Time Frequency Analysis

Final Project Report

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

---

Harsh Sharma : 20171157
Shashwat Shrivastava : 20161181

Contents:

Introduction

Vowels are primary units of the sound system of a language. These are produced by periodic impulse like excitation and possess high energy, periodicity and longer duration. Vowel region detection is a task of identifying vowel occurrences with precise boundary markings. These boundary markings are termed as vowel onset point (VOP) and vowel end-point (VEP). VOP is the time instant at which vowel region begins and VEP can be considered as the time instant at which vowel region ends in a continuous speech. The vowel regions detection is an important step in many speech processing applications. These include automatic speech recognition (ASR), speaker verification, smart audio filtering, recognition of CV units for emotion conversion, determining the duration of vowels in forensic applications, speech rate manipulation in speech synthesis, cochlear implants, and multimedia synchronization.

Some important points on Vowel:

- A speech signal is composed of voiced signal, unvoiced signal and noise. Vowels along with few consonants are 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.
- So whenever a vowel is produced there must be significant change in energy thus we now know to identify we will analyze the energy of the speech signal and process it to find Vowel Region.
- 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 takes

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

Generally, in method I, vowel region detection was carried using different acoustic modeling approaches using the combination of mel-frequency cepstral coefficients and excitation source features. Among these approaches, it is reported that the subspace GMM–HMM with discriminative training using boosted maximum mutual information produced superior performance. In method II, vowels are detected using a perceptually-enhanced spectrum matching. It explores a new model based on proposed components called matched filters. Matched filters are extracted by applying a series of perceptually-based processing operations to the speech spectra of the voiced frames. MFs are subjected to different factors leading to the variation in the speech spectra. An acoustic space representing two effective factors, namely phonetic context and speaker identity is modeled. Then, vowel and consonant MFs are conditioned to this context speaker acoustic space.

Method 1 : COMB

In this method, the combined evidence for the detection of boundary markings of a vowel region is derived from the three shreds of evidence derived from the spectral peaks, modulation spectrum and excitation source.

VOP Detection using Spectral Peaks :

The speech signal is processed in a block of 20ms with a shift of 10ms. For each block of 20ms, 256-point DFT is computed, and ten largest peaks from first 128 points are selected. The sum of these amplitudes is plotted as a function of time which represents the energy of spectral peaks. The onset of vowel is observed as significant change in sum of ten peaks in the DFT spectrum. Further, enhancing is

performed by convolving it with FOGD operator and the obtained output is the VOP evidence plot using spectral peaks. The peaks in the VOP evidence plot indicate the possible VOP locations .

VOP Detection using Modulation spectrum energy:

The speech signal is passed through approximately 18 trapezoidal critical bandpass filters between 0 and 4 kHz. An amplitude envelope of the signal is computed using half wave rectification and low pass filtering on all bands. Amplitude envelope signals are down-sampled to 80 and normalized by the average envelope of that channel, measured over the entire utterance. The modulations of the normalized envelope signals are analyzed by computing DFT over 250 ms with an overlap of 5% in order to capture dynamic properties of the signal. The 4–16 Hz components are added together across all critical bands to derive modulation spectrum energy. Thus obtained signal is enhanced and processed to obtain third evidence for detecting VOPs and VEPs.

VOP Detection using Excitation Source:

Speech is produced by the excitation of the time-varying vocal tract system with a quasi-periodic signal. The time-varying excitation information is context-dependent in terms of voicing, level of voiced energy, and associated periodicity. Linear prediction (LP) residual corresponds to the excitation source information useful in voice analysis of a speech signal. It is extracted using LP analysis.

The time-varying dynamics in the excitation characteristics are overspread in the LP residual due to its bipolar nature. So, Hilbert envelope of LP residual is estimated, which is unipolar. The smoothened Hilbert envelope of the LP residual is obtained by convolving with a Hamming window of 50 ms.

This evidence is considered for the VOP and VEP detection and enhanced using first-order difference operator. These acoustic events are detected based on the nature of the gradient of the output signal

Method 2 : FGCI

The time instants at which glottal signals produce high energy during the production of voiced speech are referred as GCIs. Spectrum estimation during glottal closure phase will be more accurate as true vocal tract resonances are present during this period. The spectral energy computed around the GCIs has been used as an evidence to detect VOP and VEP.

Firstly, the GCIs are detected using zero frequency filtering (ZFF) technique (Yegnanarayana and Murty 2009). Around these GCIs, formants are computed for 30% of speech samples using group delay function. This formant energy of speech signal is computed as the sum of first three formants, and it is plotted as a function of time. This formant energy contour is smoothed using a mean smoothing window of 50 ms and enhanced using a first order difference operator.

Significant changes in the spectral characteristics, present in the enhanced signal are detected by convolving the same with first-order Gaussian differentiator operator having 100 ms length and 25 ms standard deviation. After eliminating the spurious peaks, positive and negative peaks of this signal represent locations of VOP and VEP respectively.

Proposed Method

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 changes 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 operators. 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.

This vowel region detection method has been implemented in two stages. In the first stage, onset and end-points of the vowel are detected using the instantaneous energy

contour of the speech signal. In the second stage, the positions of VOPs and VEPs have been corrected along with the removal of spurious vowel regions. This is carried out based on the uniformity of epochs and the SoE profile.

These regions can be considered as linguistically relevant information possessing regions that can be used in the front-end of the automatic speaker recognition system.

Mathematics Involved

Stage 1: VOP and VEP detection using non-negative frequency dependent energy measure

As a part of the first step in vowel region detection, the VOPs and VEPs are detected from the continuous speech signal in the following manner: an envelope of the derivative of the signal, which is non-negative frequency dependent energy operator is applied to the speech signal x[n], to produce instantaneous energy contour. It is computed using the below equation:

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

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

This energy operator is nearly instantaneous in discrete time as energy computation of a speech signal is done with three samples at each time instant. It provides good time resolution to capture energy fluctuations of a speech signal within a glottal cycle. The fluctuations produced in the energy contour are smoothed by using mean smoothing with 50 ms window.

$y'[n]$= Mean smoothed instantaneous energy contour

Now after smoothning, we enhance the change at the acoustic landmarks present in the smoothed instantaneous energy contour of the speech signal by computing it's slope using the first order difference of the resulting signal is given by:

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

The regions associated with zero crossing points from both positive to negative and negative to positive are enhanced by normalization process given by:

$$y_N[n] = (y_d[n] - \min(y_d))/(\max(y_d) - \min(y_d))$$

The predominant energy changes present in the enhanced instantaneous energy contour associated with the vowel landmarks are detected by convolving with the first order Gaussian differentiator operator of 100 ms. Here we use the gaussian window g[n] of length L:

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

The first order Gaussian window is termed as $g_d[n]$ and given by:

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

In our case we have taken a 100ms window with 25ms standard deviation which outputs zero for constant slope, positive peak on positive slope energy transition and negative peak on negative slope of energy transition.

Stage 2: Post processing of VOP and VEP locations using uniformity of the epochs and strength of the excitation

The resulting prediction is further improved in the second stage by removing spurious vowel regions and correcting the positions of VOP and VEP locations using the uniformity of epoch intervals and the SoE of the speech. The energy changes in the speech signal during the vowel production is reflected in the excitation source information. Therefore, these changes are characterized by the SoE.

The SoE and uniformity of the epochs are computed from the zero frequency filtered (ZFF) signal, as it highlights the high information in lower frequency bands. So in stage 2 we follow this step :

Consider a speech signal s[n] and perform high frequency boosting as it is noted that higher frequencies are more important for signal disambiguation than lower frequencies.

$$x[n] = s[n] - s[n-1]$$

The speech signal is fed to a resonator centered at 0 Hz. The resonator is realized using the following transfer function. The output of cascade of two ideal second order digital resonators at zero frequency is computed as:

$$y[n] = \sum_{k=1}^{4} \alpha_k y[n-k] + x[n]$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$ and $a_4 = -1$.

The progression can be removed from the output signal using progression removal operation, which involves subtracting the local mean of the original signal at every instant of time. This is represented using the following expression:

$$\hat{y}[n] = y[n] - \tilde{y}[n]$$
$$\text{where } \tilde{y}[n] = 1/2N + 1 \sum_{n=-N}^{N} y[n].$$

Here 2N+1 is the size of the window used for computing local mean, which is typically average pitch period. The resulting output signal is called ZFF signal. The negative to positive zero crossings of ZFF signals correspond to GCIs. The gradient of ZFF signal at each GCI is termed as SoE.

The spurious vowel regions are removed based on the uniformity of the epochs and the SoE. The positions of VOPs and VEPs are corrected based on combined cues from the SoE and uniformity of epoch intervals. The SoE exhibits positive trend from a local minimum at VOP and a negative trend from a local minimum at VEP respectively. The uniformity transition points on the pitch contour also corresponds to the vowel boundaries. Therefore, the SoE contour and uniformity of epoch intervals can be used as an evidence for correcting the positions of VOPs and VEPs.

NOTE: We have implemented only till part 1 as instructed by sir, but still we are getting very good results, or in other words, after stage 1 itself we are getting very good results. As shown:

Output Plots

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

Sample 1 : *"don't ask me to carry an oily rag like that"*

We will show the output after every step for this input signal.
Step 1: Finding the Energy contour of the input speech signal.

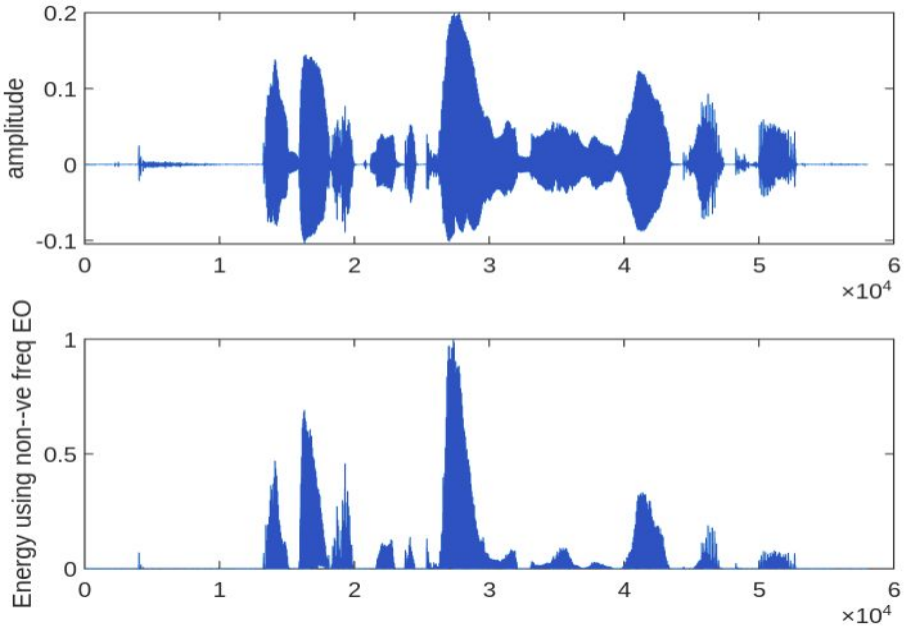


Figure 1 is the Input speech signal and Figure 2 is output after applying non-negative weighted frequency energy operator

Step2: Smoothing the energy contour with mean smoothing of 50ms window.

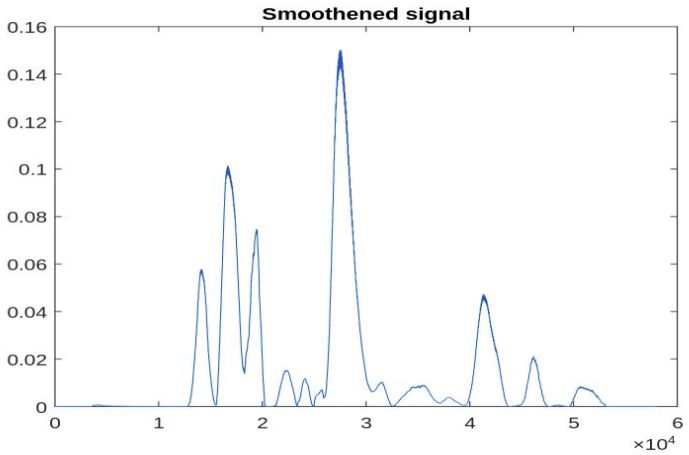


Fig: Smoothed signal

Step 3: Enhancing the smoothed signal using FOD operator.

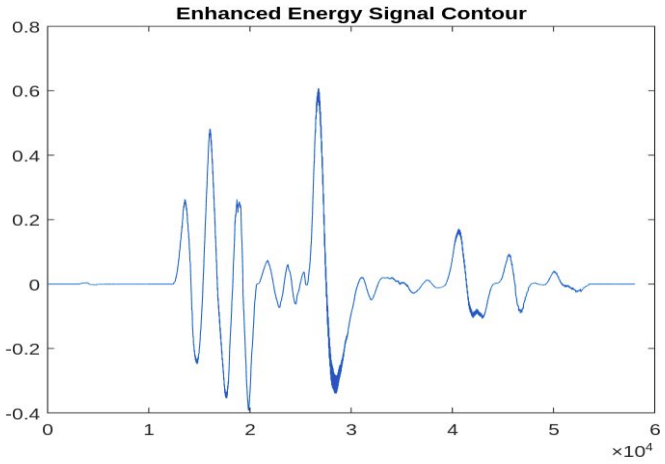


Fig: Enhanced energy signal contour

Step 4 & 5 : Finding spectral peaks using FOGD operator and after that marking vowel region.

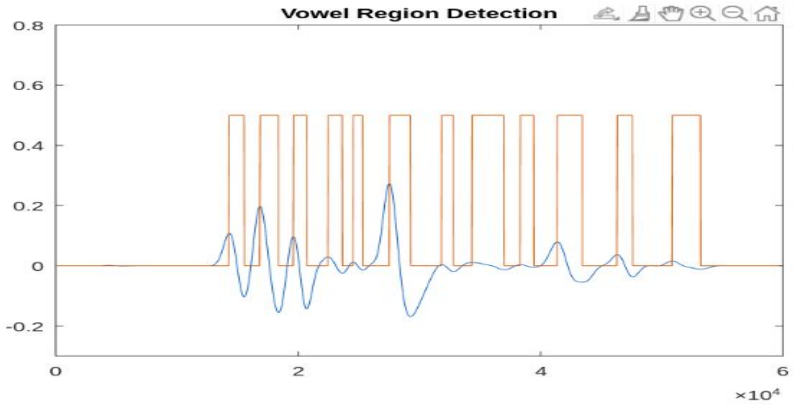
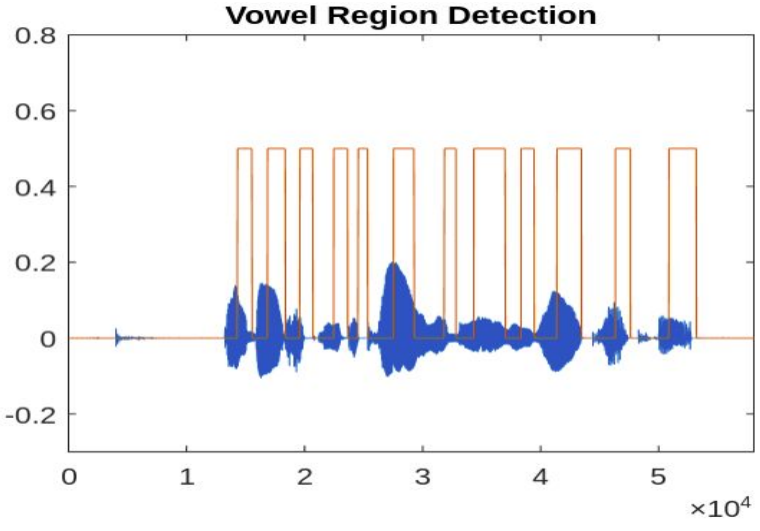
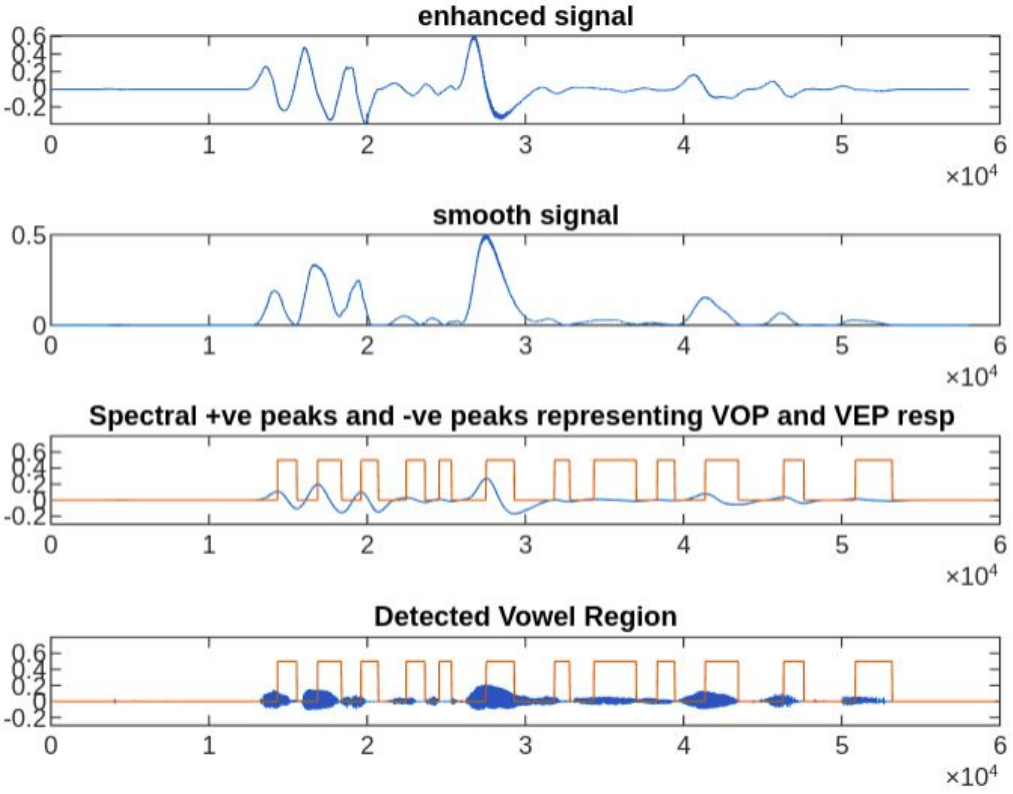


Fig : Output after applying First Order Gaussian Derivative and marking positive peaks as VOP and negative peaks as VEP.

Final Result :



Combined plot:



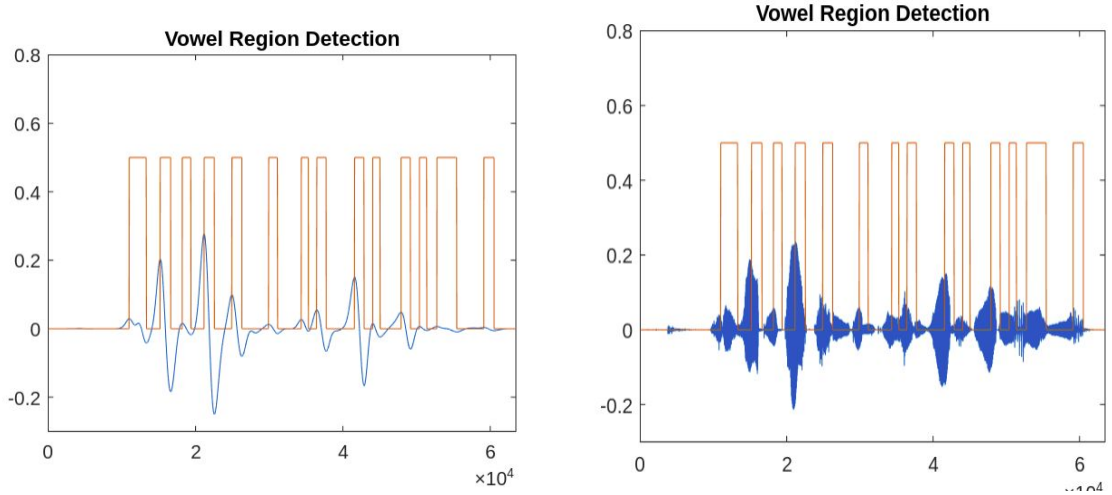
We observe in this input speech signal : Detection rate = 100%, Miss rate = 0% and False alarm = 1.

NOTE : For rest of the samples we have:

GRAPHS ON LEFT: ***Represents vowel region marked on output of FOGD output. Positive peaks represent VOP and Negative peaks represent VEP.***

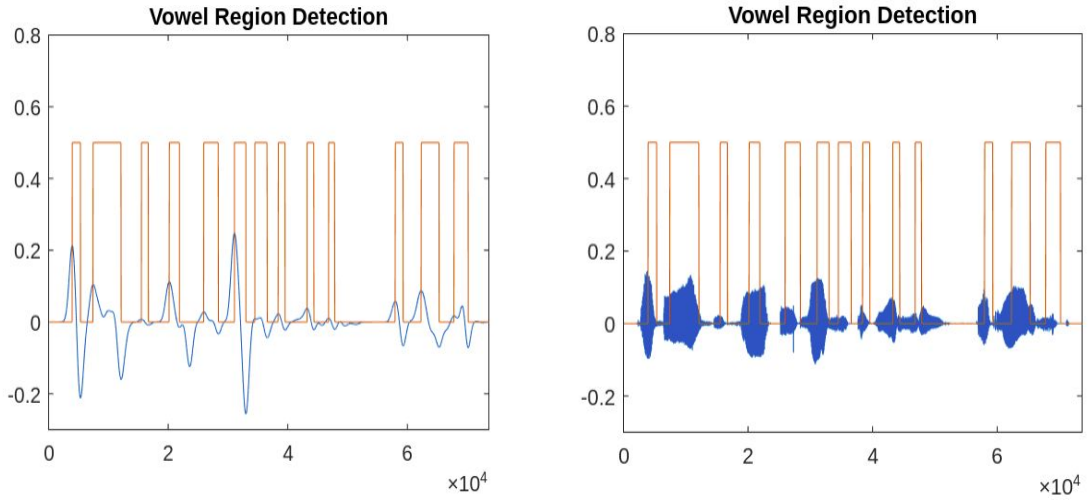
GRAPHS ON RIGHT: ***Represents vowel region marked on Input Speech Signal.***

Sample 2: *"She had your dark suit in greasy wash water all year".*



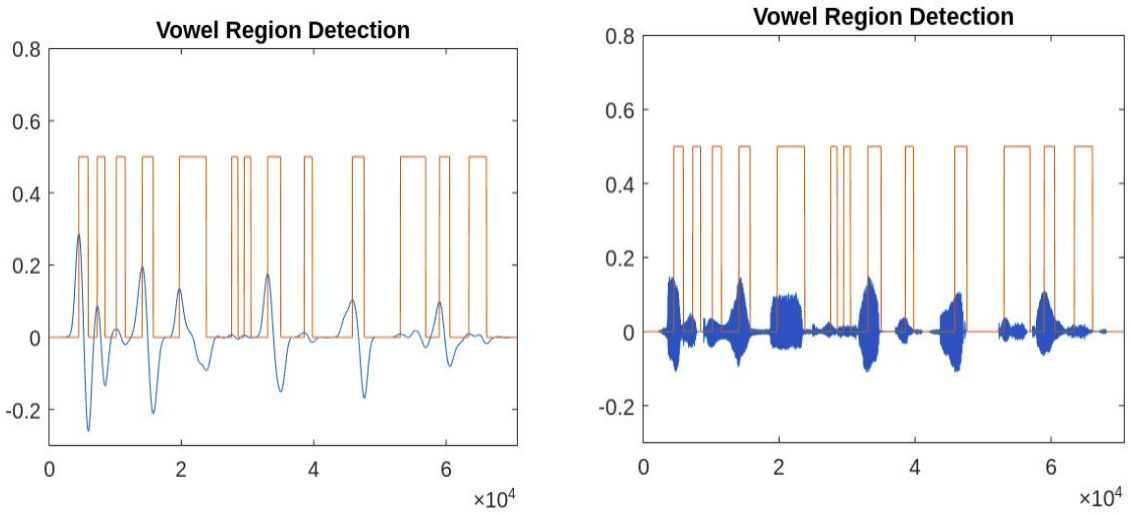
We observe in this input speech signal : Detection rate = 87.5%, Miss rate = 12.5% and False alarm = 1.

Sample 3: *"Haha he thought a lush divorce at last"*



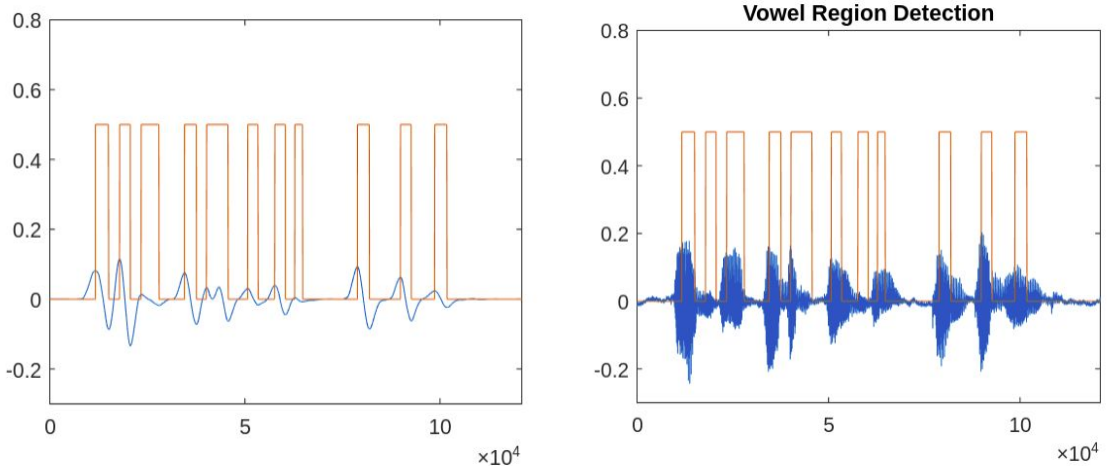
We observe in this input speech signal : Detection rate = 100%, Miss rate = 0% and False alarm = 1.

Sample 4: *"Husky young man, he said with mark distaste"*



We observe in this input speech signal : Detection rate = 100%, Miss rate = 0% and False alarm = 1.

Sample 5: *"Personal predisposition tend to blunt"*



We observe in this input speech signal : Detection rate = 91.6%, Miss rate = 8.33% and False alarm = 0.

Result Analysis

Average Detection rate = 95.82%, Average Miss rate = 4.166% and Average False alarm = 0.8.

Contribution

We have collaborated over google slides and docs to make presentation and report. We used Microsoft Teams to collaborate to code on matlab. At the end, together we made a video presentation.

Apart from this we both read papers completely and used to fix code individually as and when required after completing the initial part of coding.

Reference

Application of non-negative frequency-weighted energy operator for vowel region detection