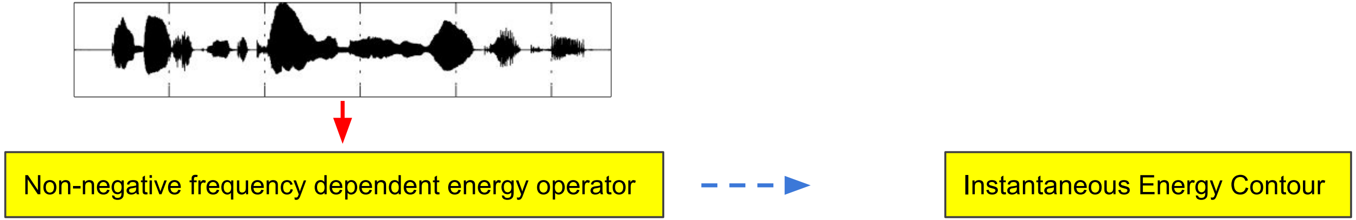
Enhance zero crossing points

FOGD

Compute slope at each sample

Vowel region

Proposed Method



Non-negative frequency dependent energy operator

Instantaneous Energy Contour

<u>Non-negative frequency dependent operator :</u>

$$Y[n] = \mathbf{\Gamma}[x[n]] = 0.25[x^2[n+1] + x^2[n-1] + h^2[n+1] + h^2[n-1] + 0.5[x[n+1]x[n-1] + h[n+1]h[n-1]]$$

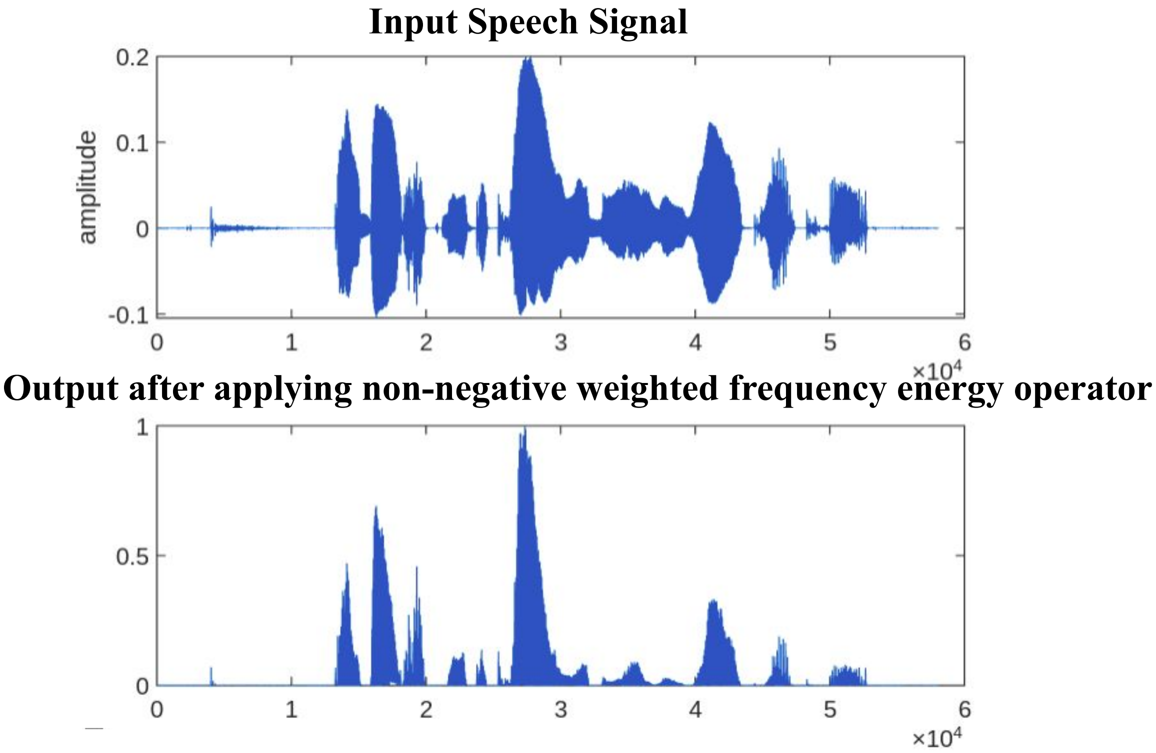
h[n] : Hilbert Transform of x[n]

This operator helps in capturing energy fluctuations of a speech signal because it provides good time resolution. Operator efficiently captures amplitude and frequency information of the signal.

Results

“don’t ask me to carry an oily rag like that”

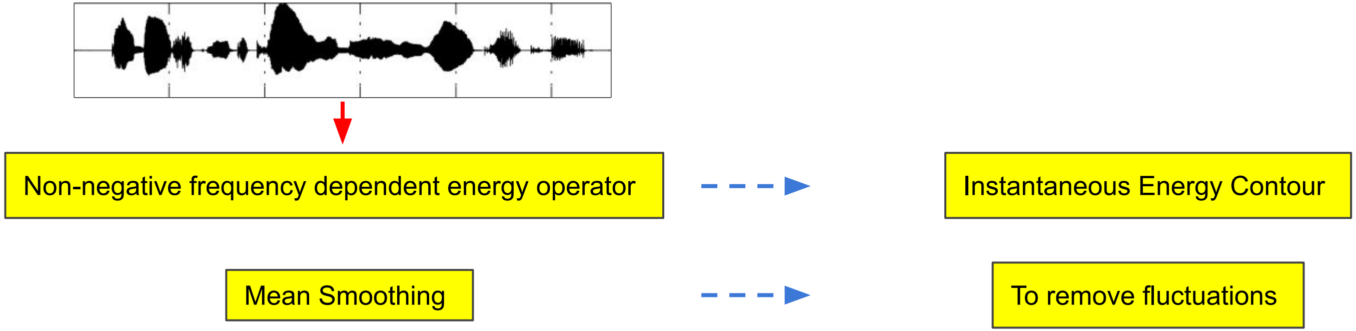
STEP 1:



Input Speech Signal

Output after applying non-negative weighted frequency energy operator

Proposed Method



Non-negative frequency dependent energy operator

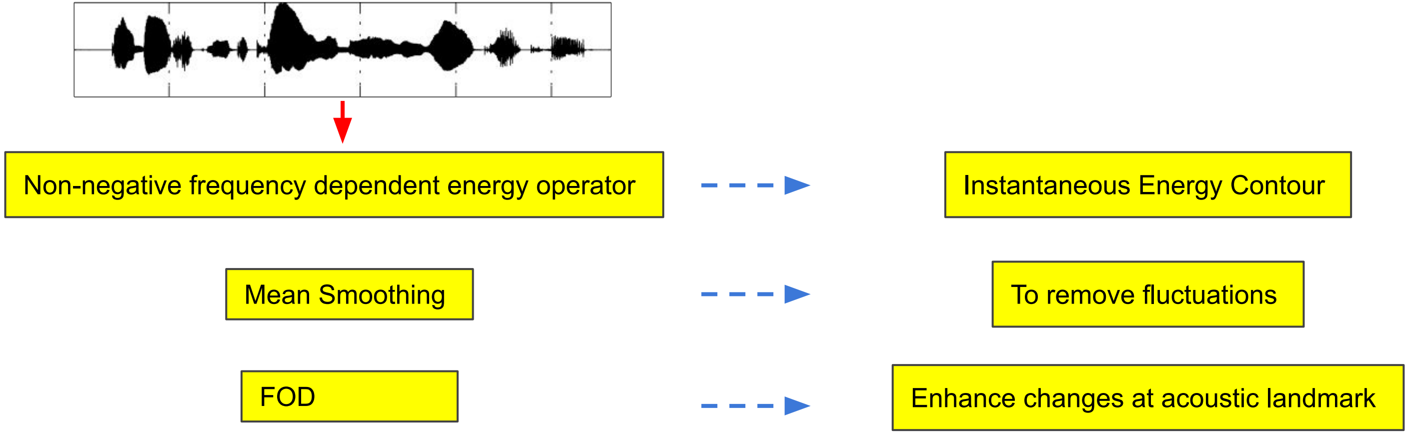
Instantaneous Energy Contour

Mean Smoothing

To remove fluctuations

Mean smoothing using 50 ms window.

Proposed Method



Non-negative frequency dependent energy operator → Instantaneous Energy Contour

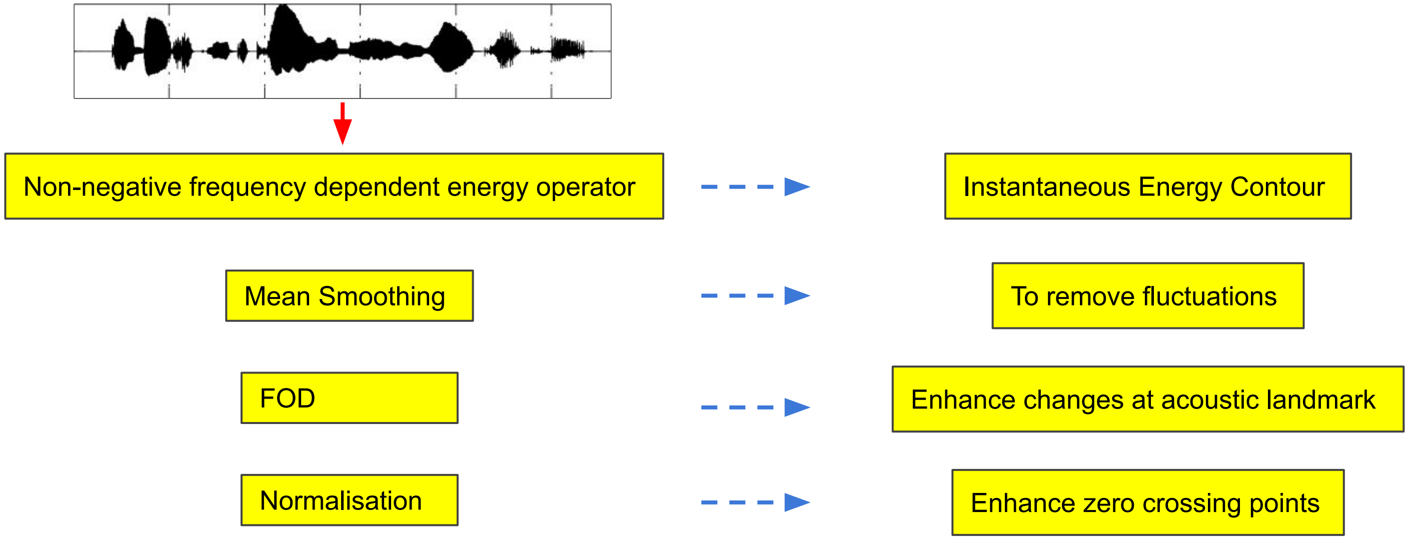
Mean Smoothing → To remove fluctuations

FOD → Enhance changes at acoustic landmark

Calculating slope using First Order Derivative:

$$y_d[n] = y'[n] - y'[n-1]$$

Proposed Method



Non-negative frequency dependent energy operator → Instantaneous Energy Contour

Mean Smoothing → To remove fluctuations

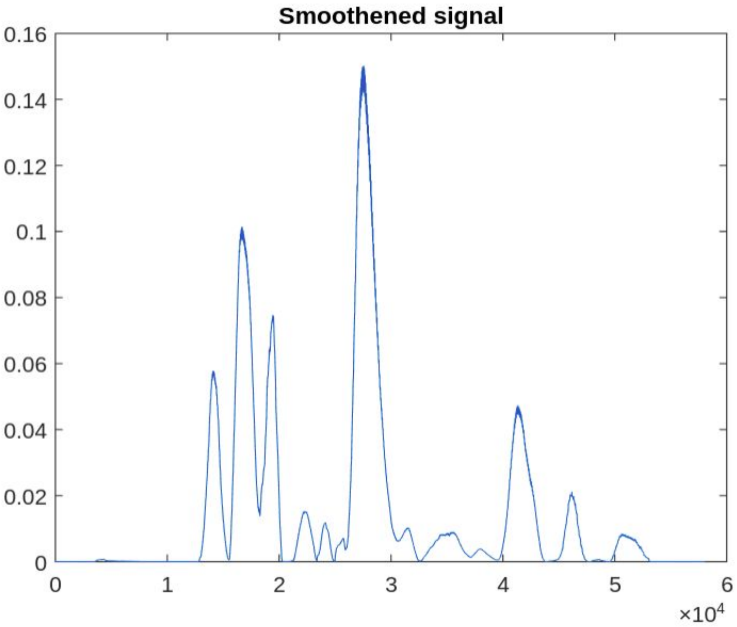
FOD → Enhance changes at acoustic landmark

Normalisation → Enhance zero crossing points

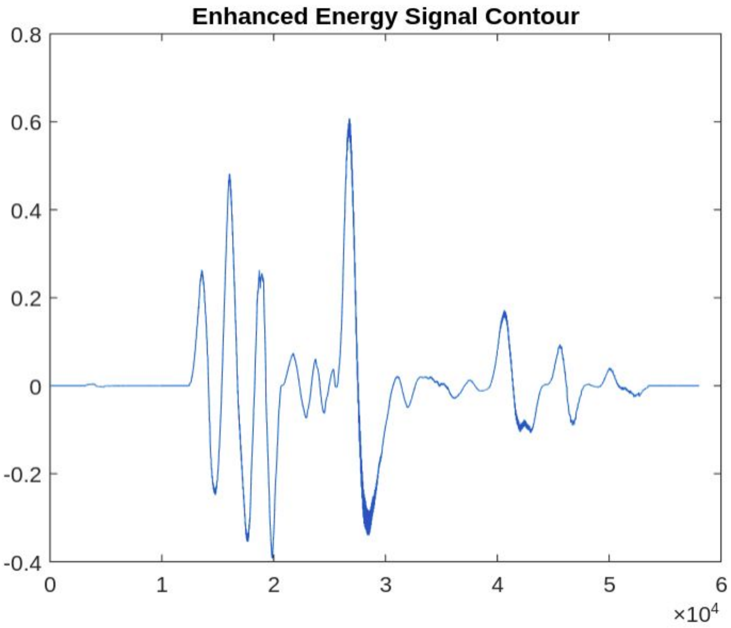
Normalizing:

$$y_N[n] = \frac{y_d[n] - \text{min}}{\text{max} - \text{min}}$$

Results

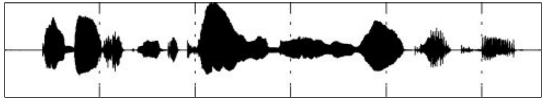


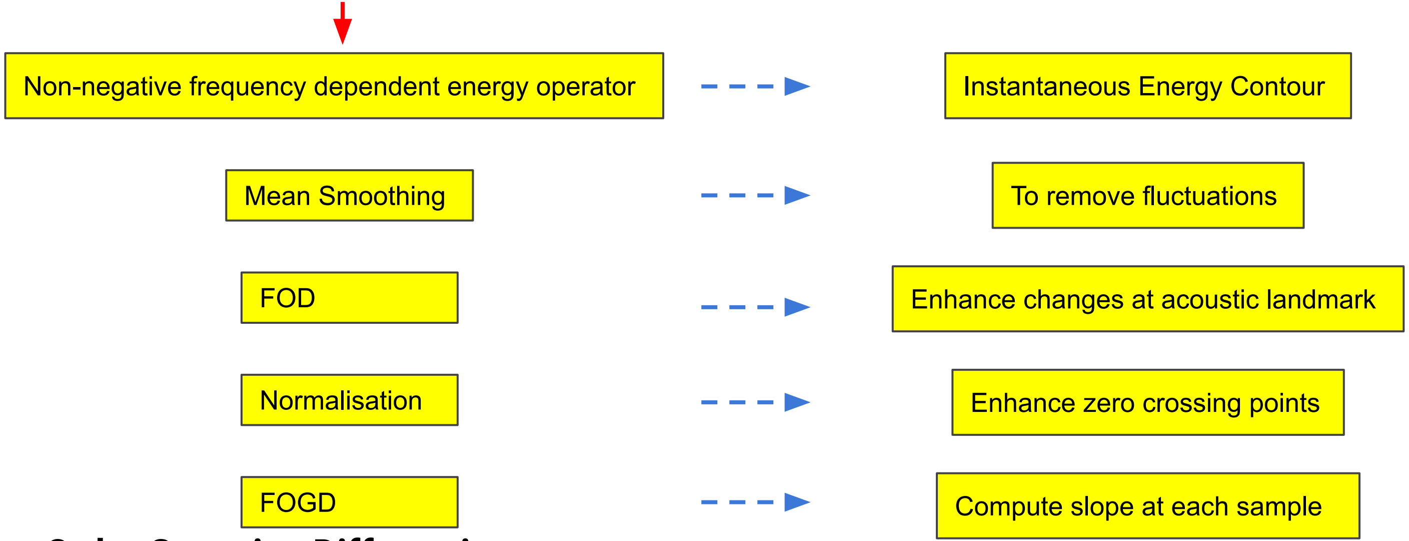
STEP 2



STEP 3

Proposed Method




Non-negative frequency dependent energy operator → Instantaneous Energy Contour

Mean Smoothing → To remove fluctuations

FOD → Enhance changes at acoustic landmark

Normalisation → Enhance zero crossing points

FOGD → Compute slope at each sample

First Order Gaussian Differentiator:

$$g[n] = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-}\frac{n^2}{2\sigma^2},\ n = 1,\ 2,\ 3\ldots L$$

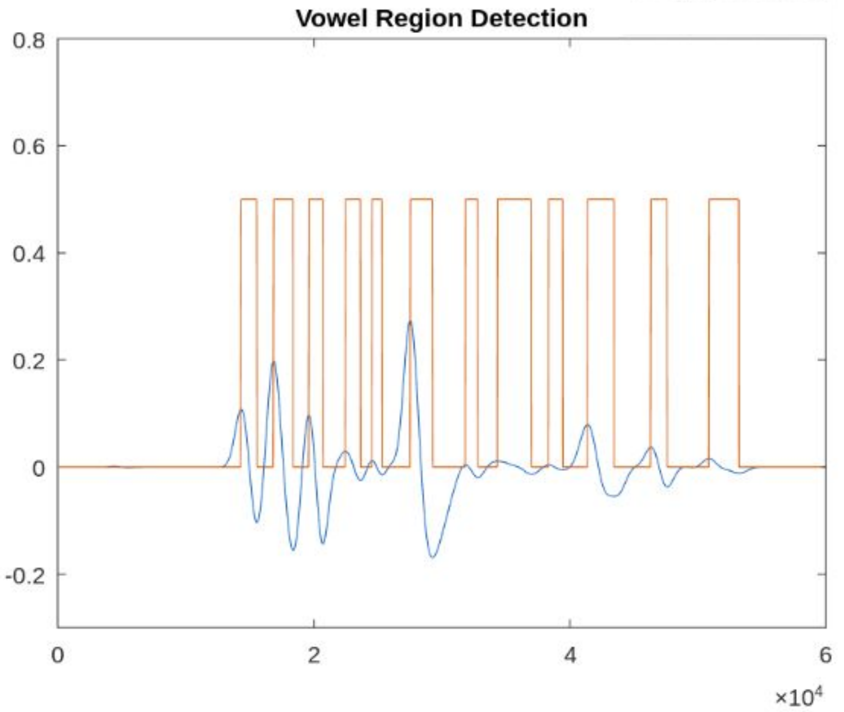
$$g_d[n] = g[n] - g[n-1]$$

A 100ms window with 25ms standard deviation

→Output zero for constant slope.
→Positive peak on on positive slope energy transition.
→Negative peak on negative slope of energy transition.

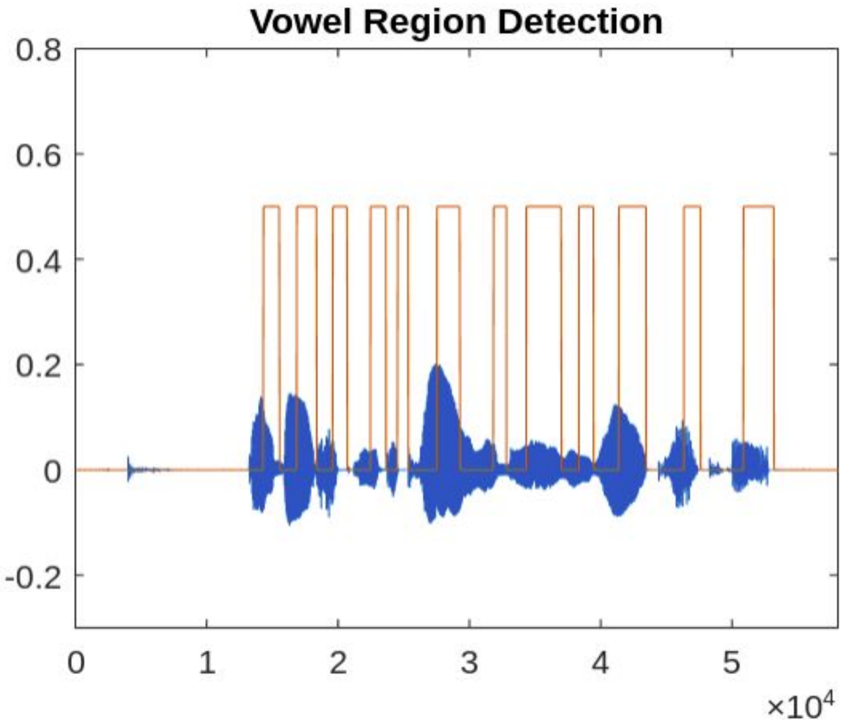
Results

STEP 4



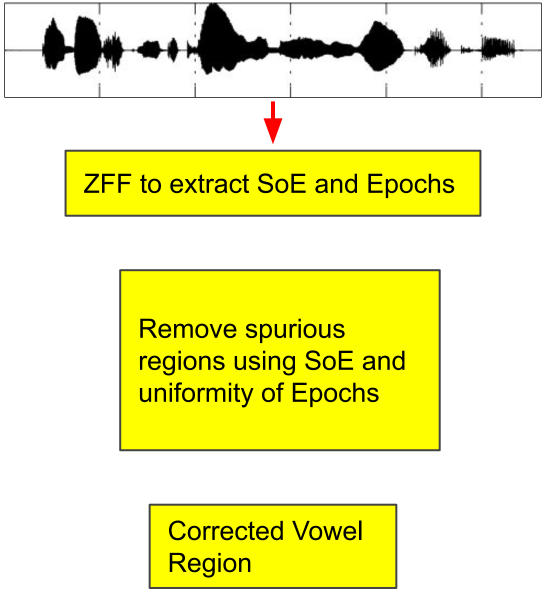
Output after applying First Order Gaussian Derivative and marking positive peaks as VOP and negative peaks as VEP.

Results



Highlighting vowel region on the input speech wave.

Stage 2



ZFF to extract SoE and Epochs

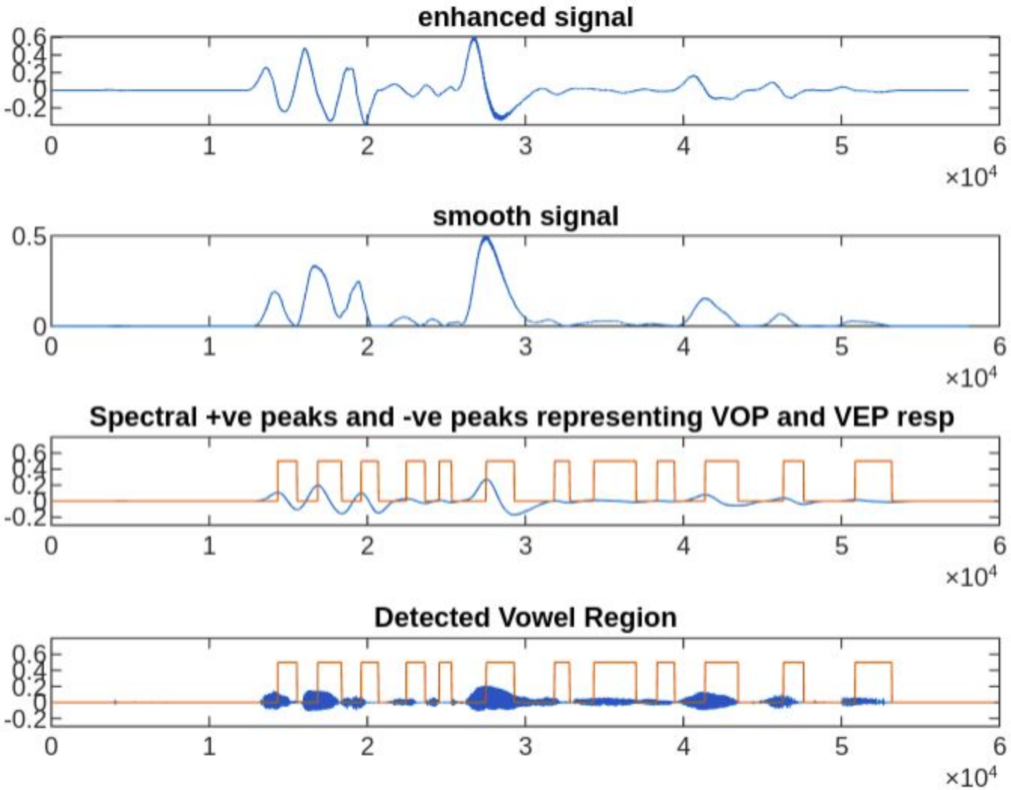
Remove spurious regions using SoE and uniformity of Epochs

Corrected Vowel Region

Glottal Closure Instants (GCIs): negative to positive zero crossings of ZFF signal

Strength of Excitation (SoE): The gradient of ZFF signal at each GCI.

Results



Combined outputs

Experiments

We tested our code on 5 speech samples out of which 4 were from TIMIT dataset and 1 we recorded manually.

Our code takes in mono channel speech input.

We have implemented the first stage which in only some cases detected spurious regions.

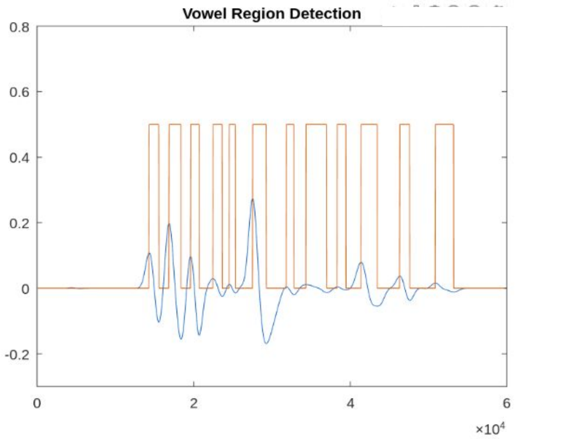
Results on 1st speech sample

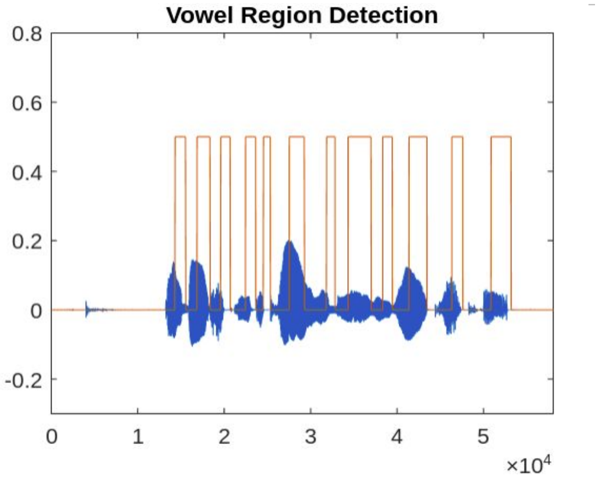
“don’t ask me to carry an oily rag like that”

Detection rate = 100%

Miss rate = 0%

False alarm = 1





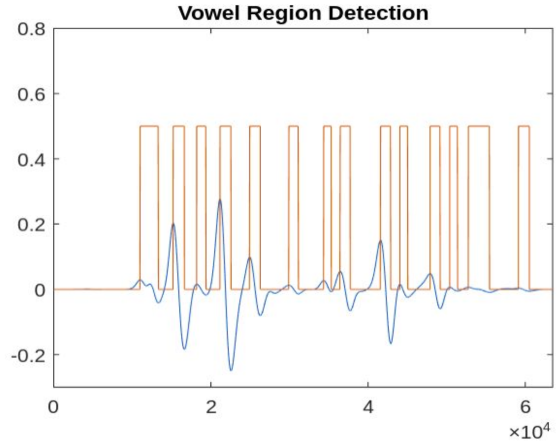
Results on 2nd speech sample

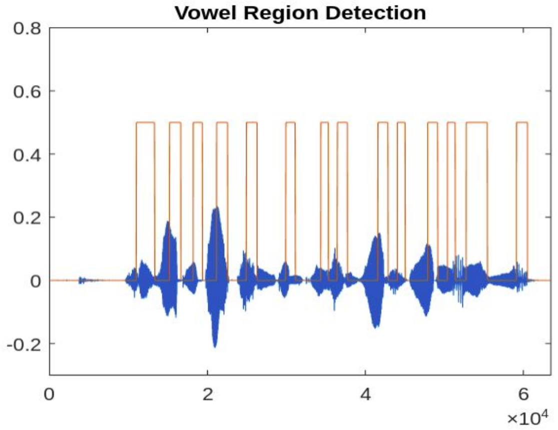
"She had your dark suit in greasy wash water all year"

Detection rate = 87.5%

Miss rate = 12.5%

False alarm = 1





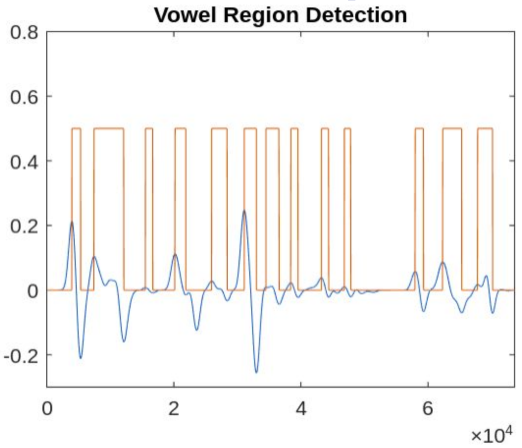
Results on 3rd speech sample

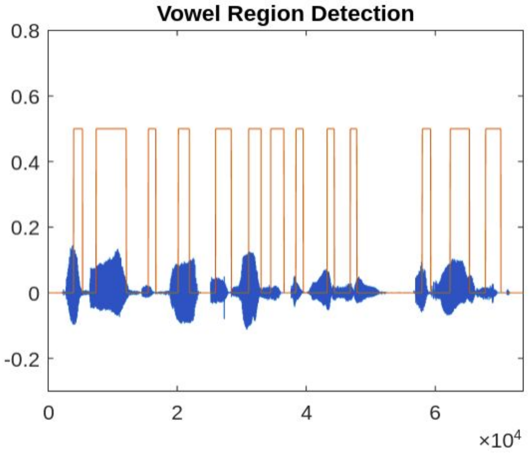
"Haha he thought a lush divorce at last"

Detection rate = 100%

Miss rate = 0%

False alarm = 1





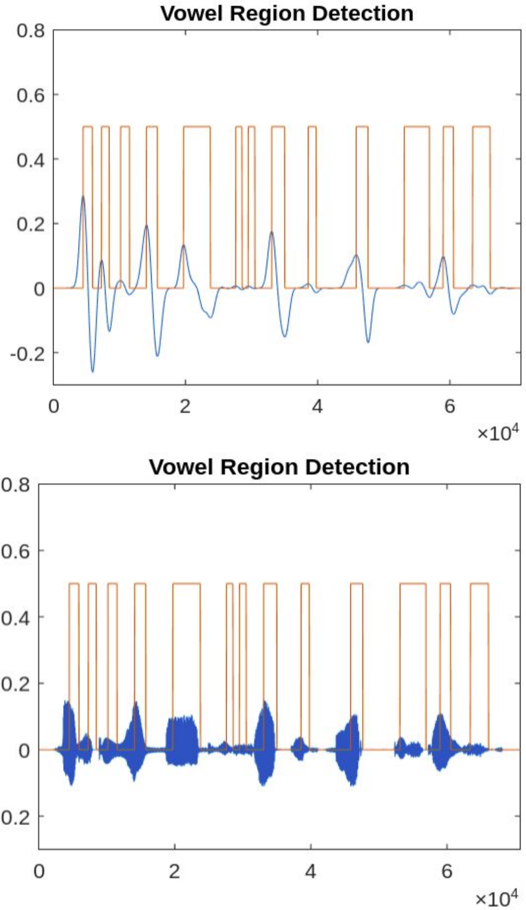
Results on 4th speech sample

"Husky young man, he said with mark distaste"

Detection rate = 100%

Miss rate = 0%

False alarm = 1



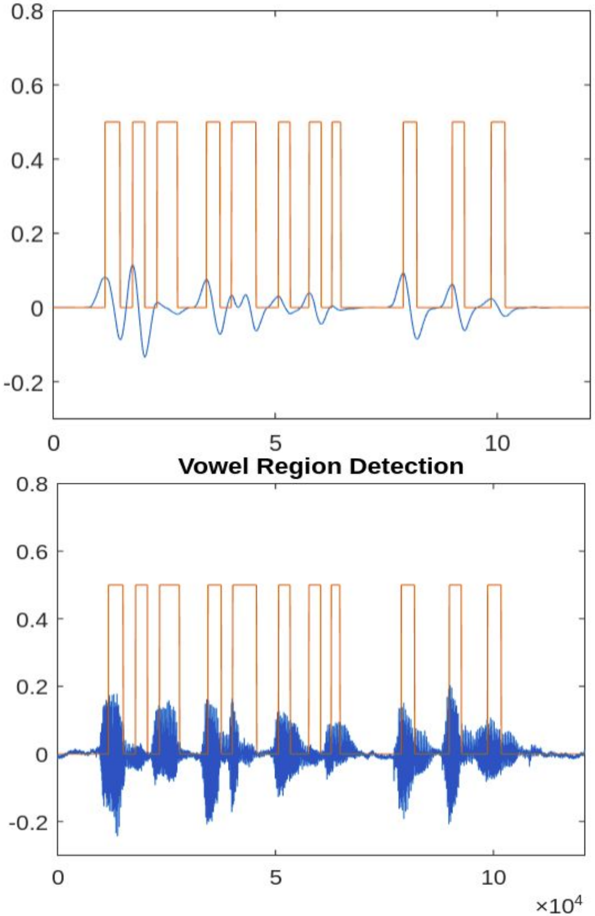
Results on 5th speech sample

“Personal predisposition tend to blunt”

Detection rate = 91.6%

Miss rate = 8.33%

False alarm = 0



Results Analysis

Average Detection rate = 95.82%

Average Miss rate = 4.166%

Average False alarm = 0.8

Thank you