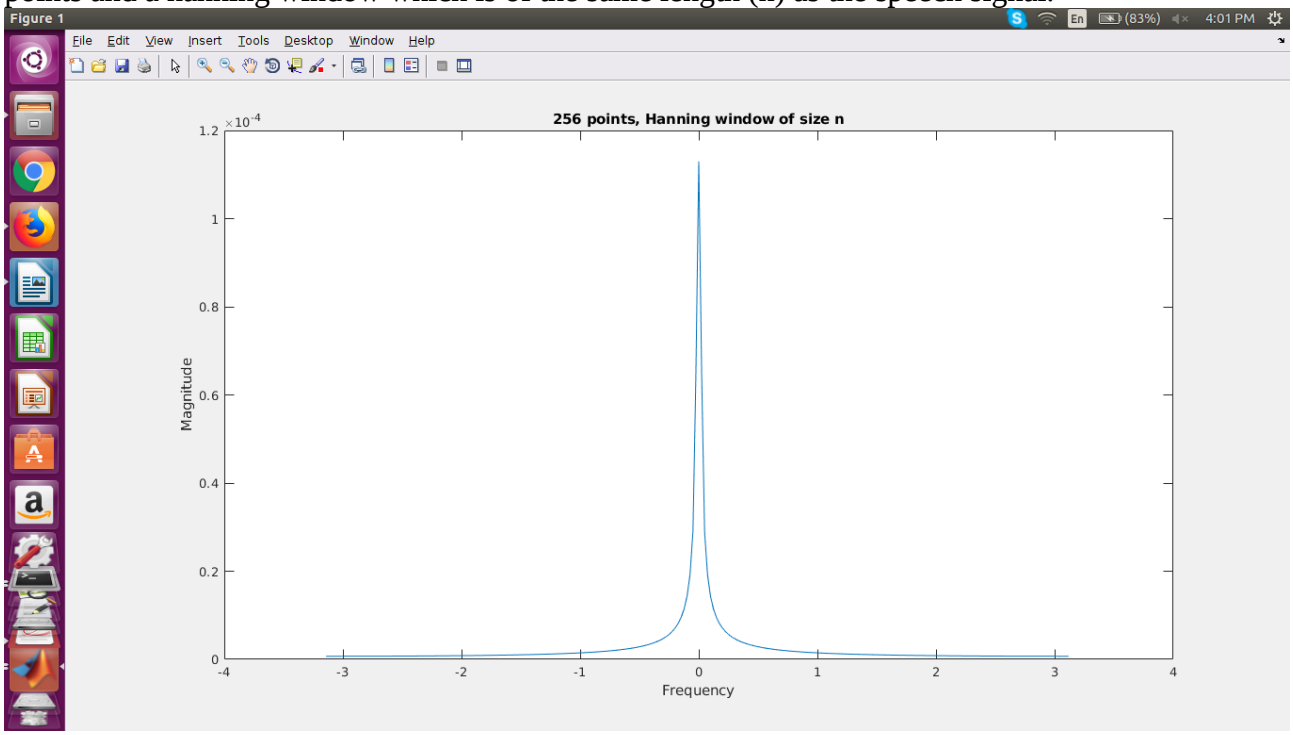
Naila Fatima
201530154
Assignment 5
Speech Signal Processing

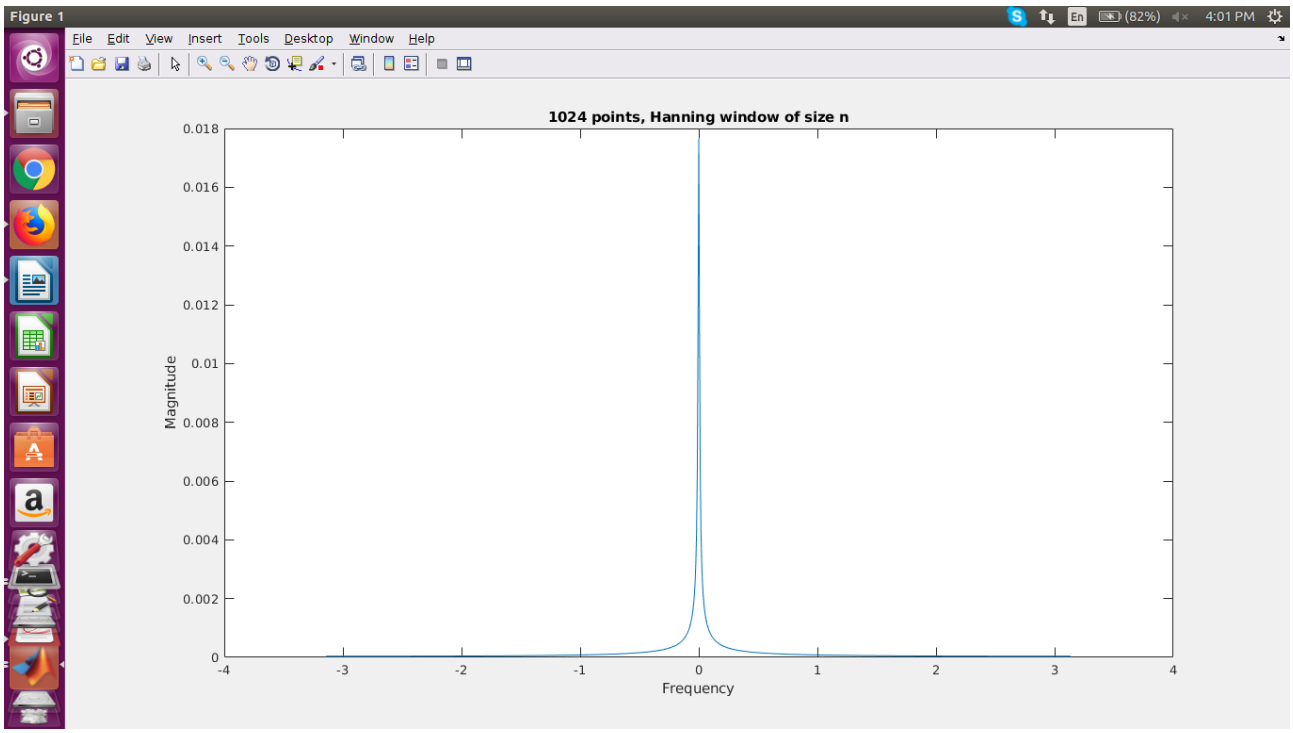
Question 2

We know that the N-point discrete time Fourier transform (DTFT) of a signal varies with the number of points, the size of the window and the shape of the window. These experiments were done on the '*arctic_a0008.wav*' file.

i) Number of points in the signal: The below plot shows the DTFT of signal taken with 256 points and a hanning window which is of the same length (n) as the speech signal.

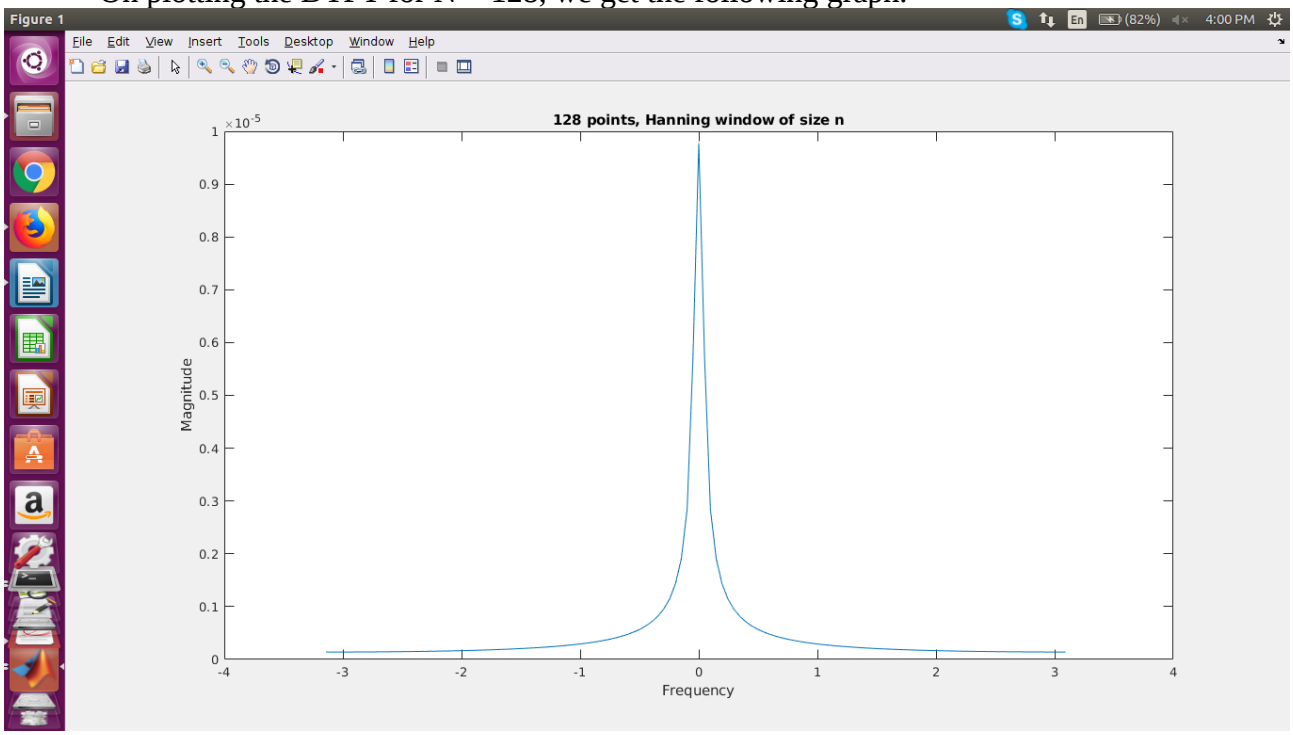


On plotting the DTFT of the same signal by changing the number of points to 1024 (while keeping the other parameters constant), we get the below plot.



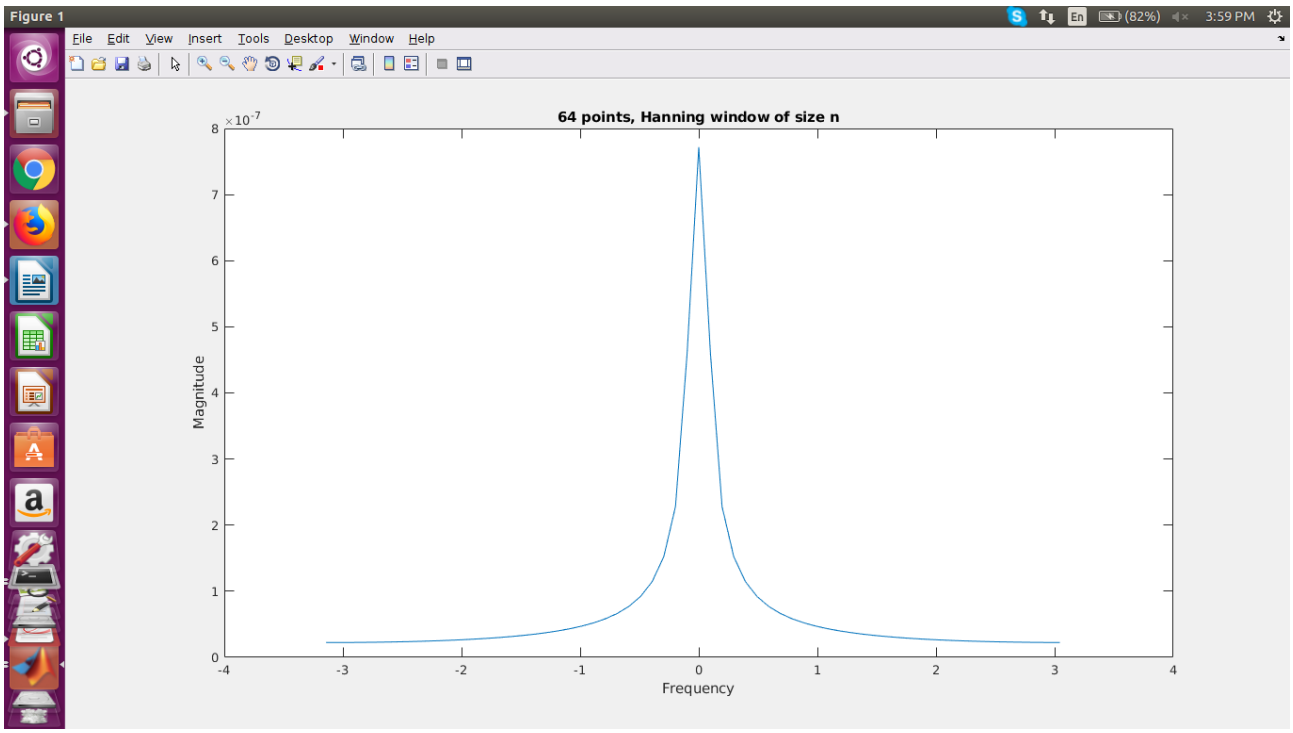
We can clearly observe that when we compare it with N = 256, the plot for N = 1024 shows a curve which has considerably greater magnitude and is more compressed in the x-direction. We can see that the 'bell-shape' of the curve for N = 256 is slightly wider that than for the curve with N = 1024. The curve for N = 1024 is more impulse-like when compared to N = 256. We can see that the magnitude for N = 1024 is close to 0.018, whereas that for N = 256 is around 1.2×10^{-4}.

On plotting the DTFT for N = 128, we get the following graph.



We can observe that the magnitude for this has dropped to around 1×10^{-5} while its 'bell-shape' has increased in width (it is less compressed in x direction).

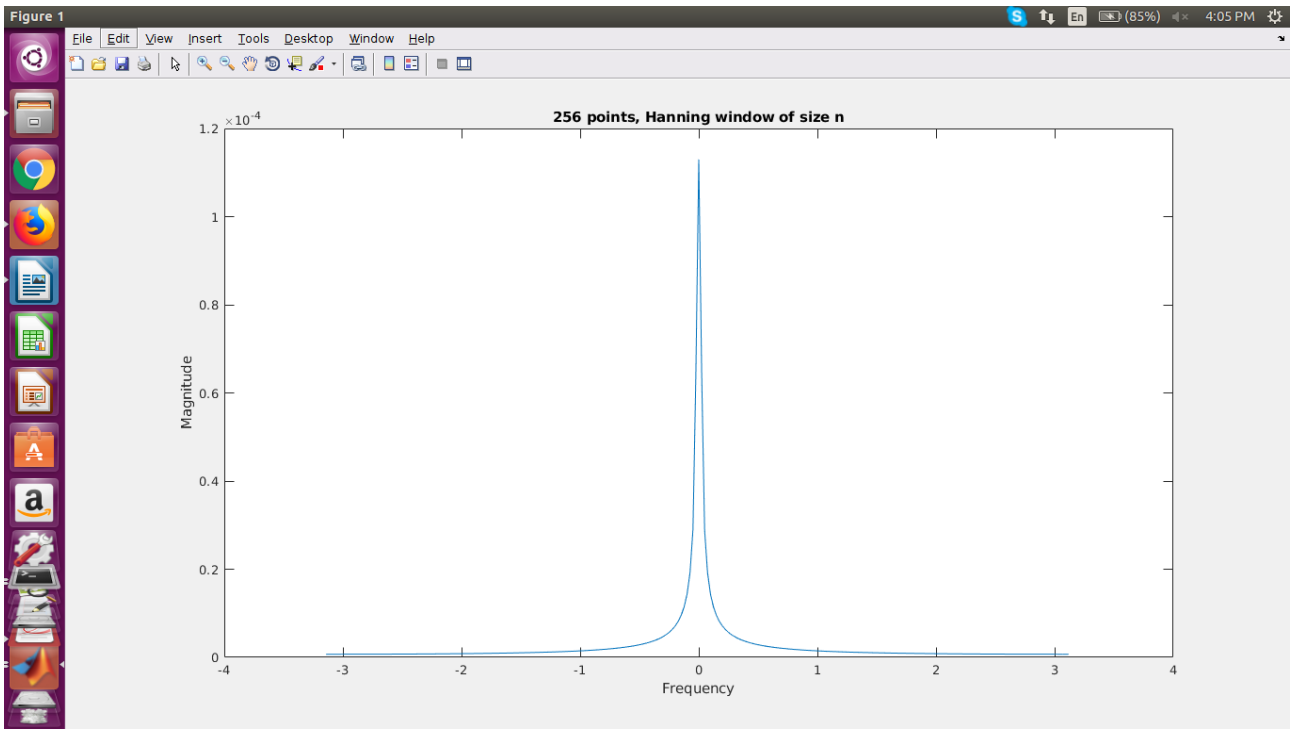
On reducing N to 64, we get the following plot.



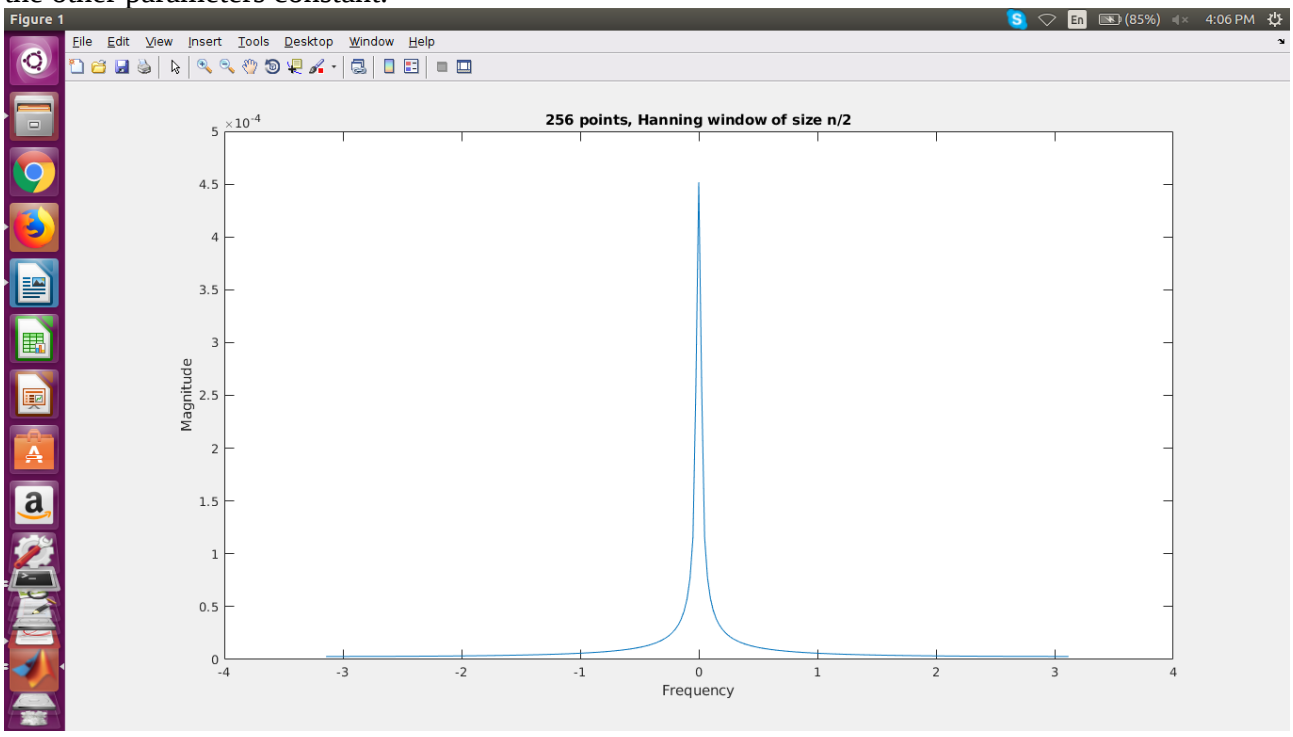
We can clearly see that the magnitude has dropped further to $8x10^{-7}$ while the width of the curve has increased further.

We can observe that on reducing the number of points N in the DTFT, the magnitude of the DTFT reduces while the width of the curve increases. The curve can be considered to have expanded in the x-direction. Similary, on increasing N, the magnitude of the DTFT curve increases while the width of the curve reduces as it becomes more narrower, it becomes more impulse-like. It should also be observed that as the value of N increases, the frequency resolution becomes worse leading to distortions in the plots.

ii) Size of the window: The plot for the DTFT by taking 256 points and a Hanning window of length n (where n is the length of the speech signal is shown below).

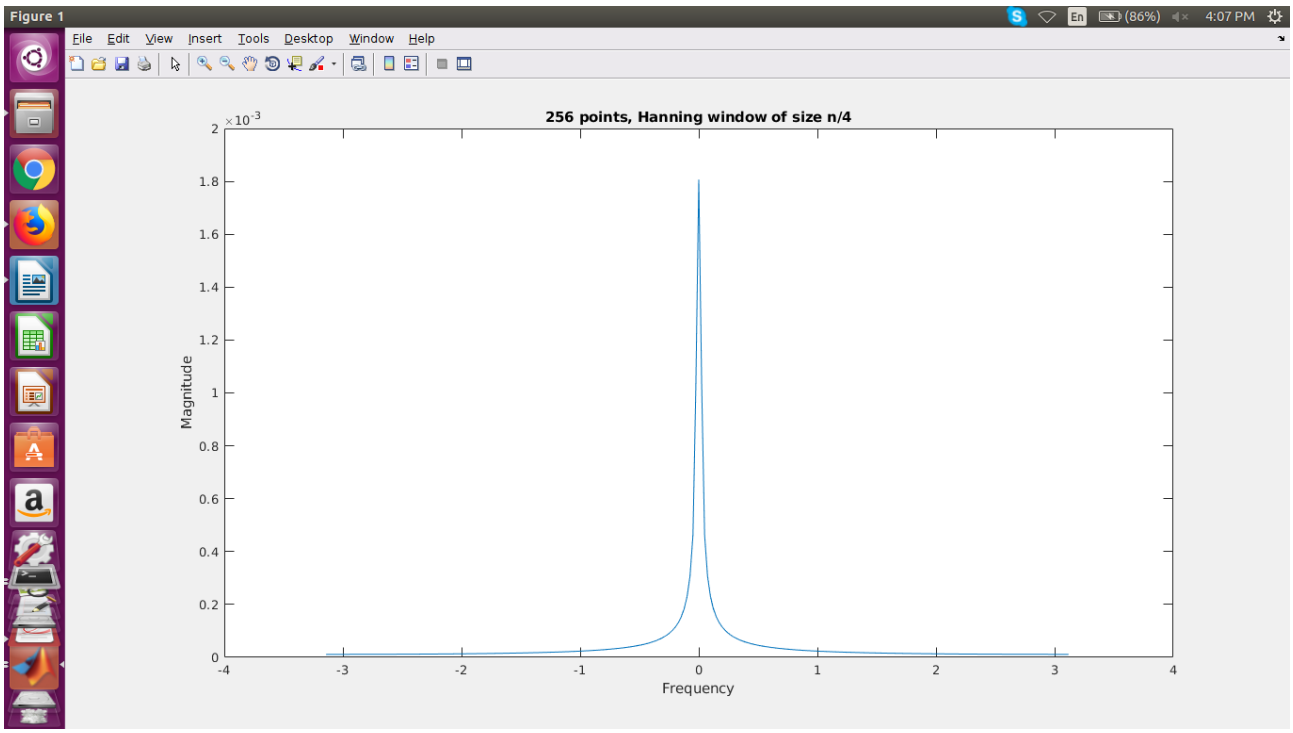


If we reduce the size of the window to n/2, we get the following plot. Note that we have kept the other parameters constant.



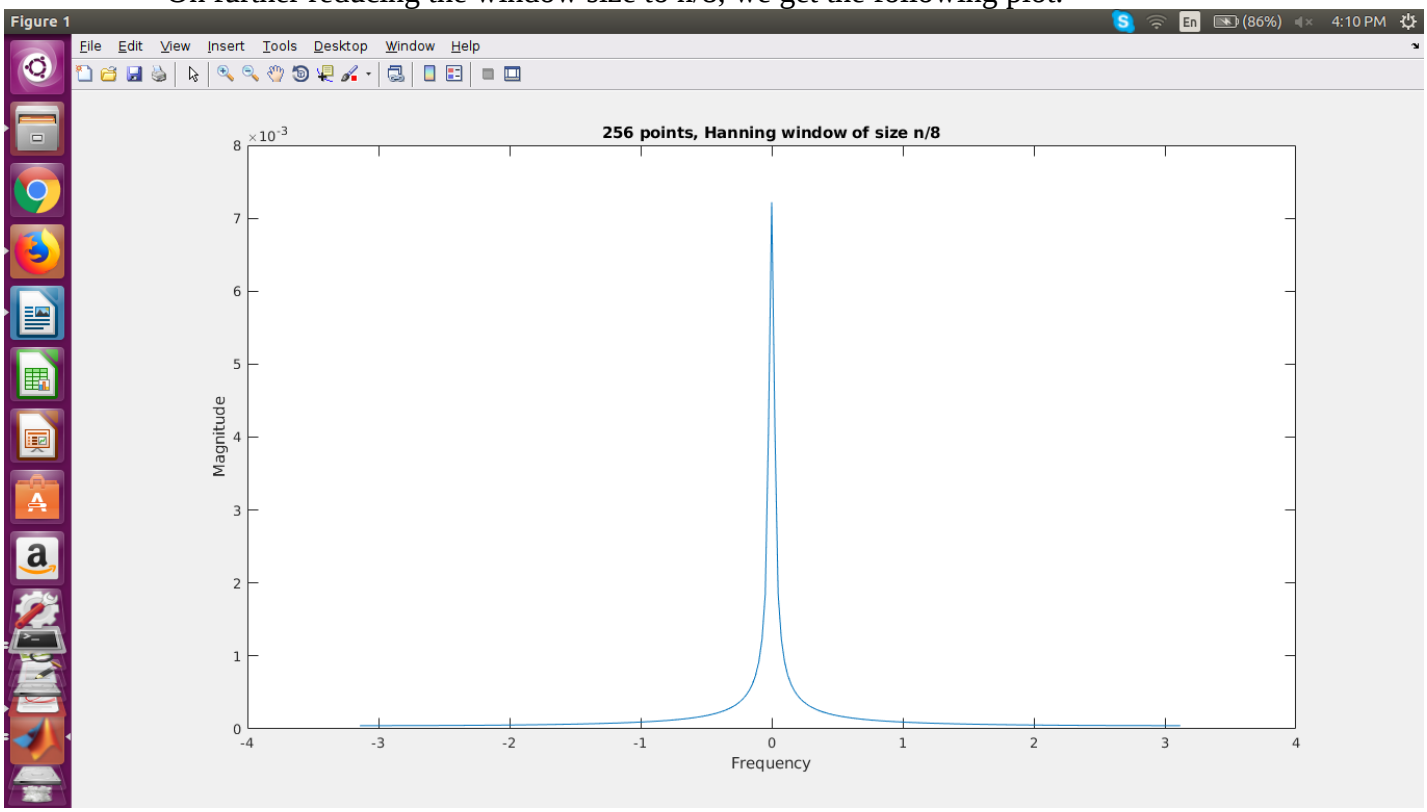
We can observe that the magnitude of the curve has slightly increased. We can see that in the curve for window size = n, the magnitude is upto 1.2×10^{-4}, whereas for the window size = n/2, the magnitude increases upto 5×10^{-4}. The width of the curve has not changed perceptibly by halving the window size.

On further reducing the window size to n/4, we get the plot below.



We can further see that the magnitude for this curve has increased upto $2x10^{-3}$ whereas the width of the curve remains almost the same.

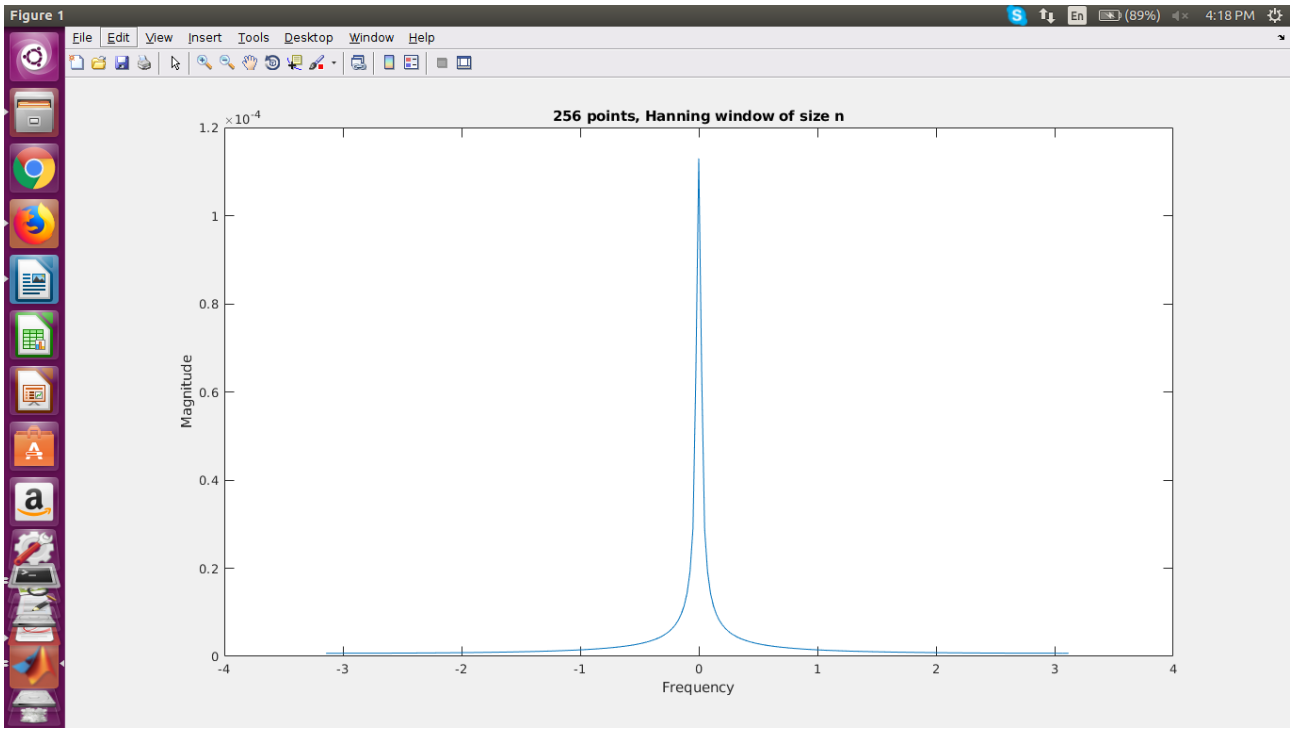
On further reducing the window size to n/8, we get the following plot.



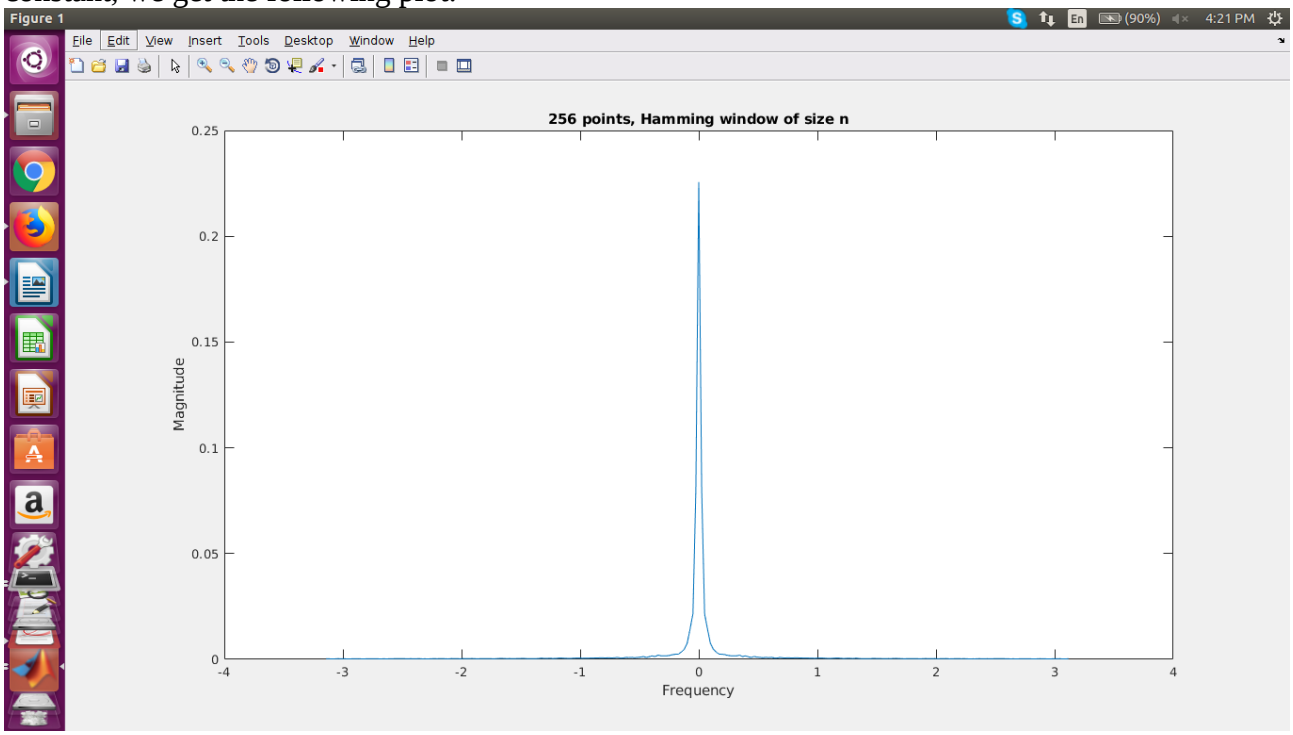
We can observe that the magnitude for this plot has increased upto $8x10^{-3}$.

On seeing the pattern above, we can conclude that by reducing the window size, the magnitude of the curve tends to increase whereas on increasing the window size, the magnitude of the curve reduces. It should be noted that a larger window gives better frequency resolution but poor time resolution whereas a narrower window gives better time resolution but poor frequency resolution.

iii) Shape of the window: The plot for the DTFT with 256 points, and a Hanning window of size n is shown below.

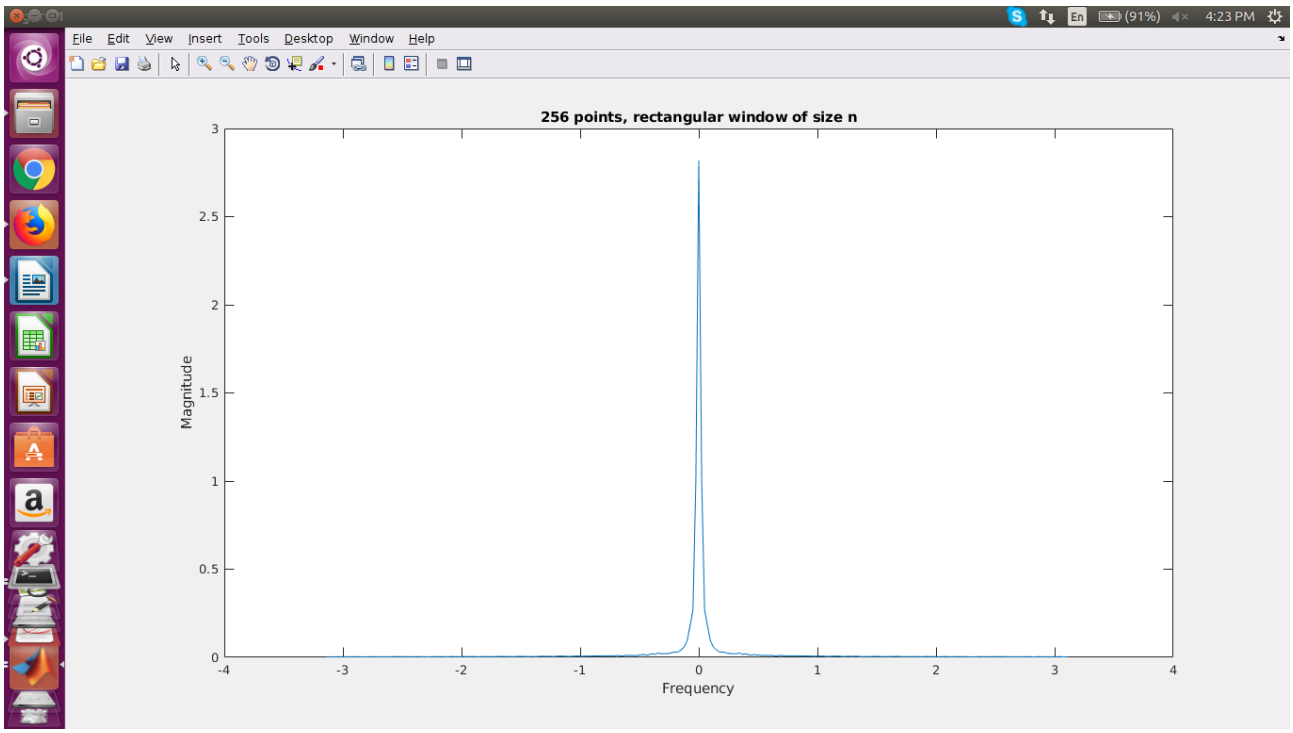


On changing the window shape to a Hamming window, while keeping the other parameters constant, we get the following plot.



We can observe that the magnitude of the curve has increased dramatically. For a Hanning window, the magnitude goes upto 1.2×10^{-4} whereas for a Hamming window, the magnitude goes upto 0.25. We can also observe that the width of the curve reduces significantly. For a Hanning window, the width was between -0.75 and 0.75 whereas for the Hamming window it is between -0.2 and 0.2. The curve is for the Hamming window is also less smooth near the bottom when compared to that for the Hanning window.

On changing the window shape to a rectangular window, we get the following plot.



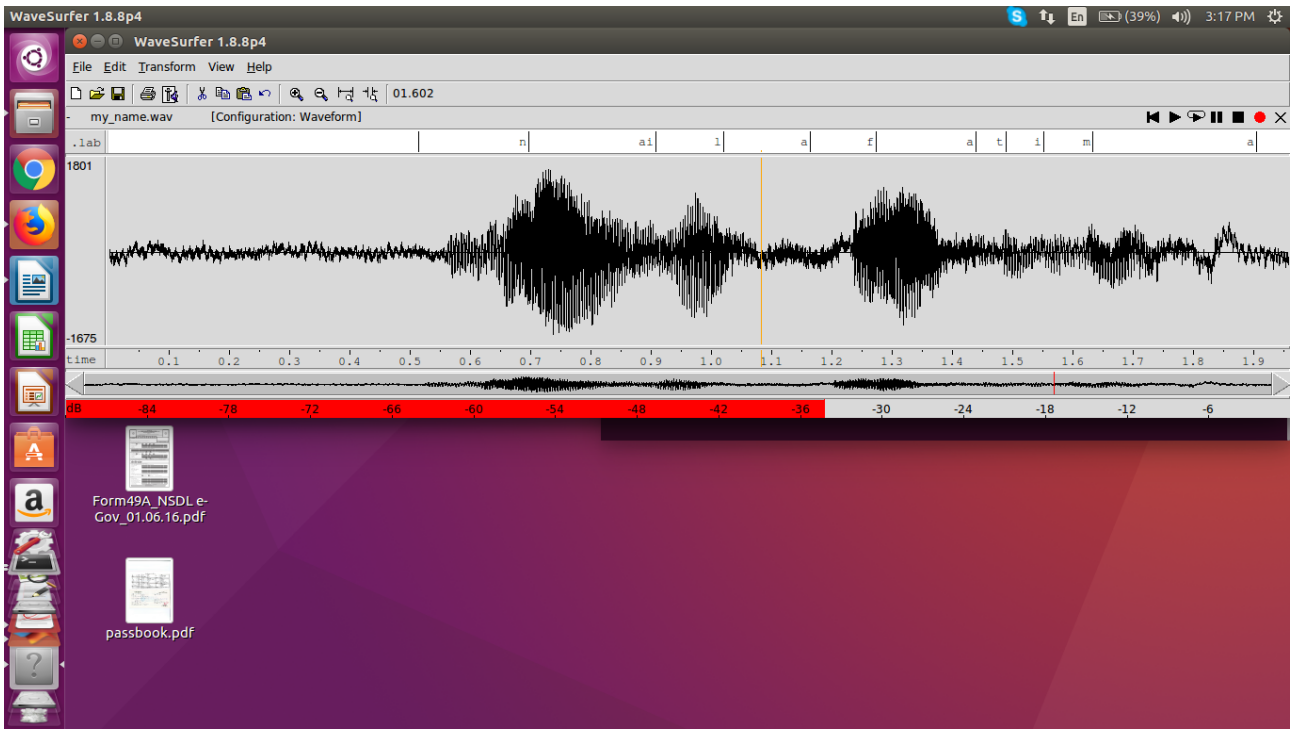
We can observe that the magnitude for the rectangular window has increased up to 3. The width seems to be as much as that for the Hamming window and considerably lesser than that for the Hanning window. The bottom is also less smooth when compared to the Hanning window.

Overall, we can observe that Hanning window has the least magnitude with the largest width (bell-shape) and that the curve is also smooth. The Hamming window has a slightly greater magnitude with lesser width and the curve is not as smooth. The rectangular window has an even greater magnitude with less width and a curve which is not as smooth as that for the Hanning window. Rectangular window is the least smooth which is why it gives such a plot. The Hanning window is more smooth when compared to the Hamming window, which is why its plots are also relatively smoother than those for the Hamming window. This is as the Hamming window has more discontinuities when compared to the Hanning window and the rectangular window has even more discontinuities.

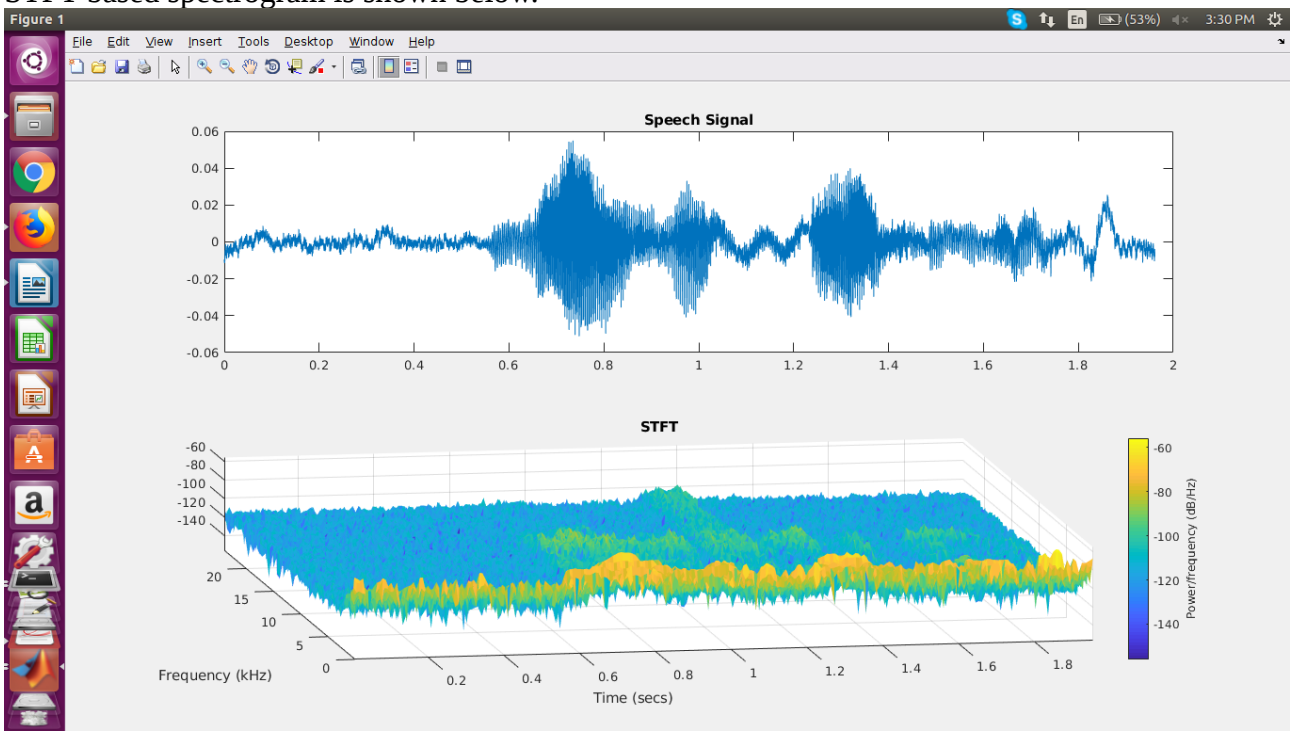
In conclusion, increasing the number of points causes the magnitude of the curve to increase while the width reduces. Increasing the window size causes the magnitude of the curve to reduce. The Hanning window has least magnitude but greatest width, the Hamming window has more magnitude but lesser width and the rectangular window has greatest magnitude but less width.

Question 1

The wavefile '*my_name.wav*' contains a recording of my name and the file '*my_name.lab*' contains the transcription for this recording. The matlab code used for this question is '*q1b.m.*' The speech signal is given below along with the transcriptions.



We know that a spectrogram is a visual representation which allows us to view how the spectrum of frequencies of a signal varies with time. The short-time Fourier transform (STFT) is a method which can be used to plot the spectrogram of a signal. The STFT of a signal is computed by taking smaller segments of the signal and computing the Fourier transform independently on each of the smaller segments. The plot showing both the speech signal waveform as well as the 3D STFT-based spectrogram is shown below.



From the image above, we can observe that the power per unit frequency is higher in the regions corresponding to higher amplitudes of the speech signal. For example, we can observe that in the interval between 0.6-0.8 seconds, the power per unit frequency is considerably higher when compared to surrounding regions. The regions in the spectrogram which show relatively higher power per unit frequency are denoted by the colours yellow and green whereas the regions which

show relatively smaller power per unit frequency are denoted by shades of blue. We can also observe that according to the plot, the frequencies which are low (0-10 kHz) have a greater power per unit frequency when compared to those which are high. This can be seen as the colours denoting higher power per unit frequency (yellow/green) are concentrated in regions with lower frequencies whereas the colours denoting lesser power per unit frequency (blue) are dominant in regions with relatively higher frequencies (>10 kHz).

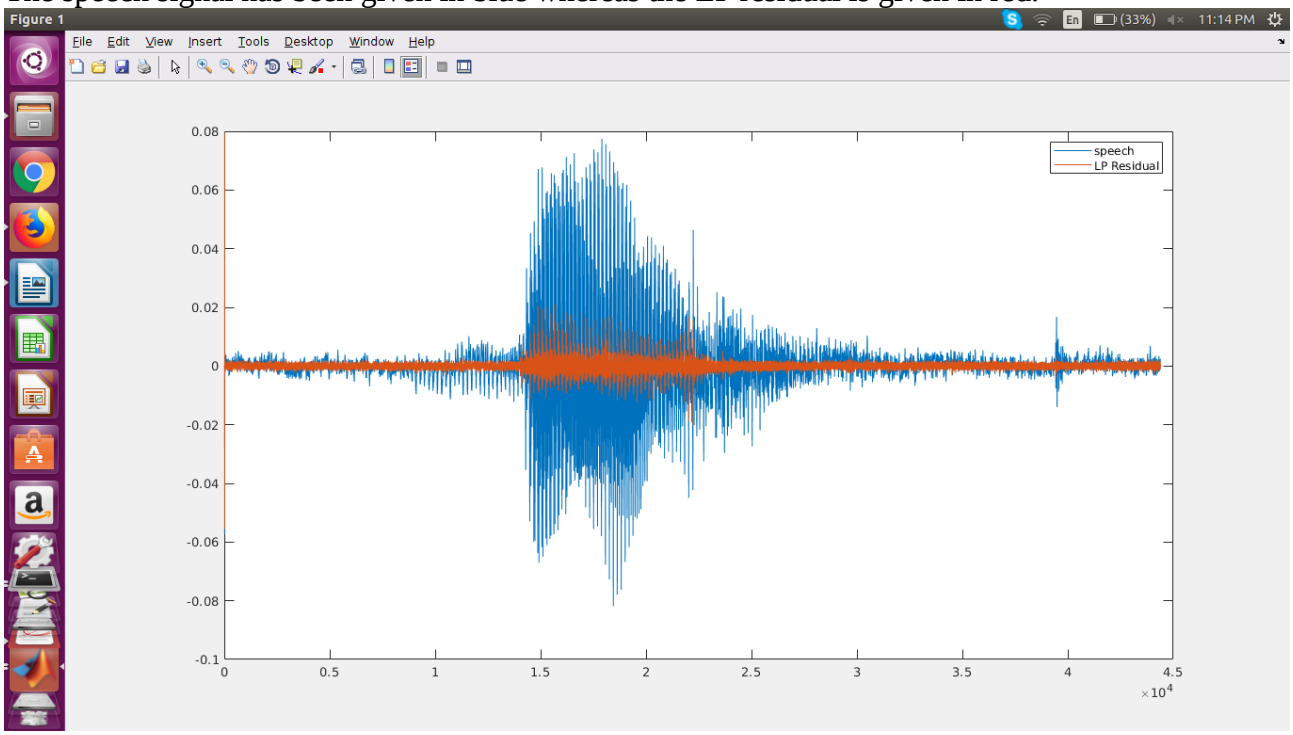
Question 3

We can find the pitch of a signal by using linear prediction (LP) analysis as well as cepstral analysis. LP analysis is a method which allows us to estimate future samples by using a linear weighted sum of previous samples. The LP residual refers to the difference between the actual signal and the estimated signal. We can find the pitch of the signal by using LP analysis. This can be done by finding the autocorrelation of the LP residual and finding the difference in samples between the two largest peaks in the autocorrelation of the LP signal. To find the pitch frequency, we take the ratio between the sampling rate and the number of samples between the two largest peaks in the autocorrelation of the LP residual.

Pitch frequency = Sampling Rate/ Difference in samples between two largest peaks of autocorrelation of LP residual

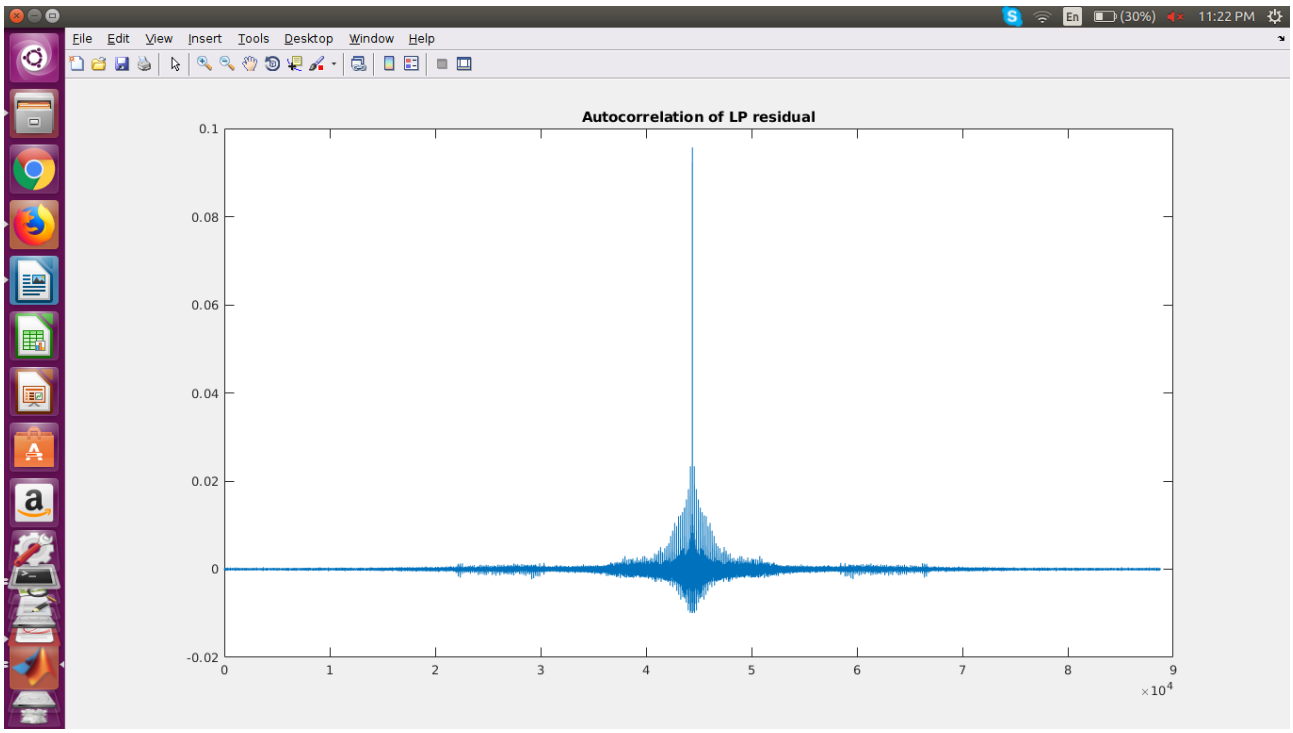
The pitch period will be the reciprocal of the pitch frequency.

The below image shows us a plot of the speech signal and the corresponding LP residual. The speech signal has been given in blue whereas the LP residual is given in red.

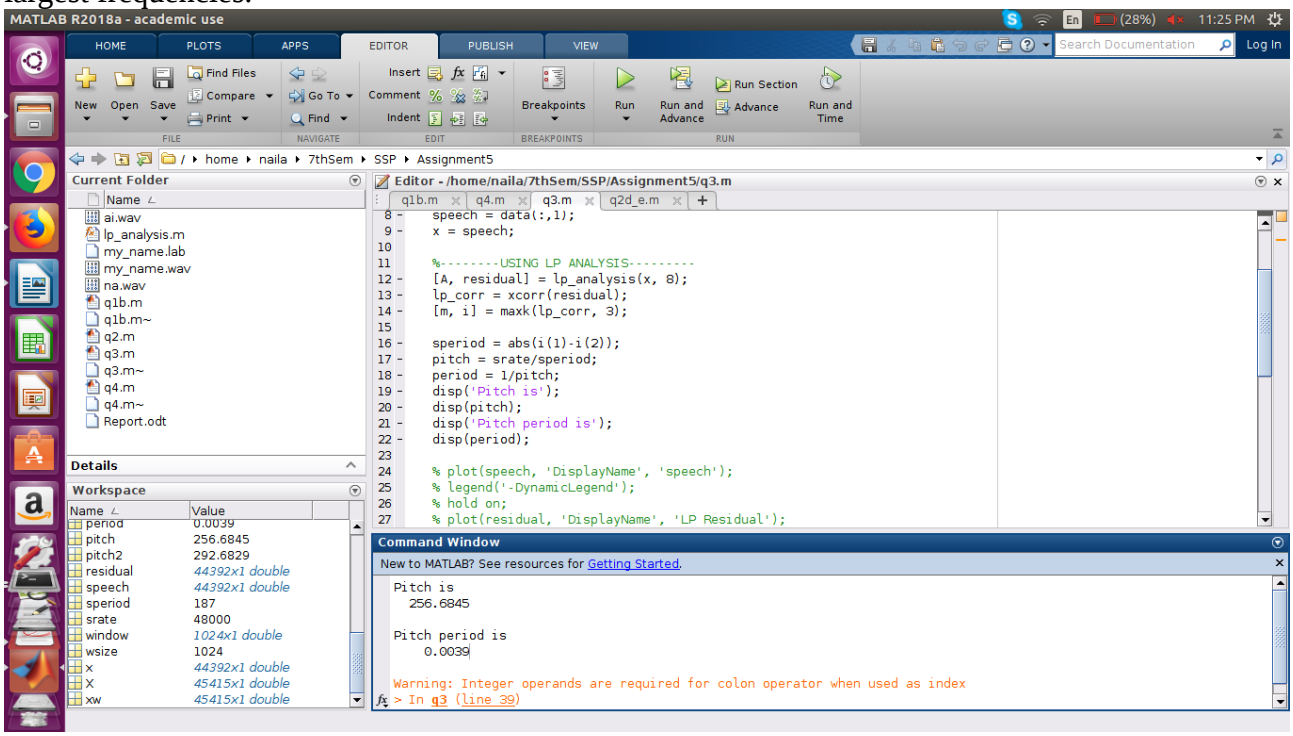


We can observe that the values of the LP residual are very small (less than 0.02) and this denotes the fact that the LP coefficients a_k are able to predict the future samples well. We can also observe that the peaks in the LP residual correspond to sudden changes in the speech signal. When the coefficients are not able to predict the sudden changes, the difference between the actual and predicted signals will be larger and this is why we have peaks at such locations.

The autocorrelation of the LP residual is given below.



As expected, the autocorrelation of the residual is symmetric. We can find the pitch from the autocorrelation of the residual by finding the difference in indices of the two largest peaks in the autocorrelation. This gives us the number of samples between the two largest peaks. We then find the pitch by finding the ratio of the sampling rate and the number of samples between the two largest frequencies.



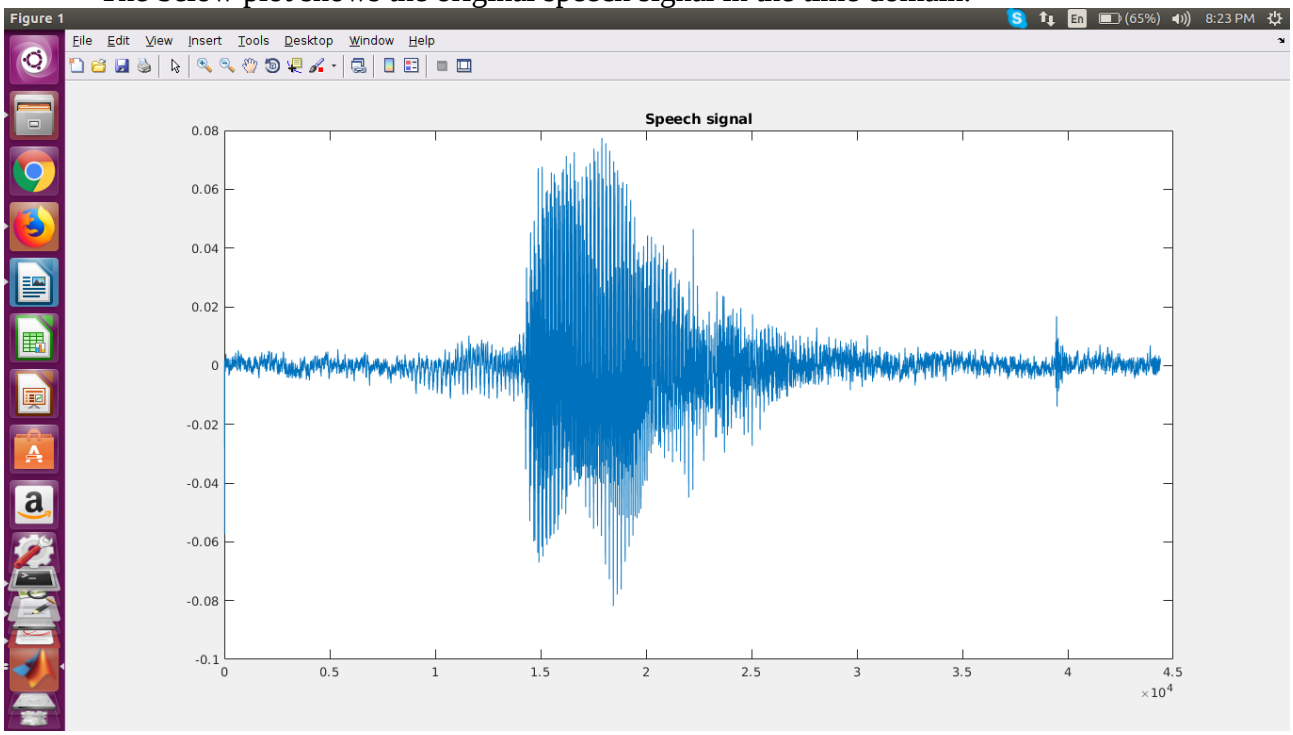
As we can see, the pitch is around 256 Hz whereas the pitch period is 0.0039 seconds. This is for '*na.wav*' recording which is a voiced syllable. It should be noted that this value of the pitch seems reasonable as the pitch of females varies from 160-300 Hz. Note that for this question, I have found the LP coefficients by using an order 8.

We can also use cepstral analysis to find the pitch of a signal. The cepstrum is the inverse DFT of the logarithm of the magnitude of the fourier transform of a signal. It is given by,

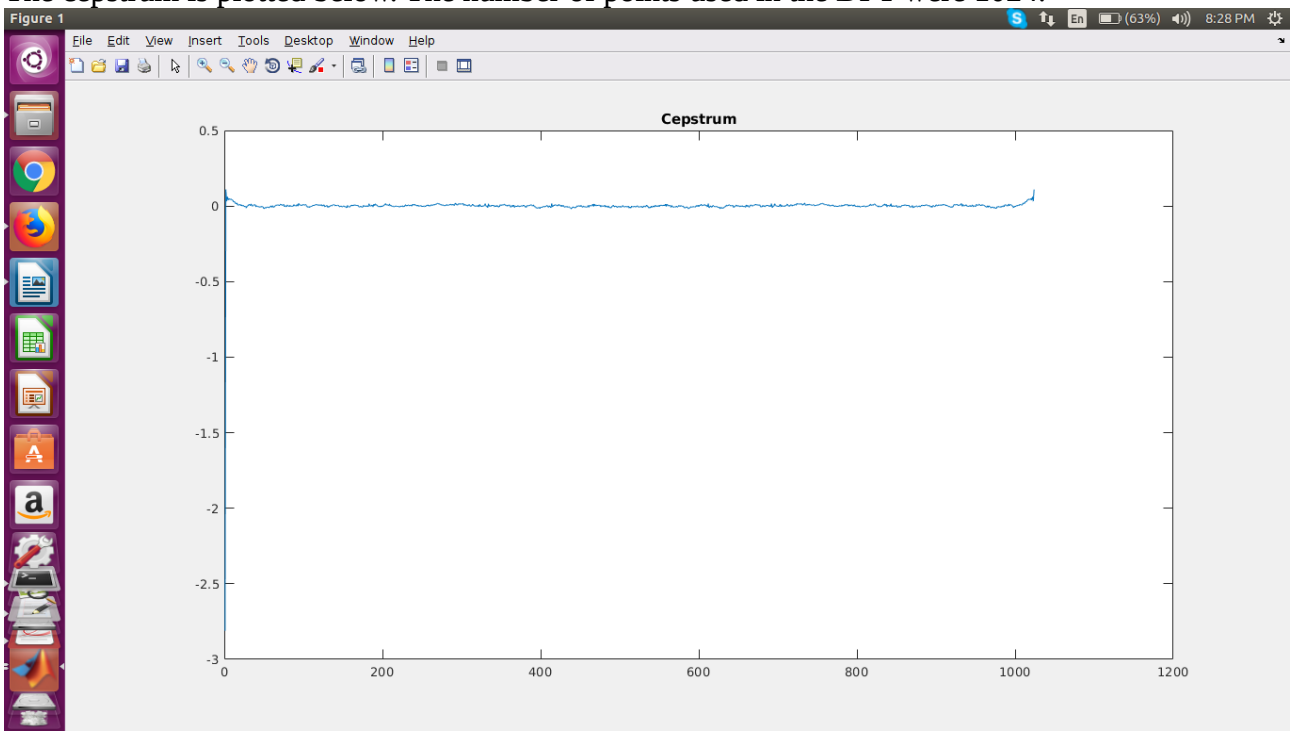
Spectrum = IDFT(log(DFT(signal)))

We can get the pitch from the cepstrum by doing a process known as liftering. When we compute the cepstrum, we get an x-axis analogous to frequency which is known as the quefrequency. A filter in the quefrequency domain is known as a lifter. We can apply a lifter to the cepstrum and the largest peak in the filtered cepstrum will correspond to the pitch frequency and we can find the corresponding pitch period as it is just the reciprocal of the pitch frequency.

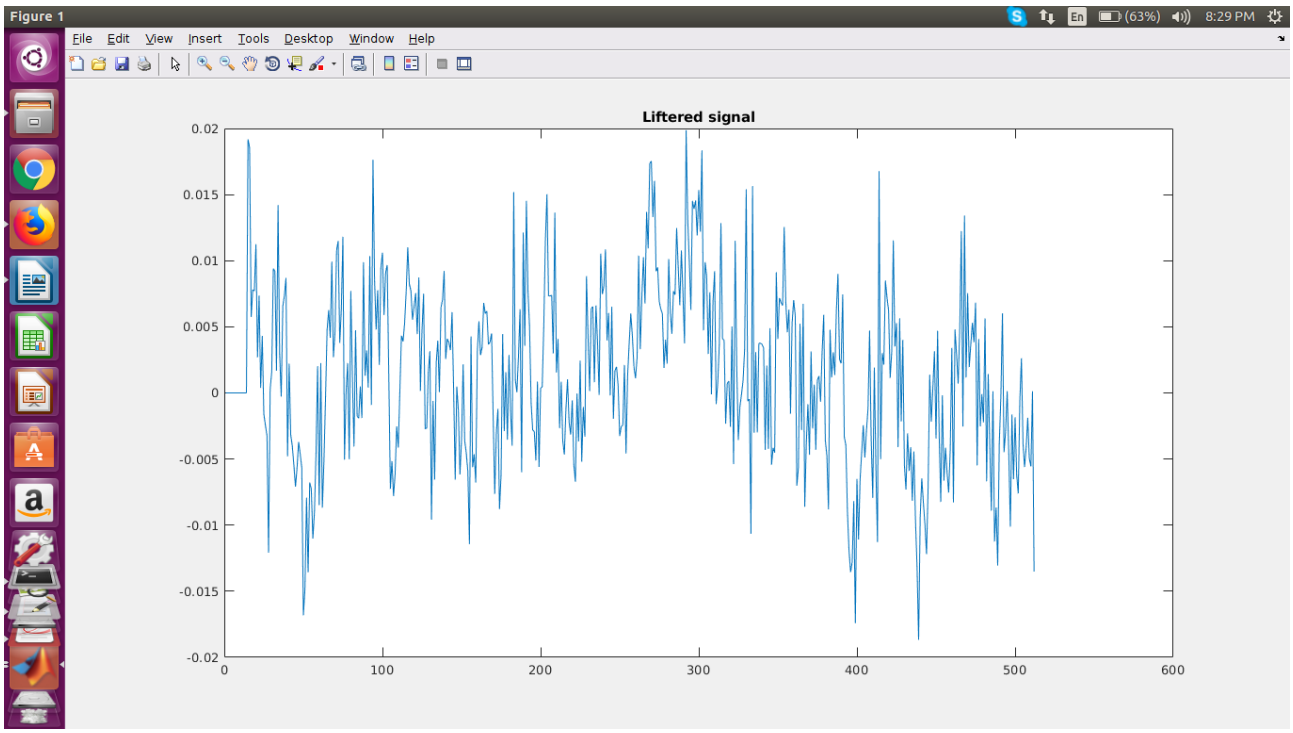
The below plot shows the original speech signal in the time domain.



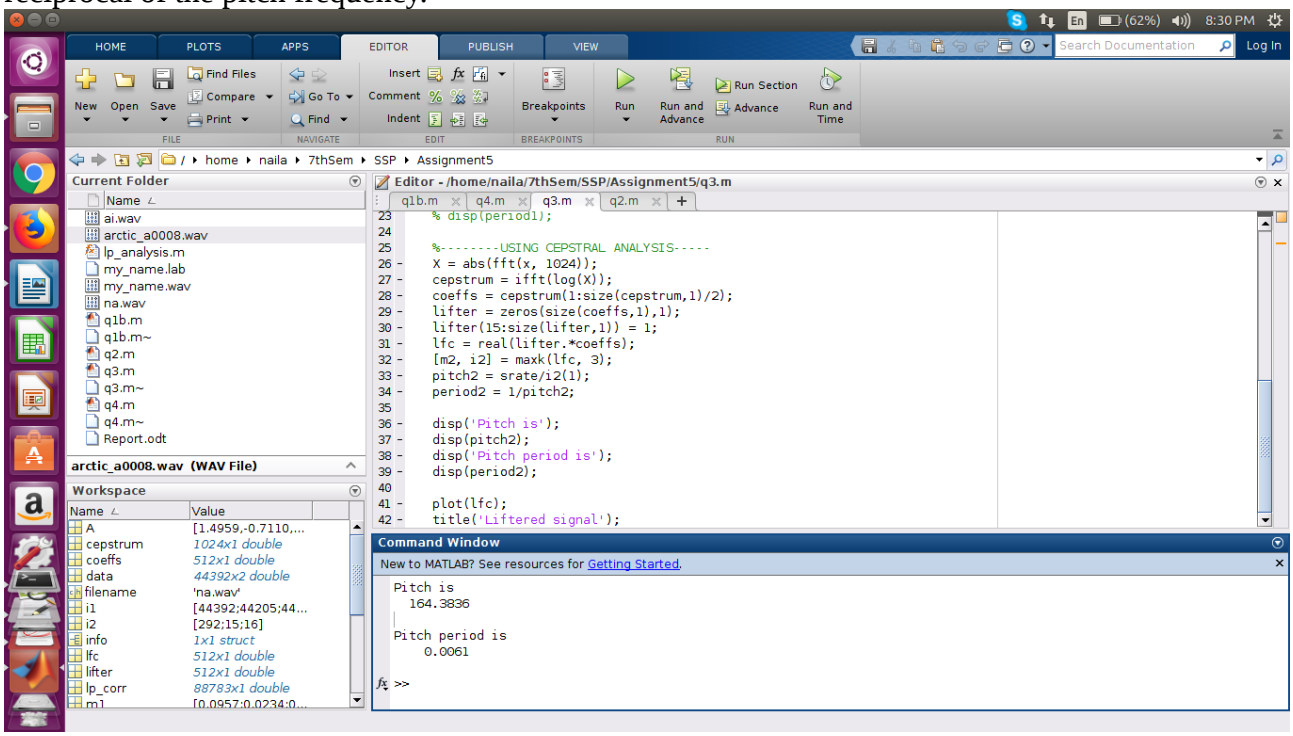
We can compute the cepstrum by finding the IDFT of the log of the magnitude spectrum. The cepstrum is plotted below. The number of points used in the DFT were 1024.



The liftered signal is shown below. This was achieved by applying a lifter to the cepstrum.



The pitch will correspond to the highest peak. We can get the pitch period by taking the reciprocal of the pitch frequency.



As shown below, the pitch is estimated to be 164 Hz whereas the pitch period is 0.0061 seconds. This makes sense as the pitch of a female is between 160-300 Hz. Note that these observations are for the '*na.wav*' recording which is a syllable.

Question 4

a) VOP and VEP refer to vowel onset point and vowel end point, respectively. Vowel onset point refers to the point at which a vowel starts whereas vowel end pooint refers to the point at which a vowel ends. It should be noted that the region between the VOP and the VEP is known as the vowel region or the sonarant region. Determining the VOP and the VEP is essential to speech signal

processing as vowels act as anchor points which can be used to extract prosody features. This allows us to do prosody modification (which involves changing the pitch and duration of speech without affecting the message or the naturalness of the signal). They can also be used for consonant-vowel (CV) recognition and speech rate modification.

b) The VOP locations of a speech signal can be determined by using methods involving the spectral energy, the LP residual, the modulation spectrum or a combination of the three. It should be noted that these methods are complementary to one another as they involve the frequency domain, time domain and the modulation domain. I have used evidence from the LP residual to determine the VOP locations.

For this method, we have to find the LP residual of each segment of a speech signal. This can be found by first determining the LP coefficients a_k for a particular order (I have used 10) and then finding the residual which is the difference between the actual and the predicted signal. The predicted signal is found by taking a linear weighted sum of previous samples where the weights are the LP coefficients.

Predicted signal $\hat{s}(n) = \Sigma a_k s(n-k)$ where k = 1 to 10

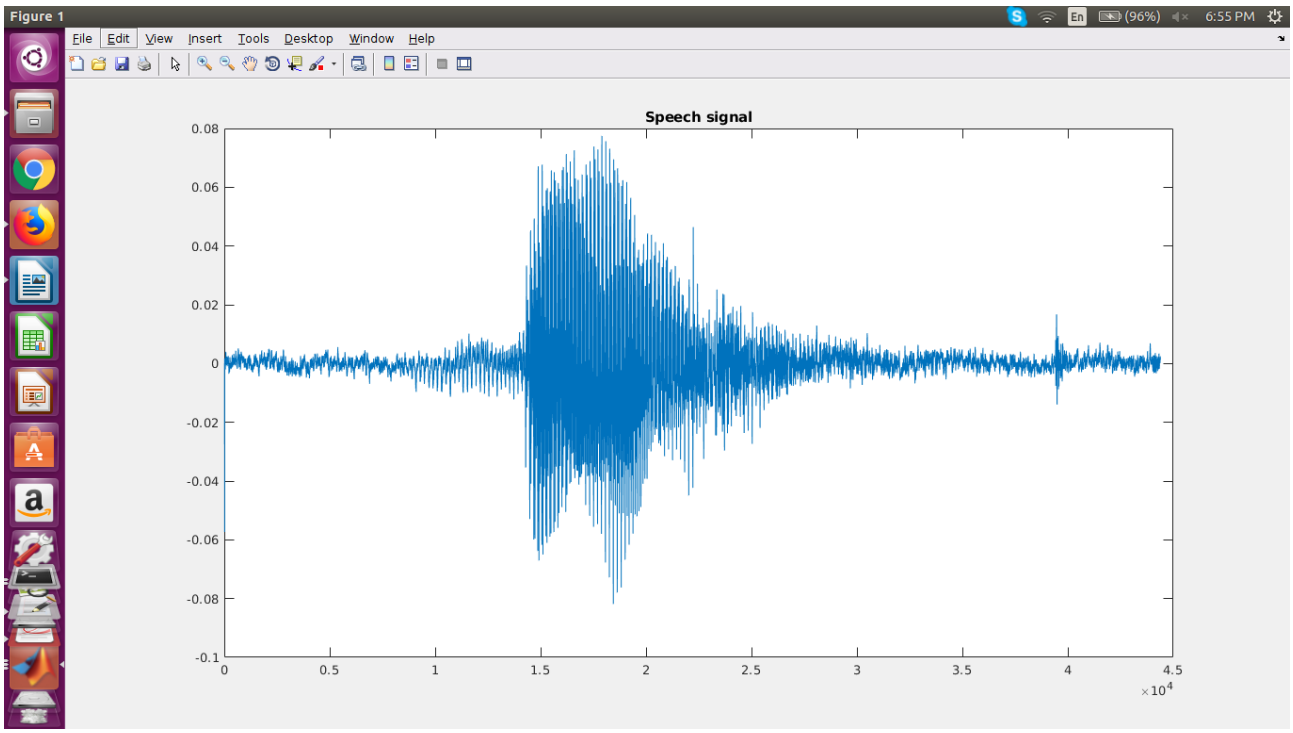
LP residual $= s(n) - \hat{s}(n)$

We divide the signal into blocks of size 20 ms and find the LP coefficients for each block and compute the residual. The LP residual contains the excitation characteristics and we enhance these further by computing the Hilbert envelope of the LP residual. After computing the Hilbert envelope of the LP residual, we smooth it by using a Hamming window of 50 ms. This smoothed version of the Hilbert envelope of the LP residual is considered to be a good representation of excitation source energy.

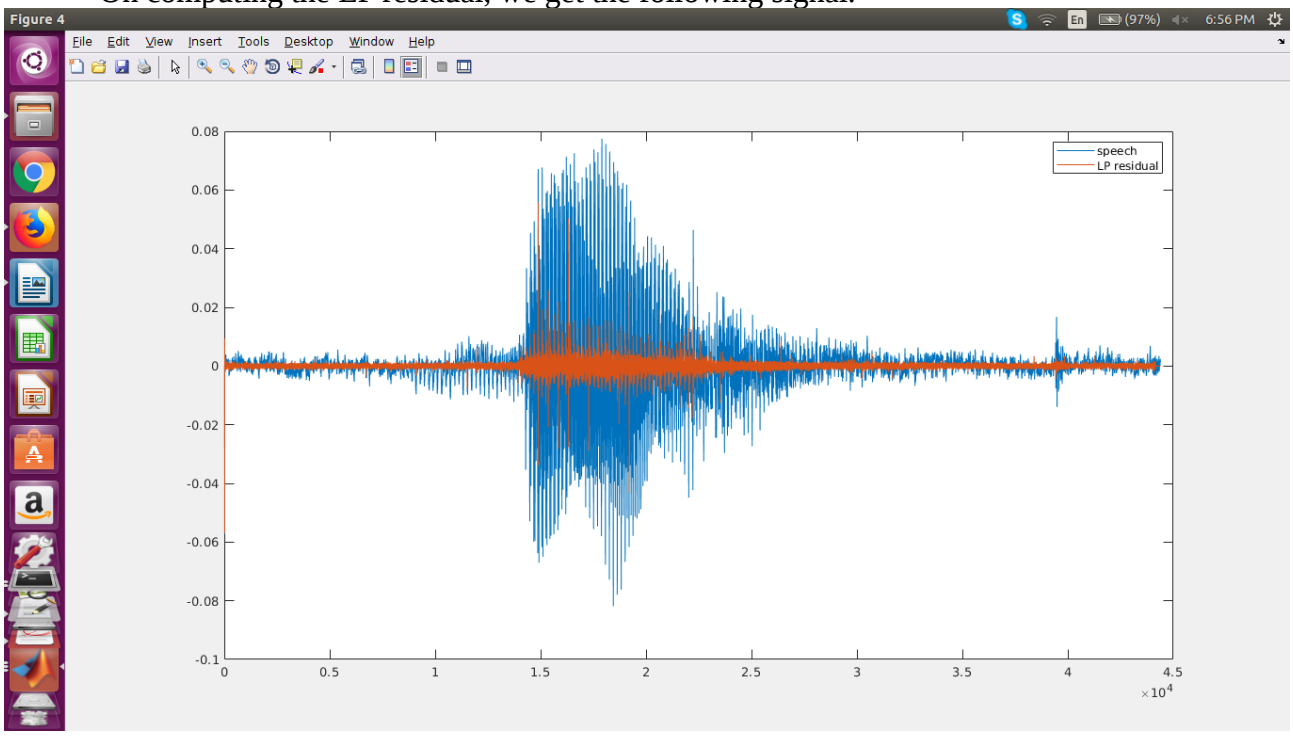
We further enhance the smoothed version of the Hilbert envelope of the LP residual by applying a first-order difference filter to it in order to calculate the slope values. Spurious zero crossings in this waveform are eliminated by finding the sum of slopes in an interval of 10 ms around each zero crossing and removing those which are less than 0.5 the mean value of the slope. In case we come across two successive peaks which are less than 50 ms away from each other, we eliminate the peak with the lower value. We then normalize the values to the range [0 1] in order to get the enhanced values.

We apply a first-order gaussian differentiator of length 100ms to the enhanced version of the Hilbert envelope of the LP residual. The peaks in the convolved output represent the locations of the VOP. A peak-picking algorithm is used along with a threshold to pick the points at which vowels start and the resulting plot is known as a VOP evidence plot.

The below picture shows the waveform for the speech signal in ‘na.wav’.

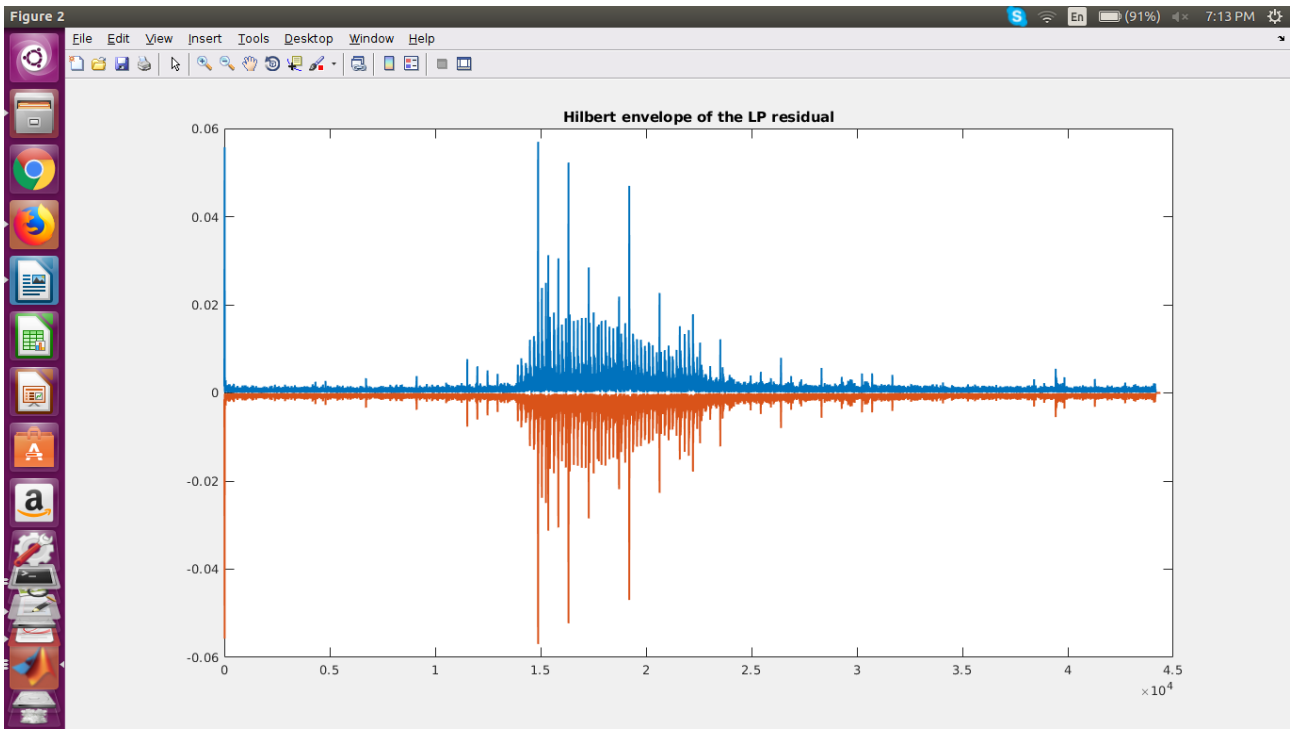


On computing the LP residual, we get the following signal.



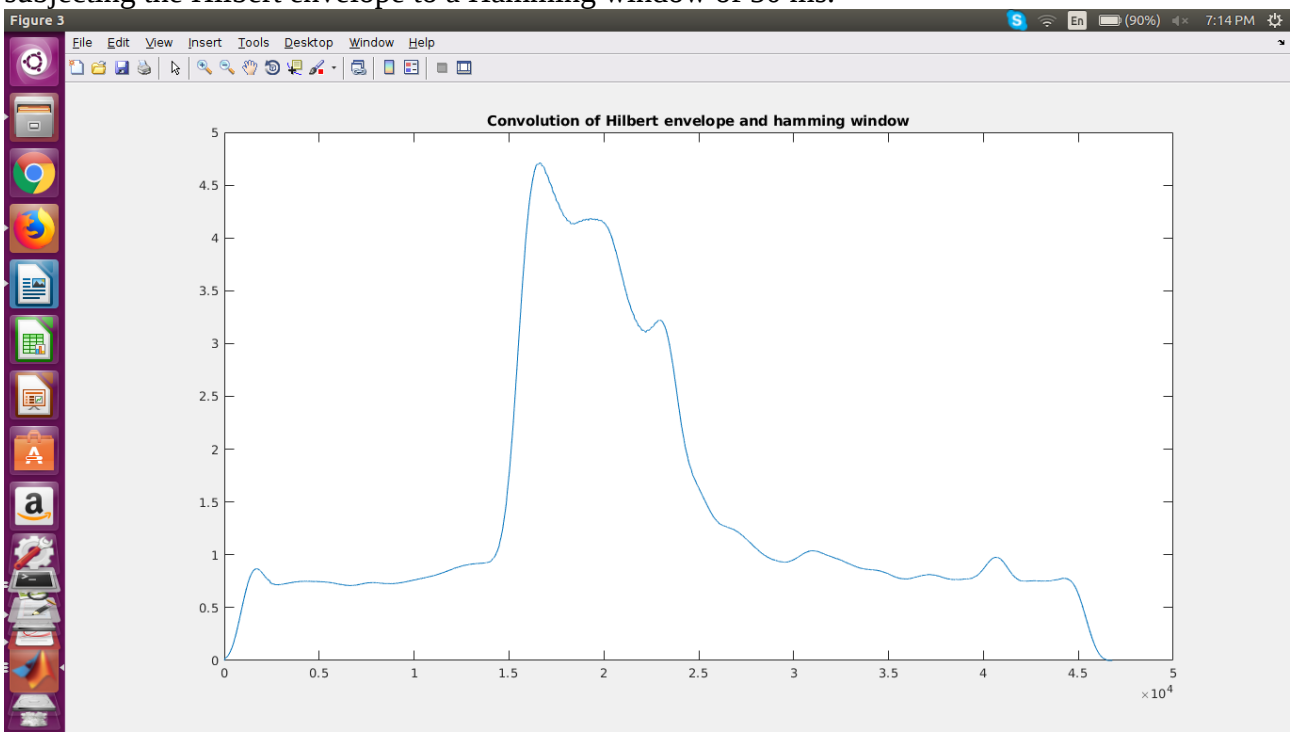
We can see that the speech signal is given by the blue plot and the LP residual by red plot. The LP residual is mainly concentrated around zero showing that the coefficients (using an order 10) do a decent job in predicting future samples.

The Hilbert envelope of the LP residual is shown below.

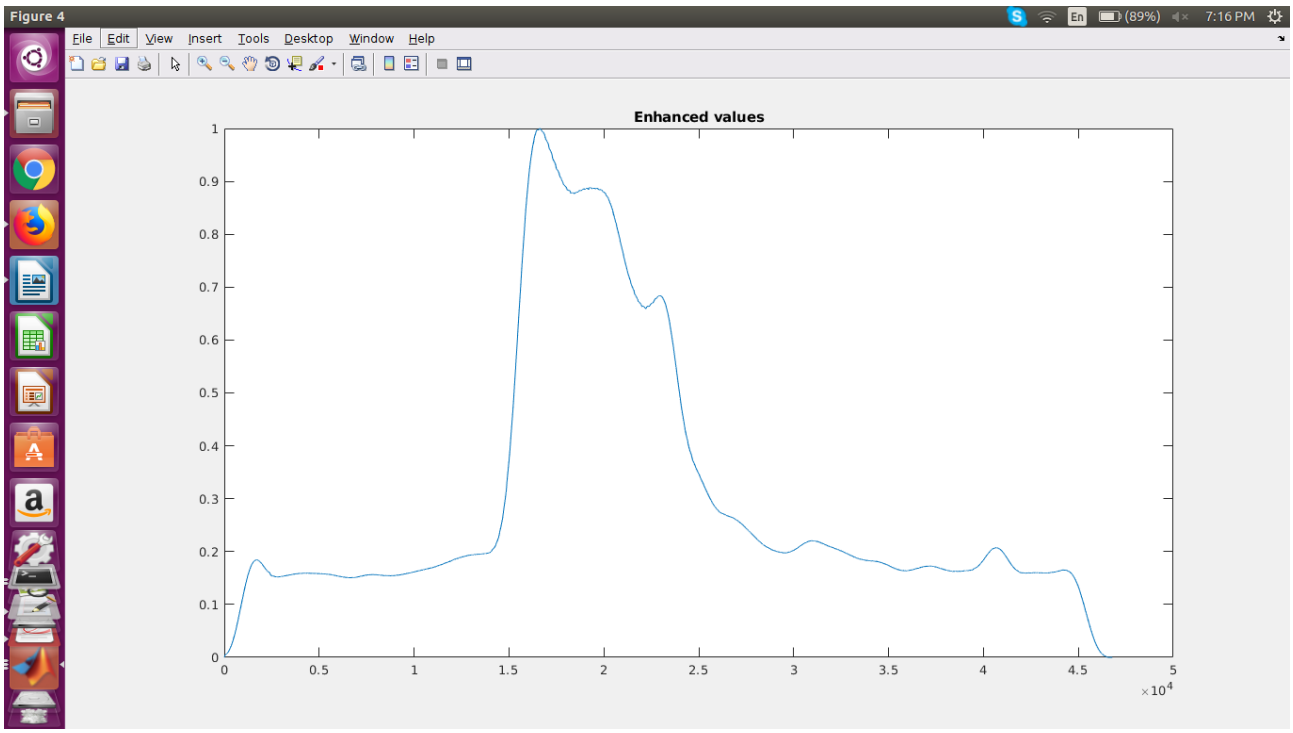


The blue part shows the upper portion whereas the red part shows the lower portion.

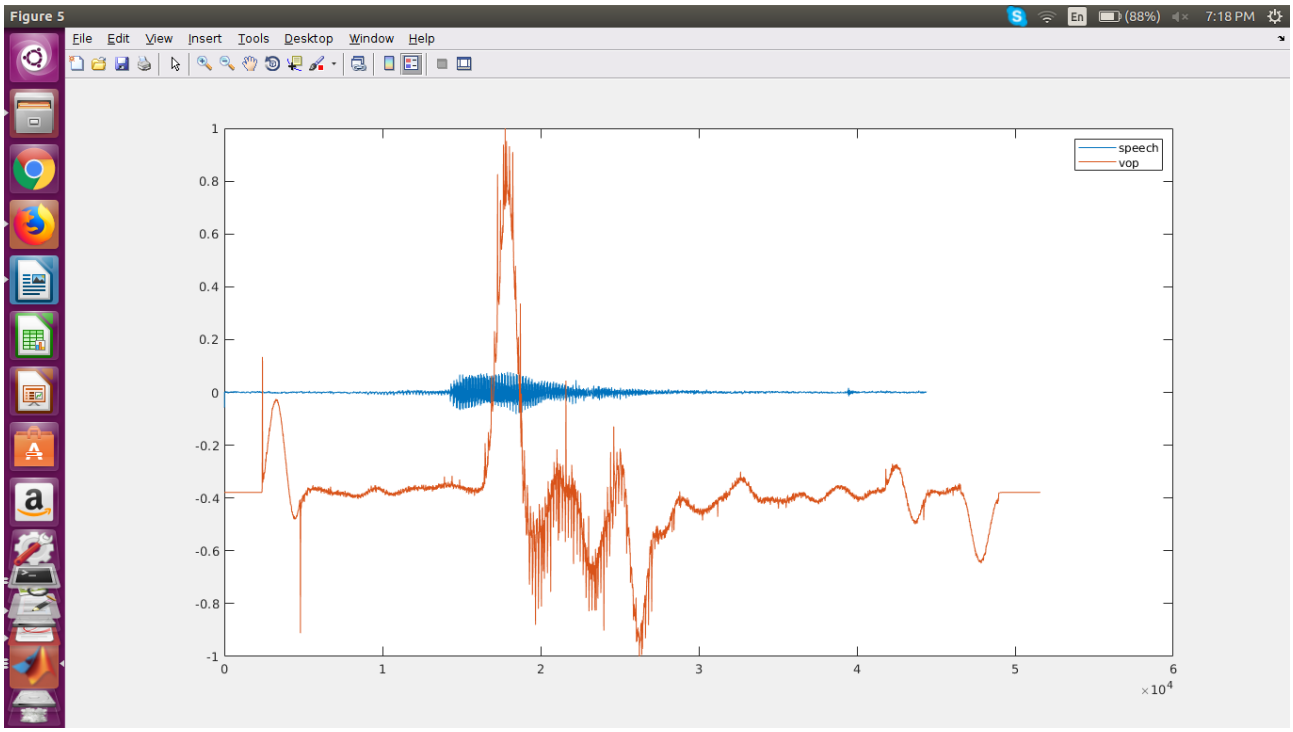
The smoothed version of the Hilbert envelope is shown below. This has been computed by subjecting the Hilbert envelope to a Hamming window of 50 ms.



The enhanced values of the Hilbert envelope has been shown below. This has been computed by applying a FOD to the smoothed Hilbert envelope and eliminating spurious peaks.

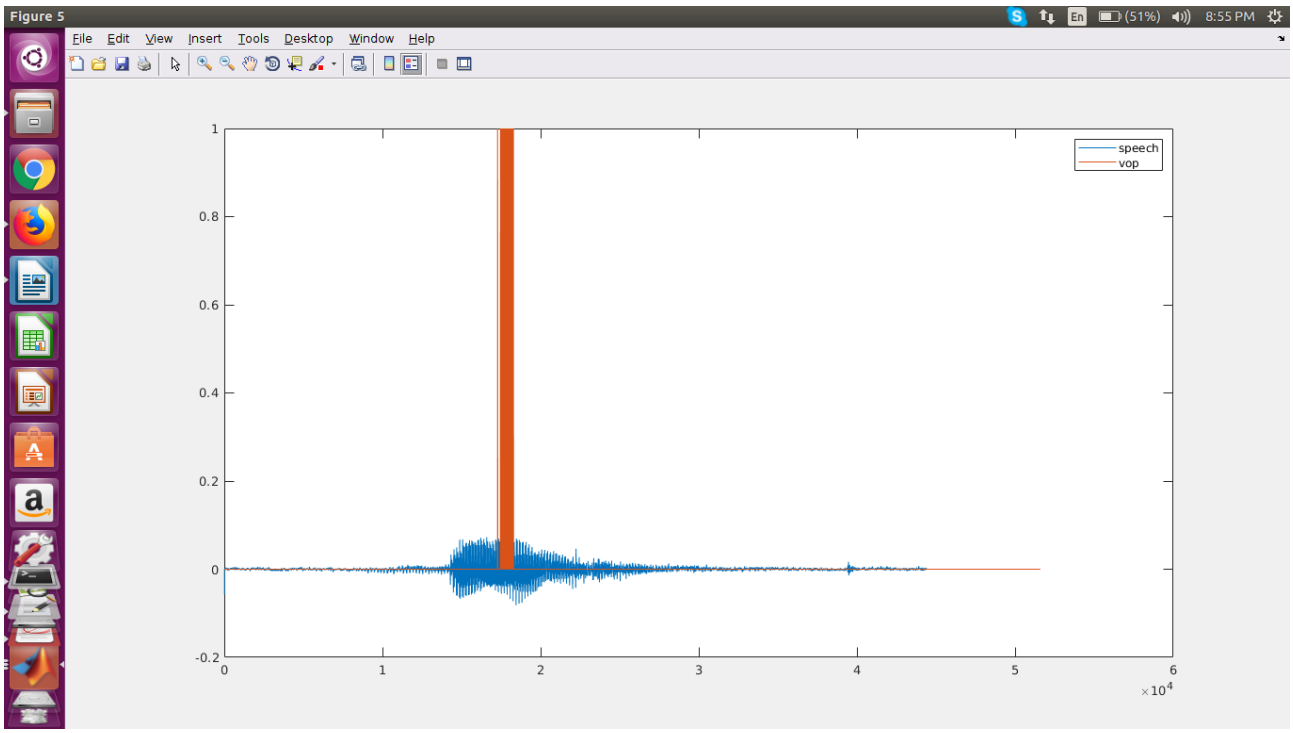


An FOGD is applied to the enhanced Hilbert envelope. The VOP evidence plot is shown below. This plot gives us the vowel likeliness contour as the peaks are the likely locations of vowels.



This VOP has been generated for the utterance of the syllable ‘na’. This seems accurate as there is one very large peak which denotes the start of the vowel ‘a’.

Using a threshold of 0.5, we get the below graph.



This single solid line corresponds to the starting of the vowel ‘a’ in the syllable ‘na’. This gives us the VOP location for the ‘na’ syllable.