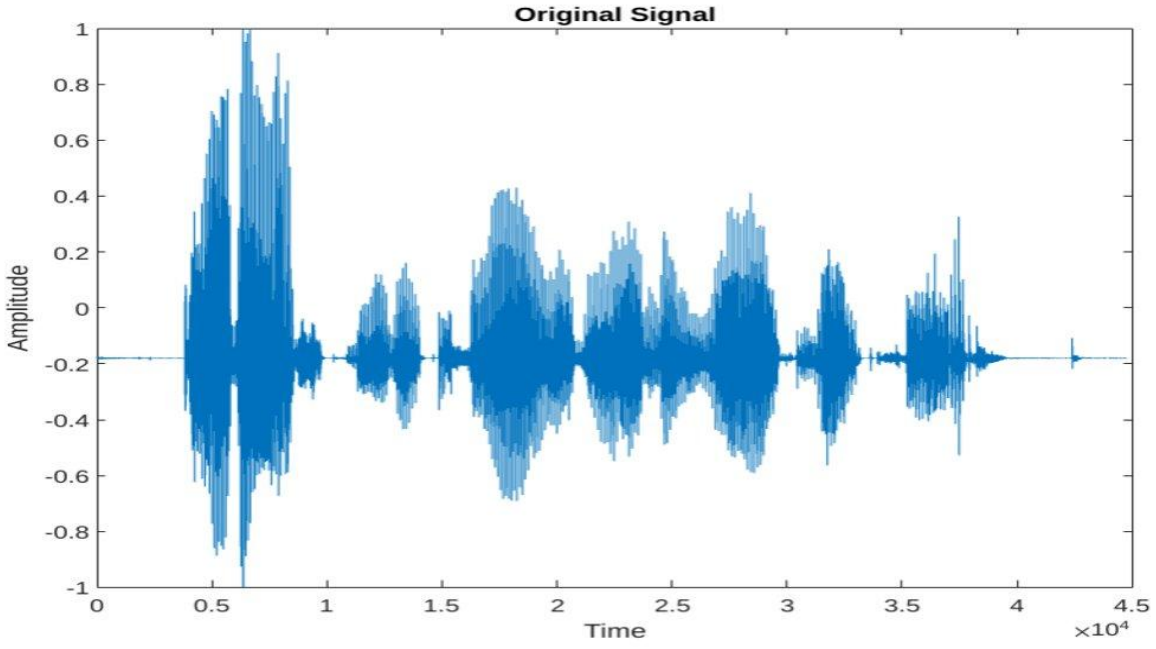
Speech Signal Processing
Project
Comparision of Vowel Detection methods

Laksh Balani (2020102019)
Yash Bhatia (2020101007)
Ishanya Sethi (2020102014)

Introduction

Vowels are long duration, periodic and high energy sound units in a speech utterance Vowel Onset Points (VOPs) are the instants of starting of a vowel, while Vowel End Point (VEPs) are the instants of ending of a vowel. Both of these are equally important for accurate detection of vowels. The frequency response of the vocal-tract system as well as the excitation source information are better manifested within the vowels A vowel sound unit may be viewed as consisting of three regions viz. onset, steady and offset regions. The change in signal energy and periodicity at the VOPs is quite sharp and comparatively higher. Signal magnitude maintains a nearly constant value within the steady regions. The signal decays slowly around VEPs. Accurate detection of vowels in a speech is important as it helps us in time scale modification due to the fact that the change in duration in these regions is comparatively more as compared to consonants and noisy regions. Vowel detection can help us extract the prosodic features of speech signal, and also help us estimate the syllables and speaking rate in the signal.

File: Dont_ask_me_to_carry_an_oily_rag_like_that.wav

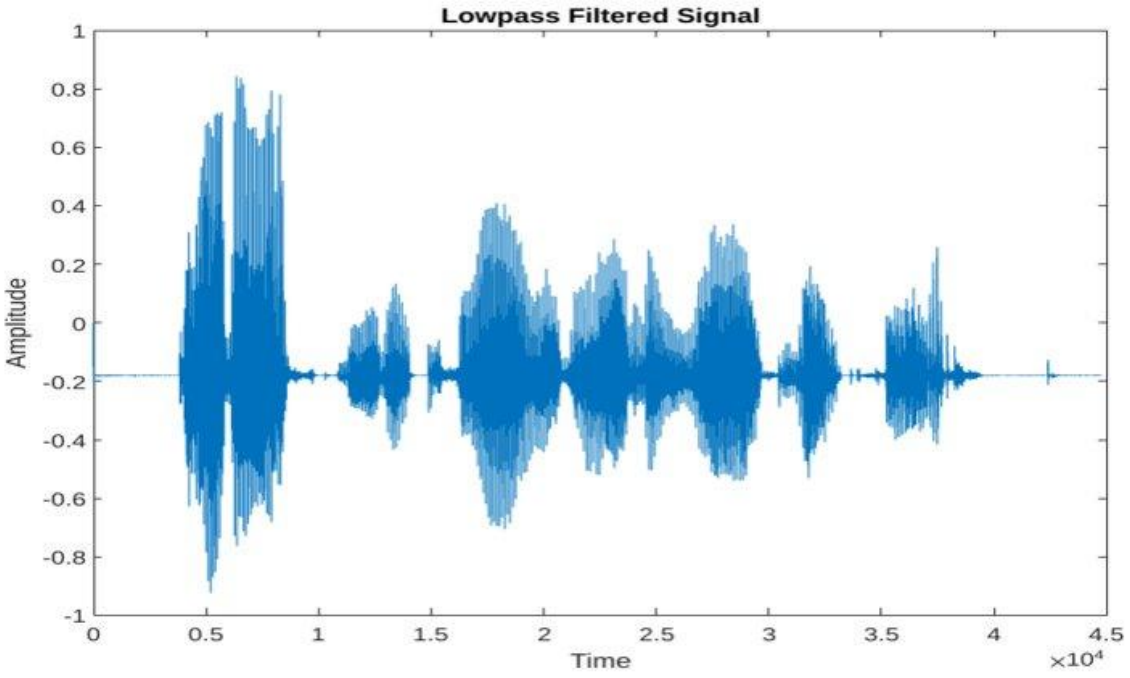


Detection of vowels using Average Magnitude Dynamics

When compared to other non vowel, quiet, and noise regions, the vowels have a much greater mean and variance of the magnitude dynamics over an analysis frame. The front-end characteristic in this study is the average magnitude dynamics (AMD) over an analysis frame. After that, the sigmoidal function is used to nonlinearly map (NL-AMD) the AMD values at each instant, sharpening the transitions at the VEPs and suppressing changes in the higher magnitude regions. Both the VOPs and the VEPs are equally discriminatory under the NL-AMD. As a result, the majority of VOPs and VEPs are discovered within a smaller variation. This method will work for both - clean as well as noisy speech.

The steps involved are as follows:

Step 1: The speech signal is passed through a low pass filter with a range of 0-2500 Hz. This is done because, for most of the vowels, the signal energy is predominantly concentrated in this range.



Step 2: After this, we construct the analytical signal of the above signal. The analytical signal is defined as:

$$x_a(n) = x(n) + jx_h(n)$$

Here, xh(n) is the hilbert transform of x(n). The Hilbert transform phase shifts negative frequency components by π/2 and positive frequency components by −π/2. Mathematically,

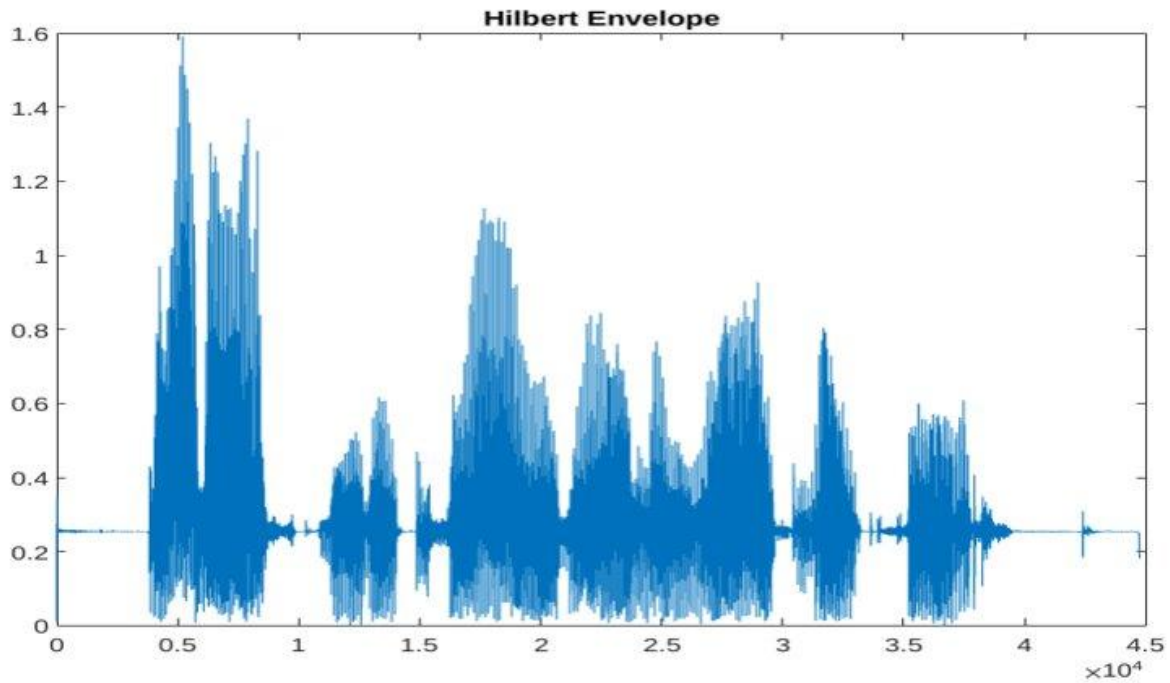
$$x_h(n) = IDFT(X_H(\omega))$$

where

$$X_H(\omega) = \begin{cases} +jX(\omega), -\pi \leq \omega < 0 \\ -jX(\omega), 0 \leq \omega \leq \pi \end{cases}$$

The magnitude of the analytical signal is referred to as the hilbert envelope. The hilbert envelope enhances the time varying nature of the filtered signal.

$$|x_a(n)| = \sqrt{x^2(n) + x_h^2(n)}$$



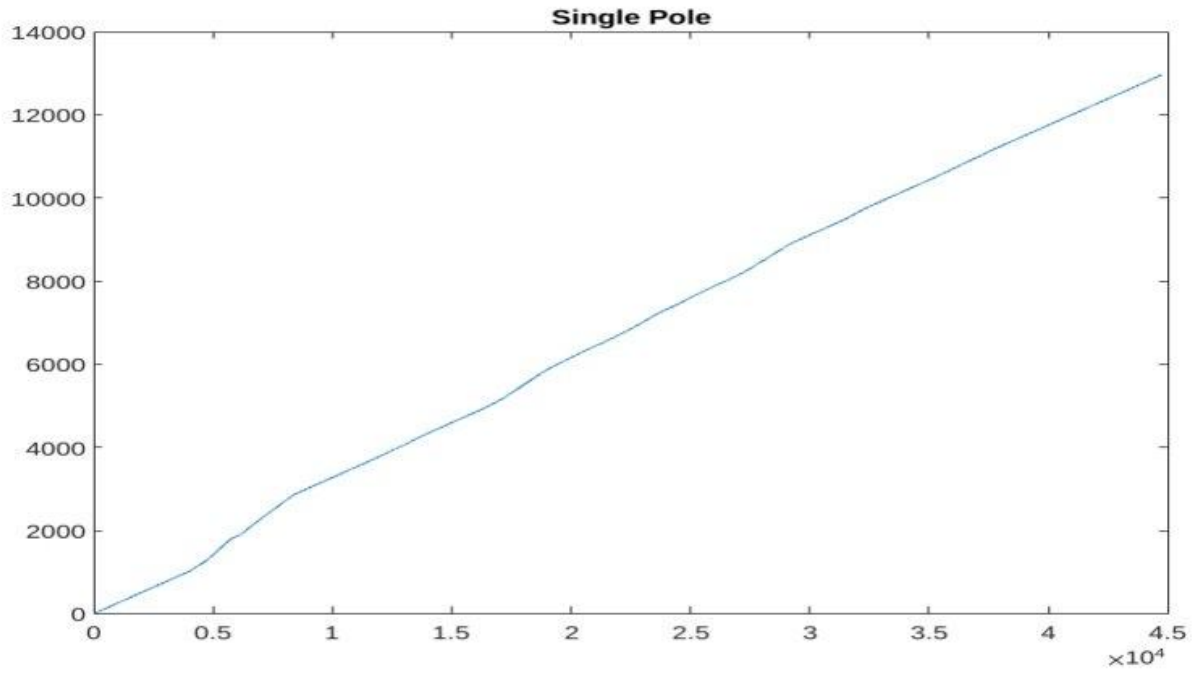
Step 3: After this, we pass it through a single pole filter , whose transfer function is given as:

$$H(z) = \frac{1}{1 - z^{-1}}$$

This is an infinite impulse response filter, which helps us enhance the stability of the hilbert envelope, and improves the ringing. Replacing H(z) by Y(z)/X(z), and then converting to time domain using the properties of z-transform yields:

$$y(n) = y(n-1) + |x_a(n)|$$

From the above equation, we realize that y(n) will be an increasing function of time (this also helps us in validating the plot we obtained for this). The stability of the filter is ensured by resetting the filter output to zero after a period of time.

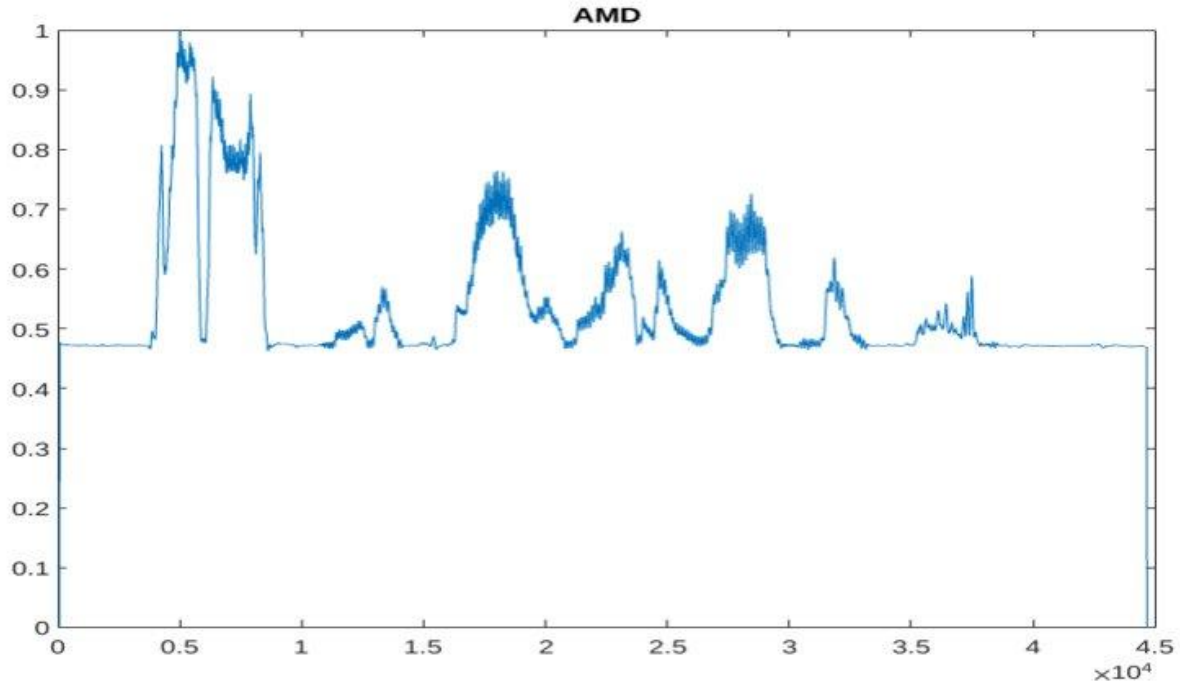


Step 4: Next, for each sample shift, we will analyze it over 2L + 1 sample points, viz : the L succeeding ones subtracted with the L preceding ones.

$$\hat{y}(n) = \frac{1}{L} \sum_{k=1}^{L} | y(n+k) - y(n+k-L-1) |$$

We choose L = 80 in this case.

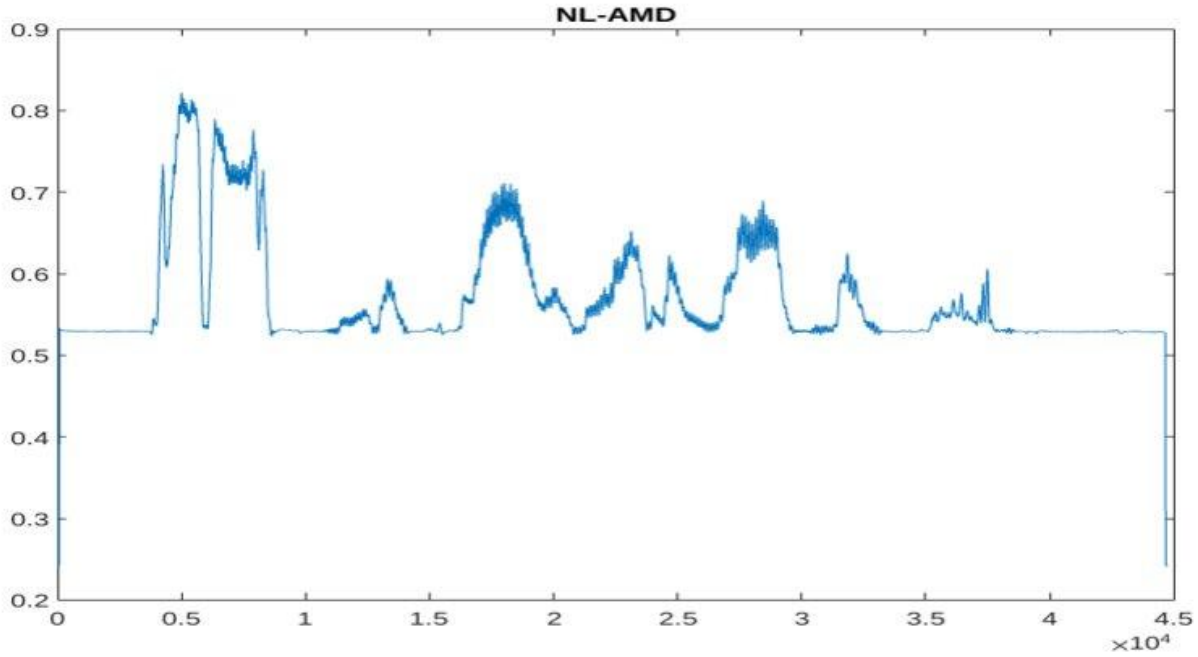
In the VOP/VEP areas, as opposed to other sound units, the AMD values will be larger due to the higher magnitude of the vowels. On the other hand, AMD will remain practically constant if the signal magnitudes are tiny or repeating (i.e., periodic) throughout time. As a result, only the VOPs and VEPs will experience substantial shifts in AMD.



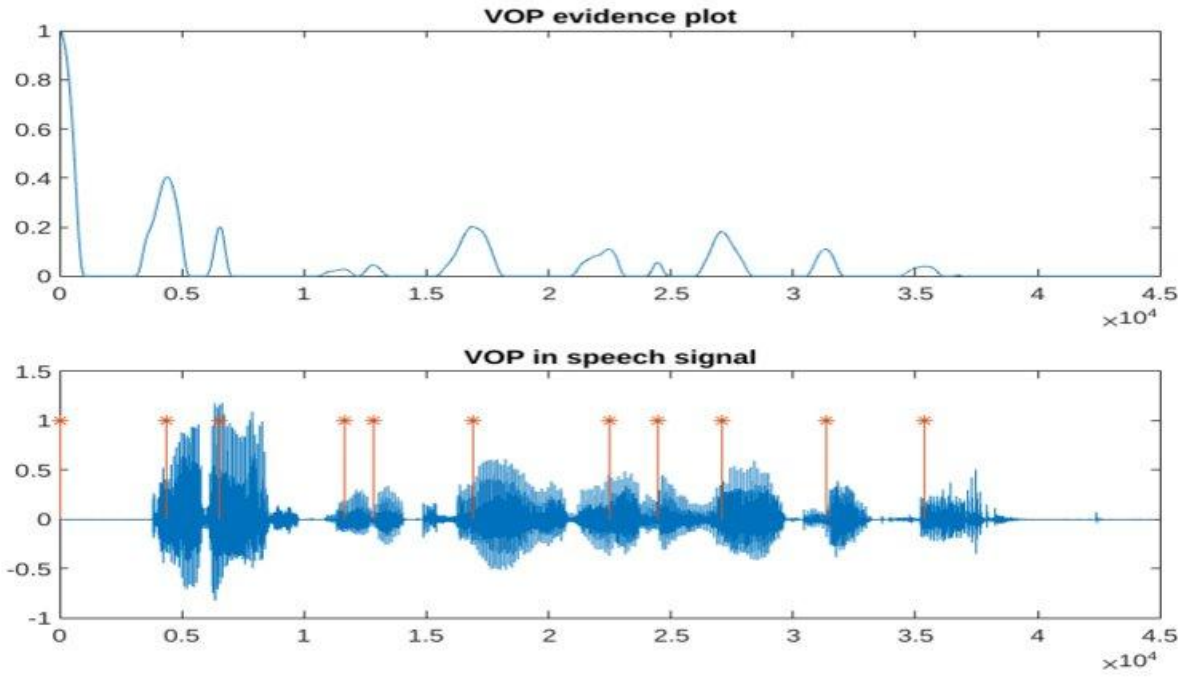
Step 5: We then non linearly map the AMDs at each time instants obtained above using a sigmoidal function, which helps us sharpen the transitions at the onset and offset regions, and suppress the variations at the higher magnitude regions.
The non linear mapping function is given by:

$$z_s(n) = \frac{1}{1 + \exp\left\{-\beta(\hat{y}(n) - t_h)\right\}}$$

Here, we take beta and th as 5 and 0.8 times mean value of $\hat{y}(n)$



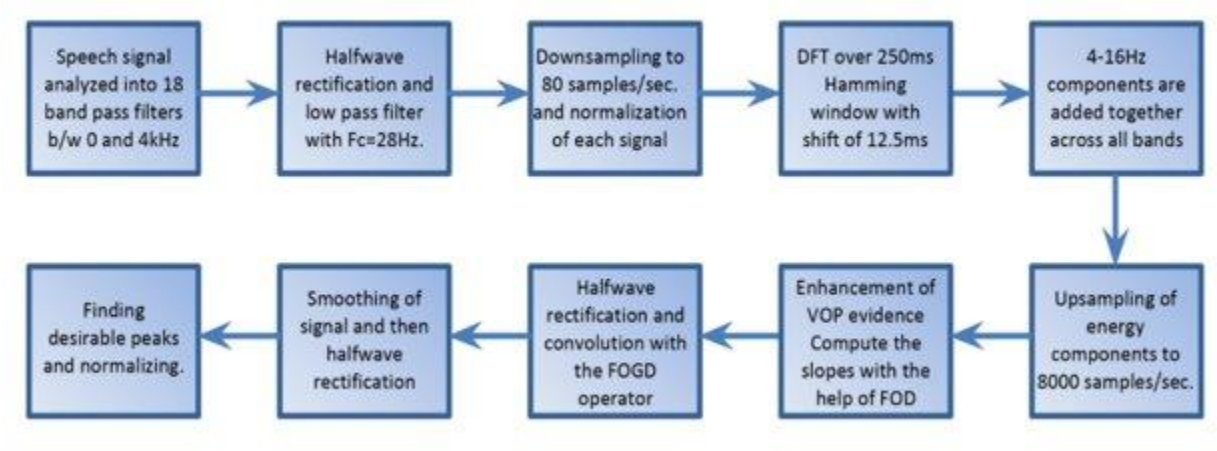
Step 6: Now, we convolve zs(n) with a first order gaussian derivative (FOGD) window, which has a standard deviation of L/6. We do this in order to get the points of significant changes in zs(n). In this convolved output, we find the peaks and valleys by keeping a proper threshold. The peaks thus obtained are the VOPs and the valleys are the VEPs.



Detection of Vowels using Modulation Spectrum

Modulation components refer to the slowly varying temporal envelope components in speech . The temporal envelope of speech is dominated by low-frequency components of several Hz. A representation of this type has compelling parallels to the dynamics of speech production, in which the articulators move at rates of 2–12 Hz, and to the sensitivity of auditory cortical neurons to amplitude modulations at rates below 20 Hz.

The generation of modulation spectrum follows the given steps:



Step 1: The speech signal is analyzed into approximately 18 critical band filters between 0 and 4 kHz. The filters are trapezoidal in shape, and there is minimal overlap between adjacent bands.

Step 2: In each band, an amplitude envelope signal is computed by halfwave rectification and low pass filtering with cutoff frequency of 28 Hz.

Step 3: Each amplitude envelope signal is then downsampled to 80 samples/s and normalized by the average envelope level in that channel,

Step 4: The modulations of the normalized envelope signals are analyzed by computing the DFT over 250-ms Hamming windows with shift of 12.5 ms, in order to capture the dynamic properties of the signal.

Step 5: Finally, the 4–16 Hz components are added together, across all critical bands.

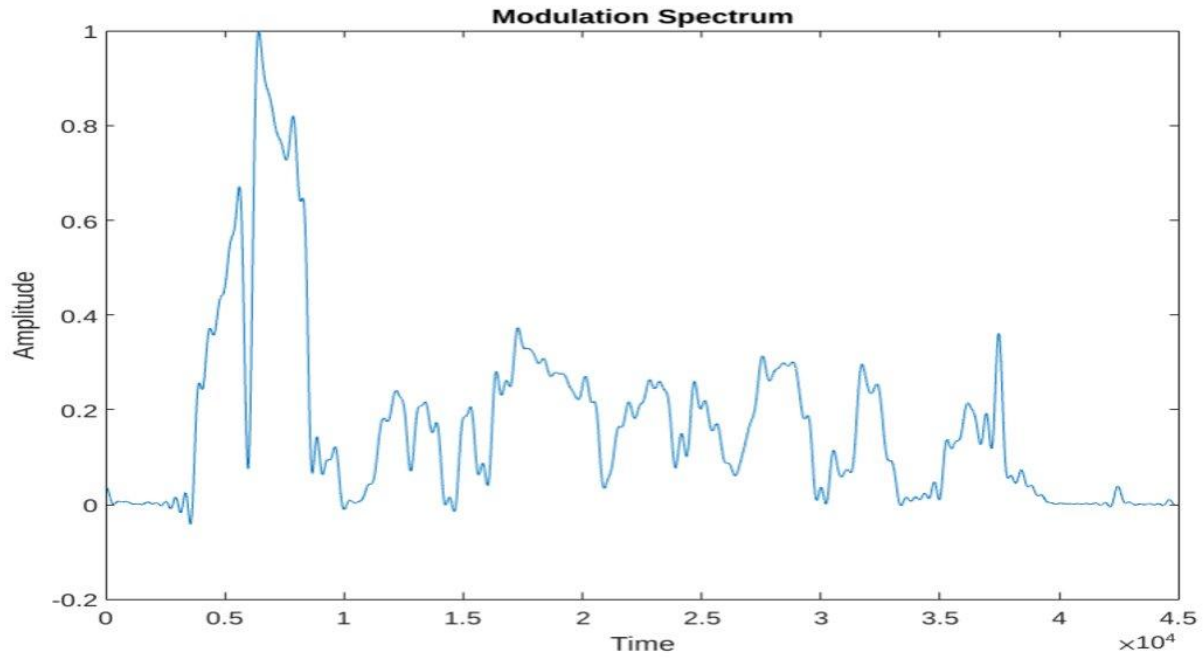
Mathematically the modulation transfer function energies are expressed as

$$m(i) = \sum_{p=1}^{18} \left[\sum_{k=k_1}^{k=k_2} \left| \hat{X}_p(k,i) \right|^2 \right]$$

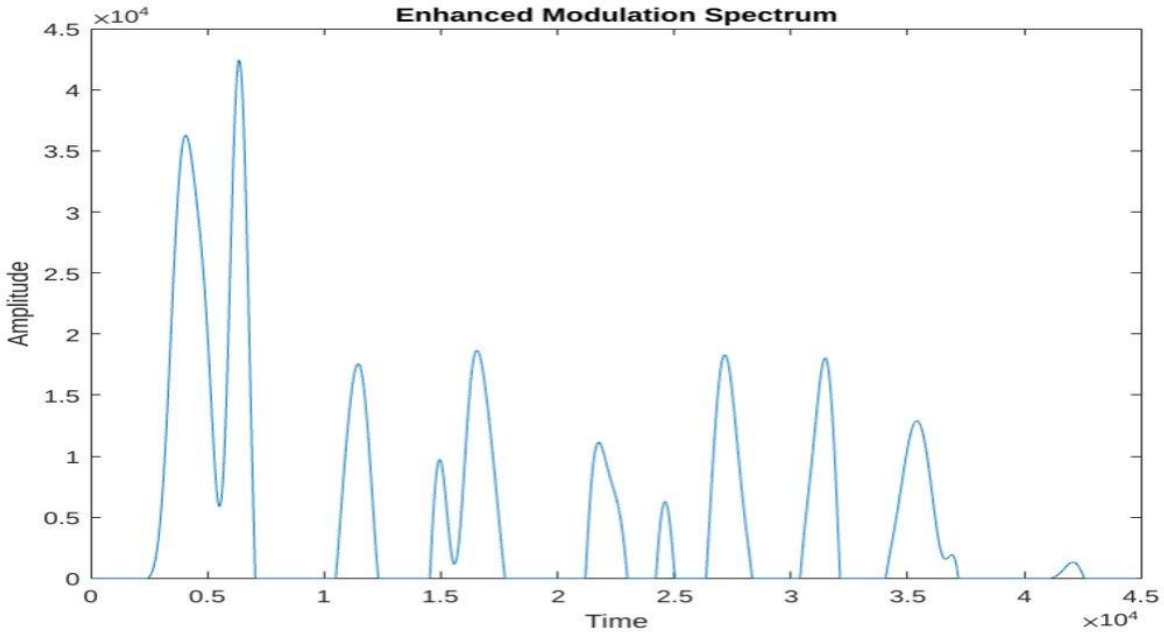
Where i is the frame index, p represents the critical band number,and k1 and k2 represent frequency index of 4 Hz and 16 Hz, respectively. Xp is computed as

$$\hat{X}_p(k) = \sum_{n=0}^{N-1} \hat{x}_p(n) w(n) e^{-\frac{j2\pi nk}{N}}; \quad p = 1, 2, \ldots, 18.$$

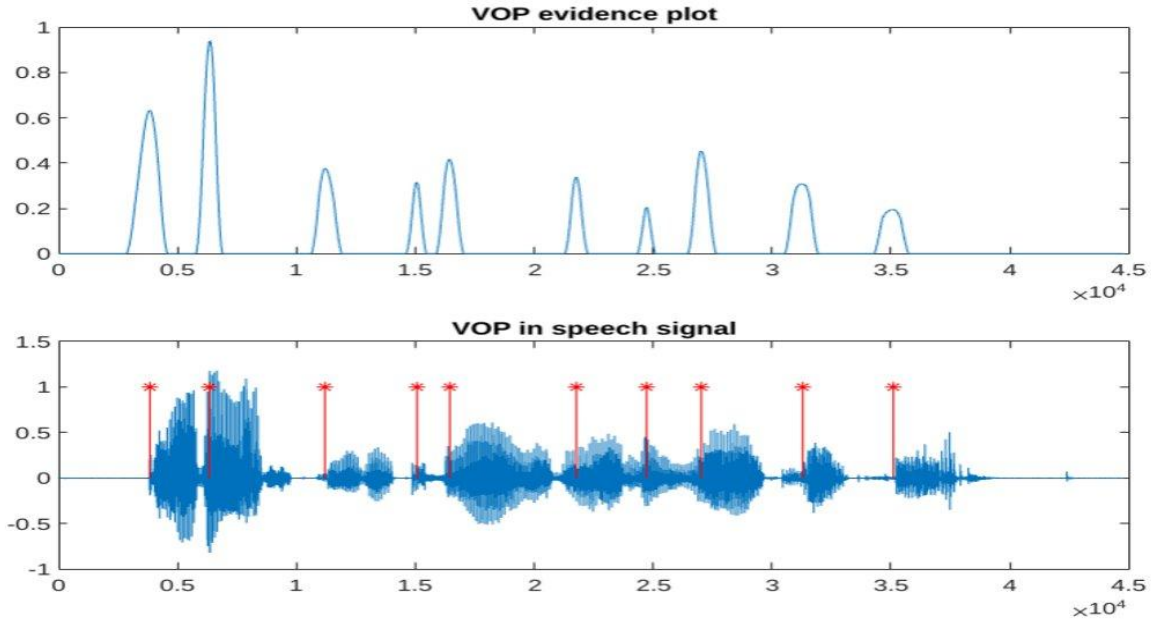
where Xp represents the normalized envelope of filter output, is a Hamming window, and N is the number of points used for computing the DFT. The modulation energy components computed for each frame are then upsampled to 8000 samples/s and plotted as function of time.



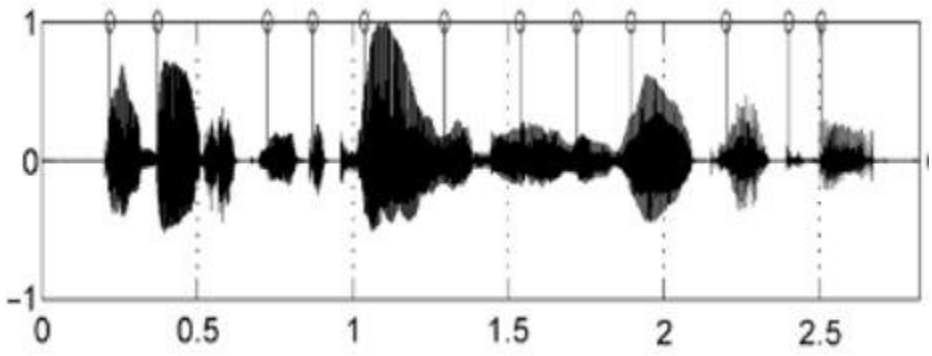
The onset of vowel can be observed as significant change in the modulation spectrum energy. The change may be further enhanced by computing the slope. This is called the enhanced modulation spectrum.



The significant change may be detected by convolving the same using FOGD operator of length 100 ms. The convolved output is the VOP Evidence Plot using Modulation Spectrum. The peak in the VOP evidence plot selected on a threshold basis indicates the location of the VOP.



Manually marked VOPs

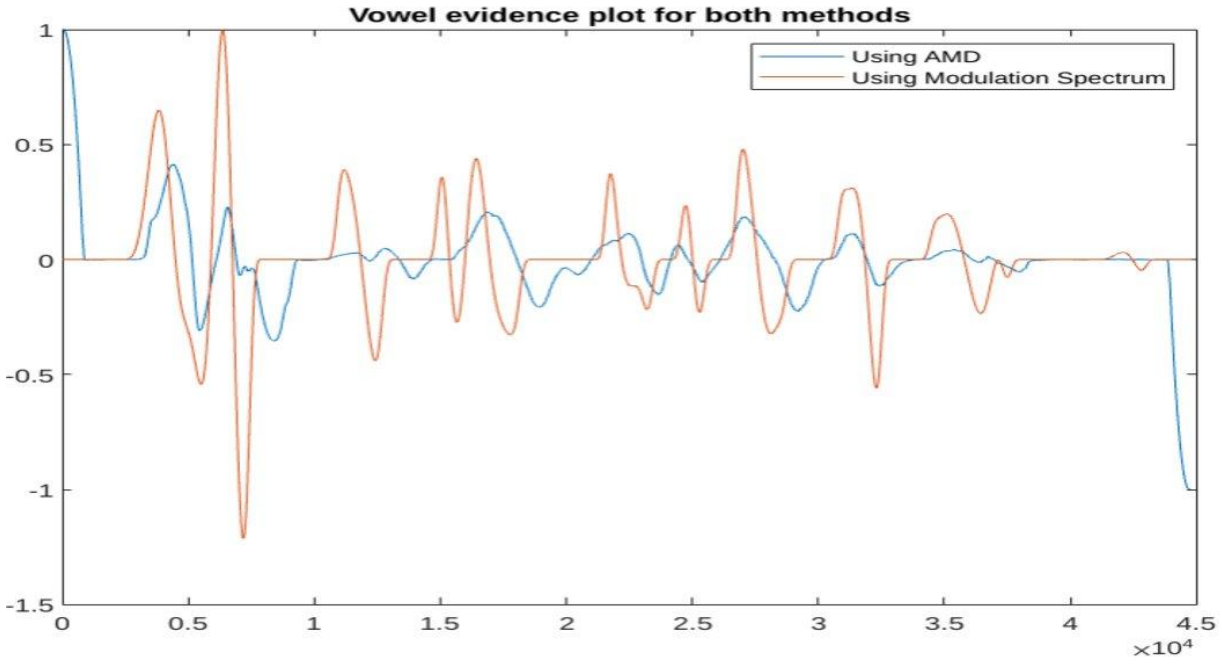


There are 12 VOPs in the above plot.

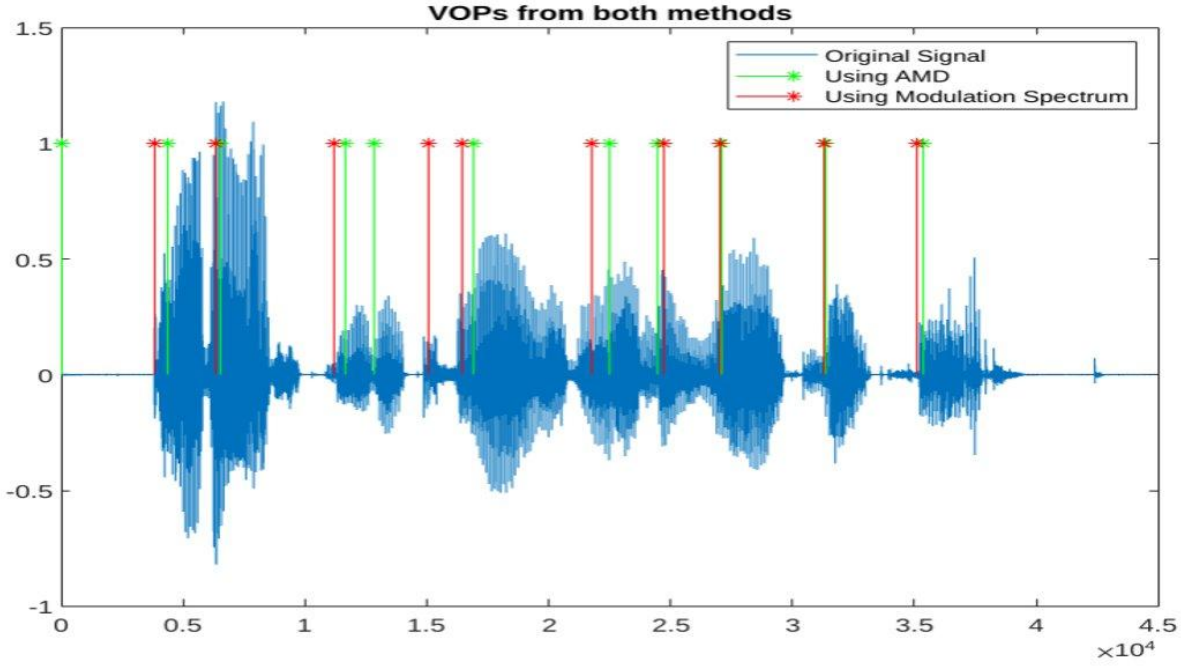
Comparison of both the methods for VOP detection

- Number of VOPs Detected:
    - 10 using average magnitude dynamics(AMD)
    - 10 using Modulation Spectrum Energies
- The VOPs detected are slightly delayed due to the filters used in the MATLAB code.
- The method using AMD gives a peak at t=0. This is a falsely detected VOP. This is due to calculations at initial points of the signal.
- Both the methods give 9 common VOPs (slightly delayed).
- The method using Modulation Spectrum Energies gives a spurious VOP

VOP evidence plot envelope for both methods:



Final VOPs detected



References

- Vowel Onset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies by S. R. Mahadeva Prasanna, Member, IEEE, B. V. Sandeep Reddy, and P. Krishnamoorthy, Student Member, IEEE
- An efficient approach for detecting vowel onset and offset points in speech signal by Sarmila Garnaik Avinash Kumar Gayadhar Pradhan Kabiraj Sethi