**Naila Fatima**
**201530154**
**Assignment 6**

**Question 1**

*a)* We know that in order to produce speech, we require the excitation source (which gives us the energy utilized in speech) as well as the vocal tract system (which causes variation in the sounds produced). Speech is a combination of vocal tract system characteristics and excitation source characteristics. Air which is exhaled acts as the excitation source and speech is produced during the exhalation of air. Lungs and associated structures provide required energy for this action and the vocal folds inside the larynx are the main excitation source. Respiration is the process in which the lungs provide the energy source and phonation is the process in which vocal folds convert this energy into audible sounds. We know that the shape of the vocal tract affects characteristics such as formants and pitch and is dependent on the positions of the articulators (such as tongue, lips, etc). On changing the vocal tract shape, we change the system characterisitcs. It should be noted that articulation is the process in which the position of articulators determines the speech produced. We can observe that respiration is the stage where the vocal tract gains its energy.

*b)* The glottal volume velocity (GVV) represents the process of vocal fold vibrations. It models the parameters related to vocal fold vibrations such as the rate at which air is exhaled, rate at which the vocal cords open and close as well as how the air is chopped. The parameters associated with a GVV waveform are useful in applications such as detection of loudness and emotion recognition. The parameters include opening phase, closing phase, closed phase, slope of opening phase, slope of closed phase and ratio of slope of opening phase to slope of closed phase.

The EGG and the GVV both capture information related to the excitation source of speech. They are different as the EGG records the activity of the glottis whereas the GVV is a rough estimation of glottal activity. The EGG tends to be smoother when compared to the GVV.

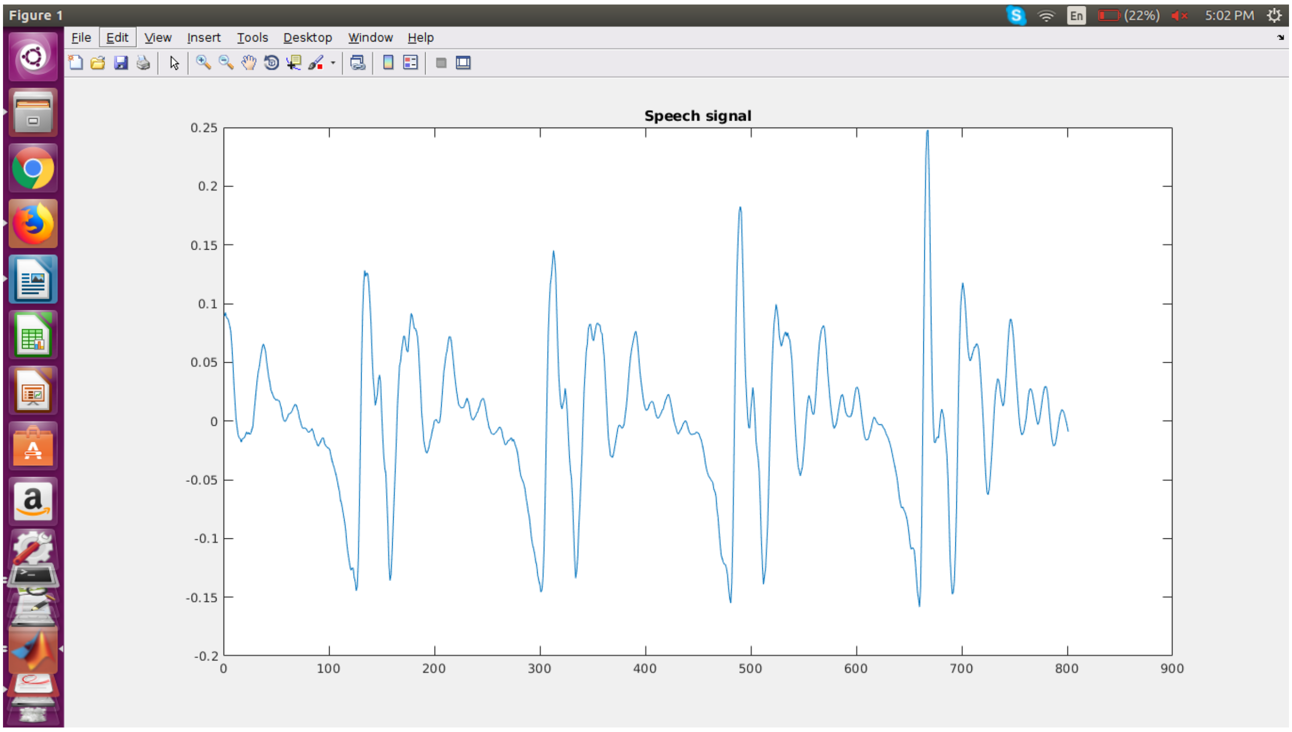
*c)* The glottal volume velocity waveform can be extracted from the speech signal waveform by using linear prediction (LP) analysis. This is done by using the inverse filtering where a speech signal is passed through an analysis filter which is constructed by using the coefficients $a_k$ computed from LP analysis.

In order to find the GVV from the speech signal waveform, we have to find the LP coefficients in all the voiced regions of the signal. After finding the LP coefficients $a_k$, we have to construct the analysis filter. The analysis filter is given by

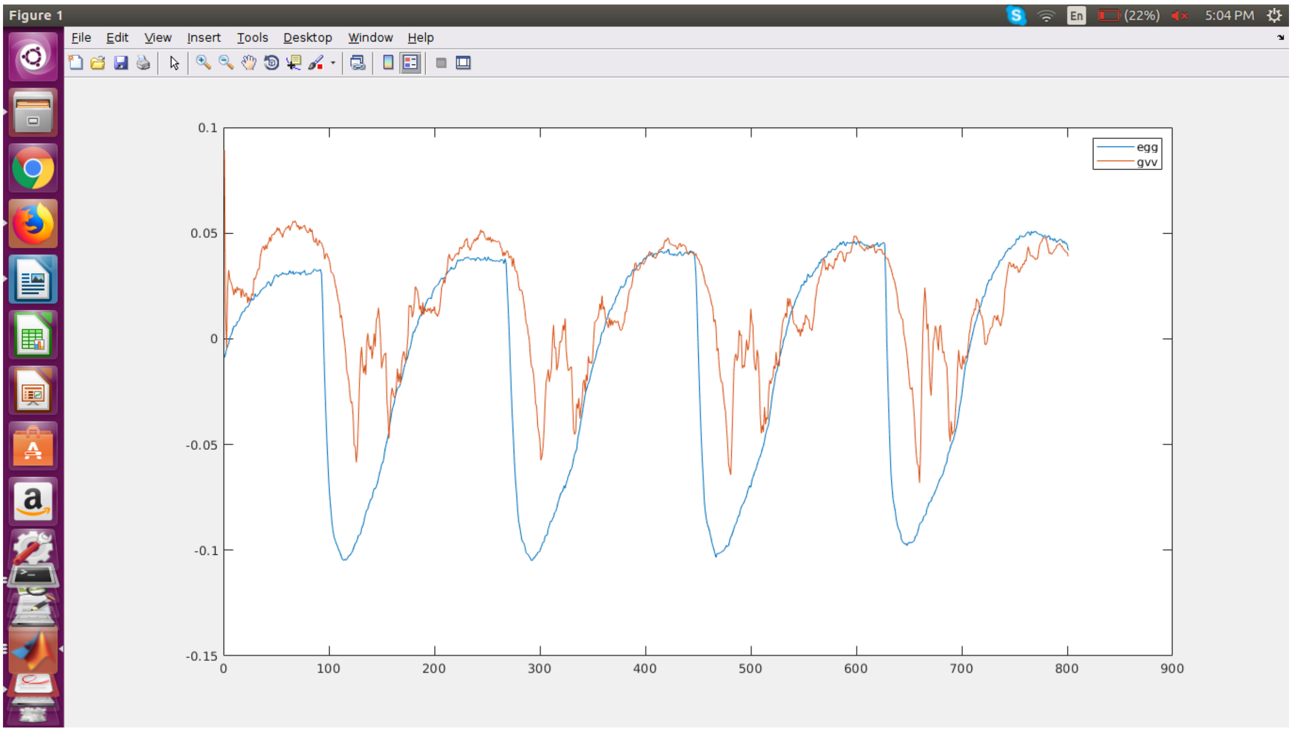
$$H(z) = 1/(1-\Sigma a_k z^{-k}) = E(z)/S(z)$$

The application of the analysis filter to the speech signal gives us the LP residual. We then pass the LP residual through an integrator in order to nullify the effect of the radiator. The output signal is the GVV as it gives us the response of the glottis.

Using the wav file '*arctic_a0008.wav*' and the code in *q1c_2.m*, we get the following results. The speech signal is shown below for a voiced region.

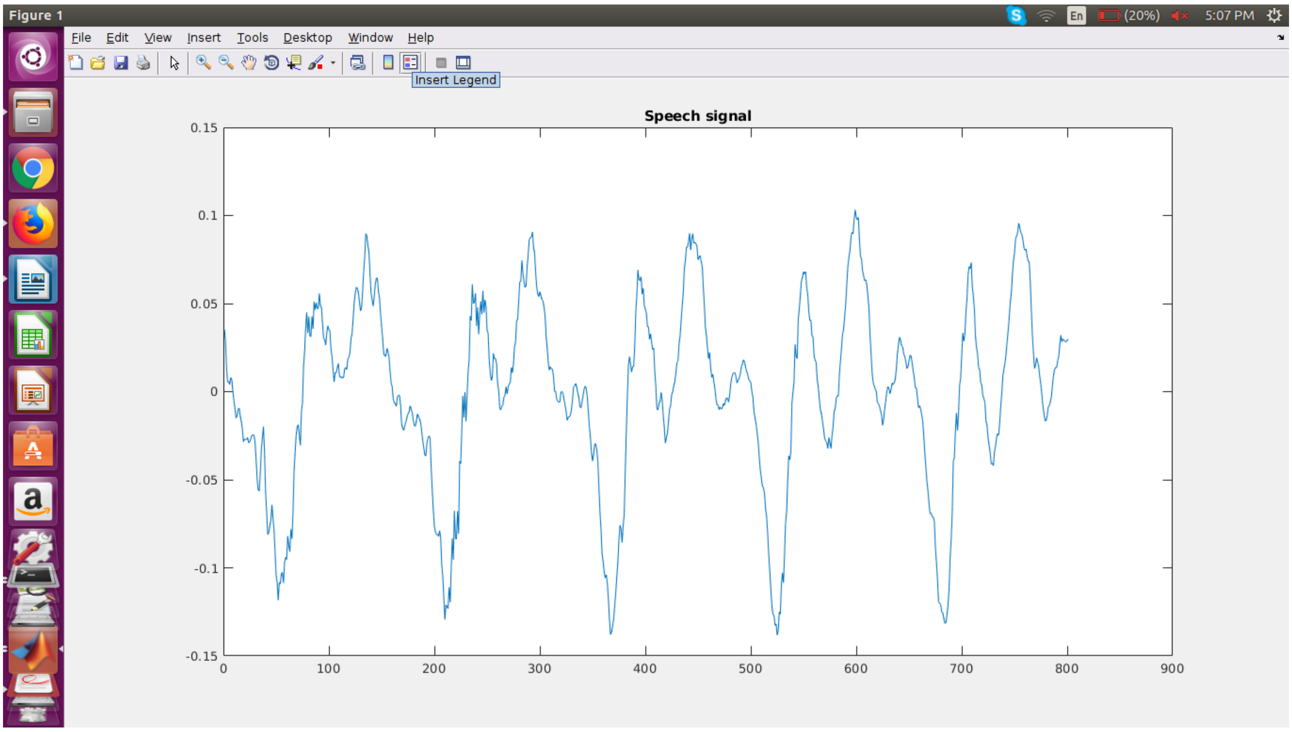


For this signal, the corresponding EGG signal (which is taken as the ground truth) and GVV waveform are shown below.

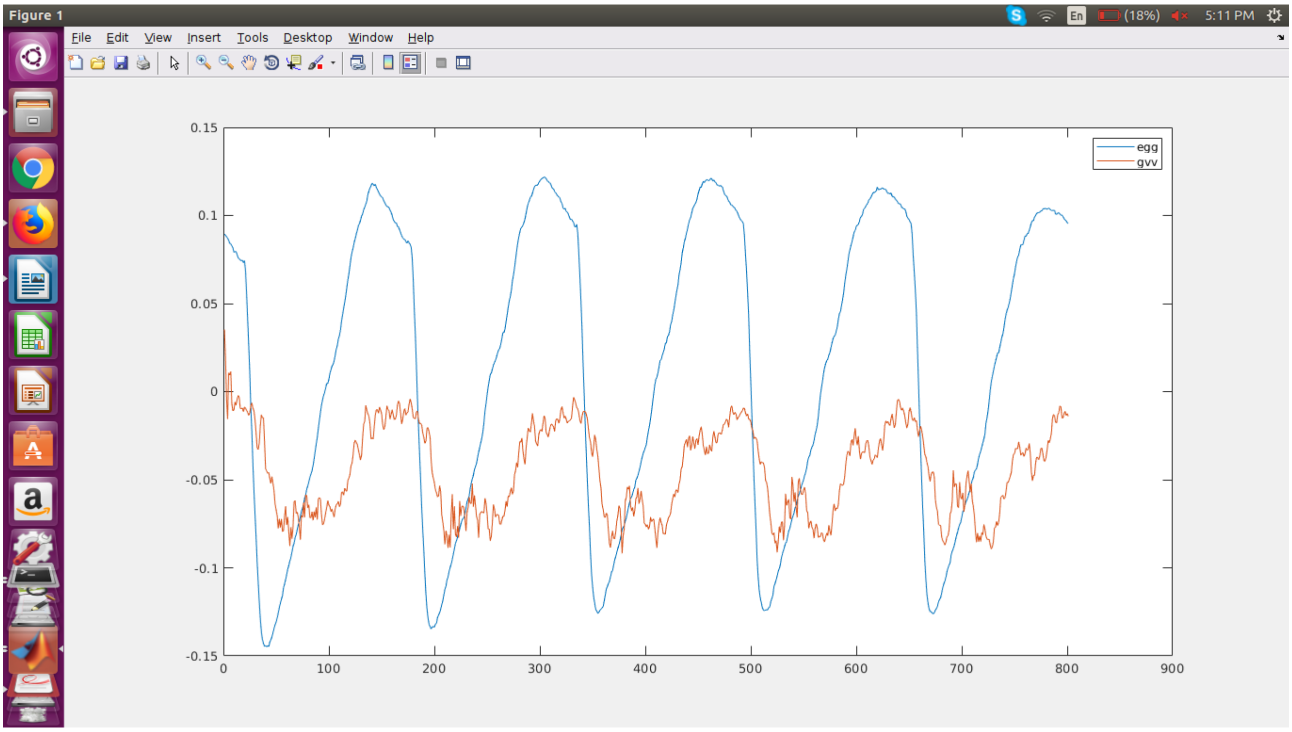


The blue plot is for the EGG signal (which is ground truth) and the red plot is for the estimated GVV waveform. We can see that the two plots are very similar to one another, just how they should be. For finding the LP coefficients, we have used the order 8. We can observe that the GVV is not equivalent to the EGG signal but it is very similar as it mimics its shape.

The corresponding results for the audio file ‘*arctic_a0033.wav*’ are shown below. The speech signal for a voiced region has been plotted below.



The corresponding EGG and GVV signals are shown below.



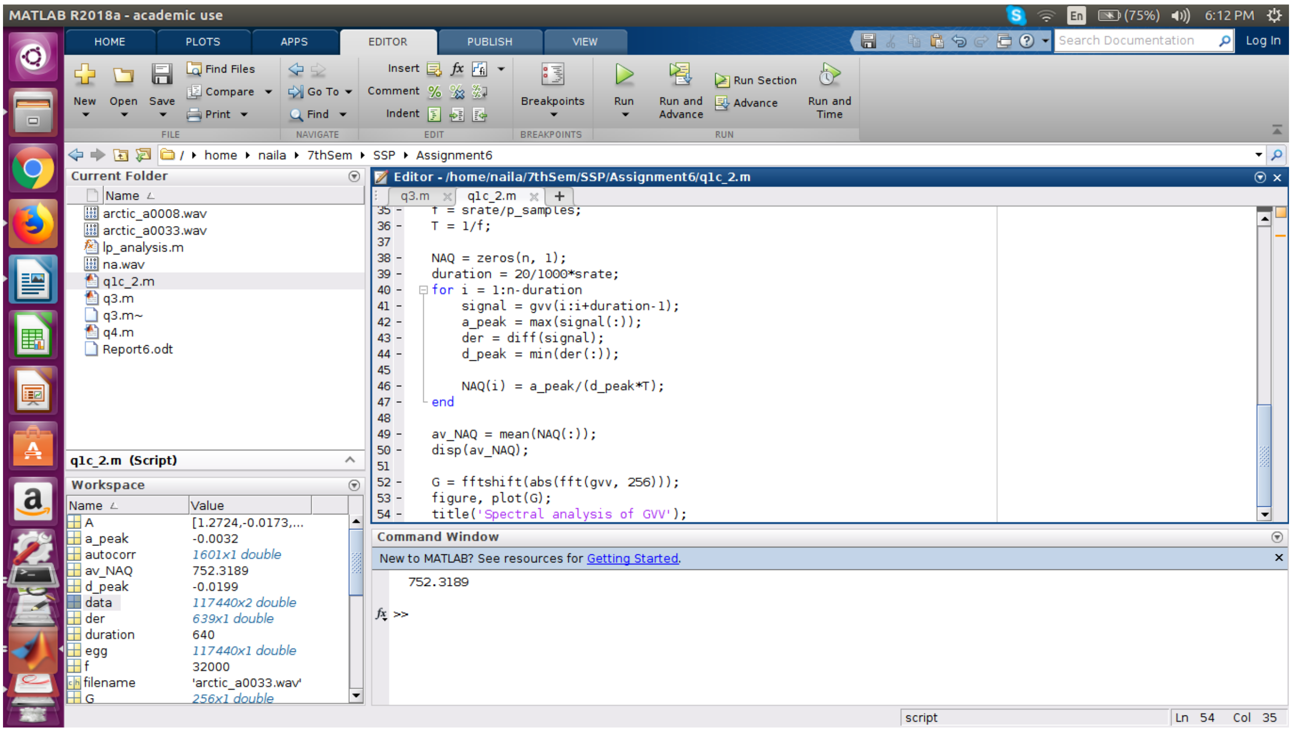
The blue plot is the EGG signal (ground truth) whereas the red plot is the GVV waveform. We can observe that the two waveforms are very similar in shape and periodicity.

## Question 2

For the glottal waveform extracted from a speech signal, we can find the parameters average Normalized Amplitude Quotient (NAQ) as well as H1-H2. We know that the GVV waveform contains excitation source information and certain features can be extracted from it. The types of features which can be extracted from a GVV waveform are time-based parameters, amplitude-based parameters and spectral domain-based parameters. The time-based parameters include open quotient, speed quotient, closing quotient, closed quotient and return quotient. The amplitude-based parameters include Normalized Amplitude Quotient (NAQ) which is the ratio of the peak of amplitude glottal flow to the product of glottal cycle time period and the minimum peak of the derivative of the glottal flow.

NAQ = ac/(d_peak * T) where ac is the peak of amplitude glottal flow, T is the glottal cycle and d_peak is the minimum peak of the derivative of the glottal flow.

For the wavefile ‘*arctic_a0008.wav*’, the average NAQ has been computed by averaging the NAQ values for blocks of length 20 ms. As shown below, the average NAQ for this file is around 752.



arctic_a0008.wav
arctic_a0033.wav
lp_analysis.m
na.wav
q1c_2.m
q3.m
q3.m~
q4.m
Report6.odt

q1c_2.m (Script)

Editor - /home/naila/7thSem/SSP/Assignment6/q1c_2.m

```
f = srate/p_samples;
T = 1/f;

NAQ = zeros(n, 1);
duration = 20/1000*srate;
for i = 1:n-duration
    signal = gvv(i:i+duration-1);
    a_peak = max(signal(:));
    der = diff(signal);
    d_peak = min(der(:));

    NAQ(i) = a_peak/(d_peak*T);
end

av_NAQ = mean(NAQ(:));
disp(av_NAQ);

G = fftshift(abs(fft(gvv, 256)));
figure, plot(G);
title('Spectral analysis of GVV');
```

Workspace

| Name | Value |
|---|---|
| A | [1.2724,-0.0173,... |
| a_peak | -0.0032 |
| autocorr | 1601x1 double |
| av_NAQ | 752.3189 |
| d_peak | -0.0199 |
| data | 117440x2 double |
| der | 639x1 double |
| duration | 640 |
| egg | 117440x1 double |
| f | 32000 |
| filename | 'arctic_a0033.wav' |
| G | 256x1 double |

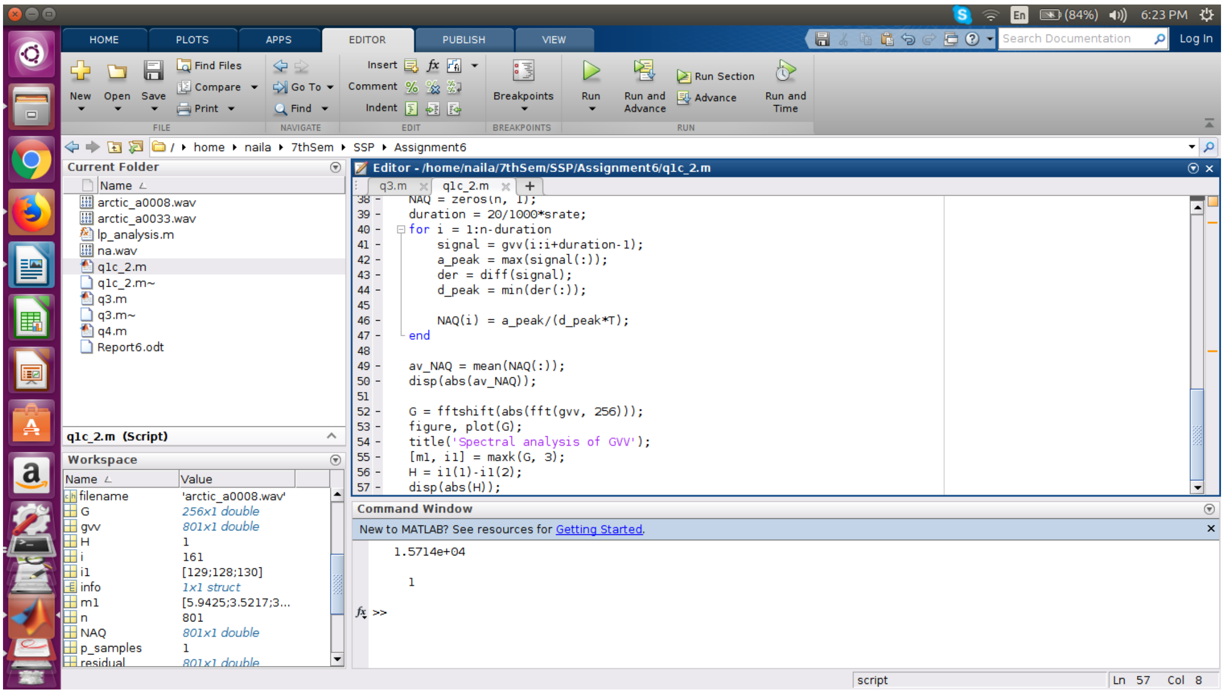
Command Window

New to MATLAB? See resources for Getting Started.

```
  752.3189

>>
```

script Ln 54 Col 35

For the wavefile ‘*arctic_a0008.wav*’ , the average NAQ has been computed using 20 ms blocks. The average NAQ is 1.57 * $10^4$.



arctic_a0008.wav
arctic_a0033.wav
lp_analysis.m
na.wav
q1c_2.m
q1c_2.m~
q3.m
q3.m~
q4.m
Report6.odt

q1c_2.m (Script)

Editor - /home/naila/7thSem/SSP/Assignment6/q1c_2.m

```
NAQ = zeros(n, 1);
duration = 20/1000*srate;
for i = 1:n-duration
    signal = gvv(i:i+duration-1);
    a_peak = max(signal(:));
    der = diff(signal);
    d_peak = min(der(:));

    NAQ(i) = a_peak/(d_peak*T);
end

av_NAQ = mean(NAQ(:));
disp(abs(av_NAQ));

G = fftshift(abs(fft(gvv, 256)));
figure, plot(G);
title('Spectral analysis of GVV');
[m1, i1] = maxk(G, 3);
H = i1(1)-i1(2);
disp(abs(H));
```

Workspace

| Name | Value |
|---|---|
| filename | 'arctic_a0008.wav' |
| G | 256x1 double |
| gvv | 801x1 double |
| H | 1 |
| i | 161 |
| i1 | [129;128;130] |
| info | 1x1 struct |
| m1 | [5.9425;3.5217;3... |
| n | 801 |
| NAQ | 801x1 double |
| p_samples | 1 |
| residual | 801x1 double |

Command Window

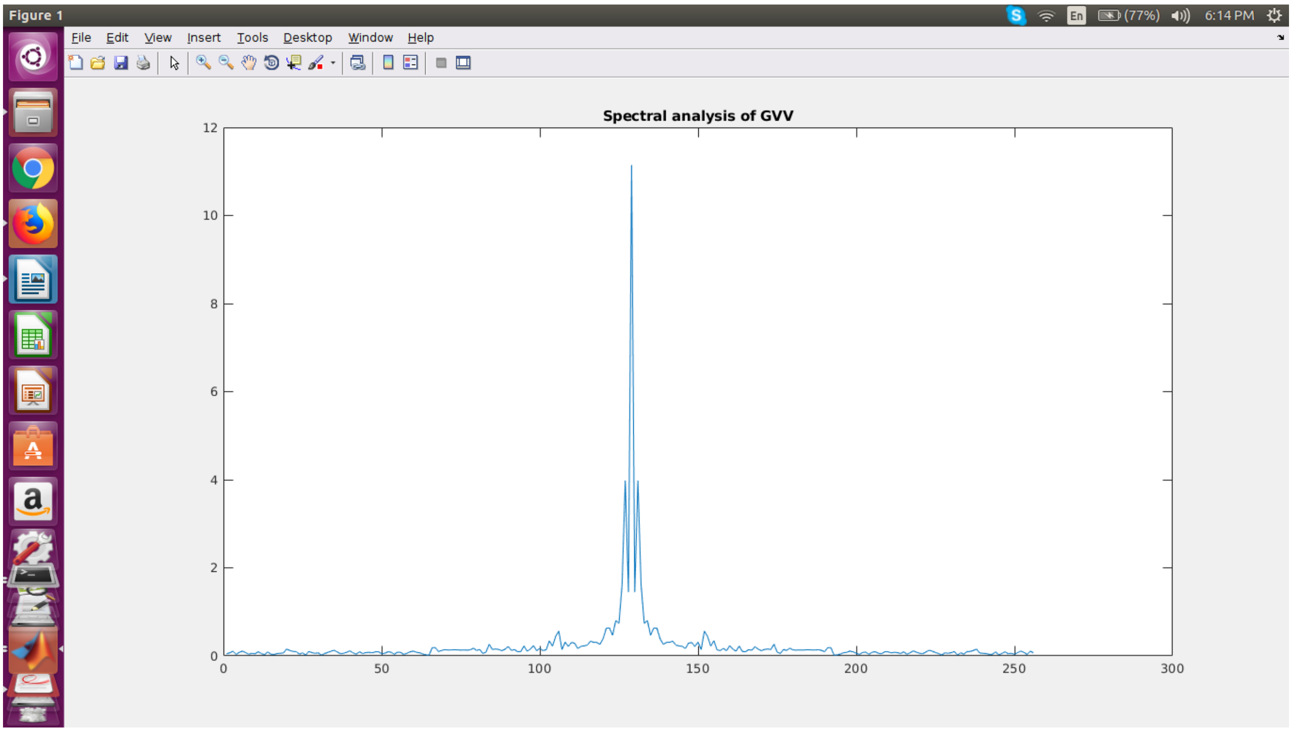
New to MATLAB? See resources for Getting Started.

```
  1.5714e+04

     1

>>
```

script Ln 57 Col 8

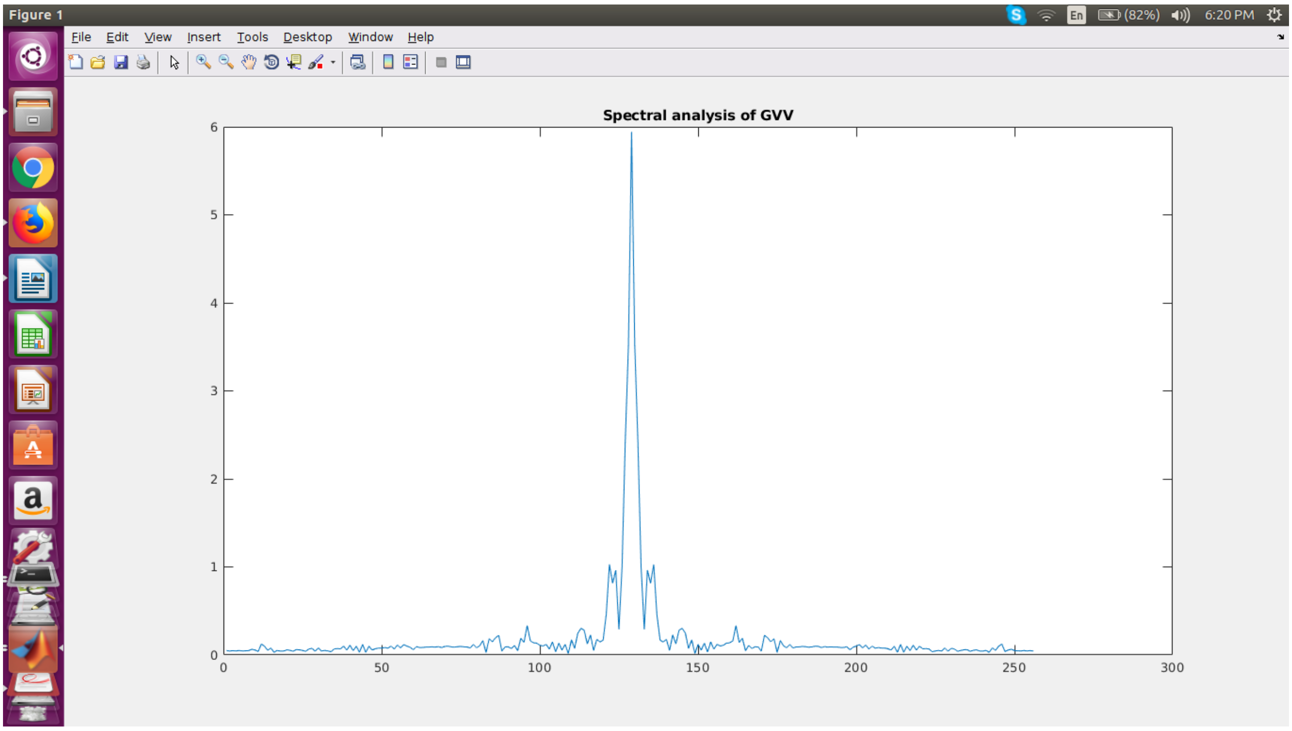
H1-H2 is a spectral-domain parameter which can be extracted from the GVV signal. This can be computed by finding the first (H1) and second (H2) dominant peaks in the magnitude response of the GVV signal and taking their difference, H1-H2.

For the wavefile '*arctic_a0033.wav*', the magnitude response of the GVV signal is shown below.



The first two dominant peaks are at 129 and 131 so the H1-H2 parameter will be 2.

For the wavefile '*arctic_a0008.wav*', the magnitude response of the GVV signal is shown below.

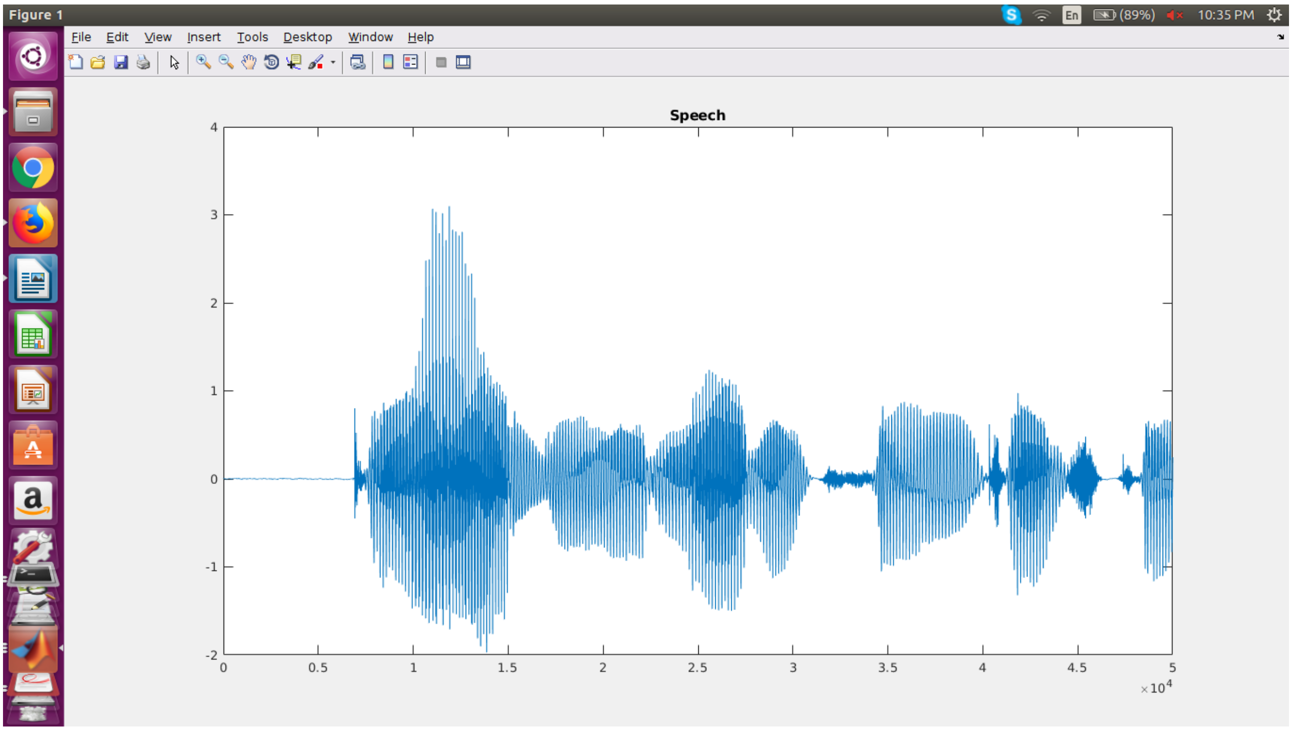


The dominant peaks are at 130 and 129 so the H1-H2 parameter will be 1.

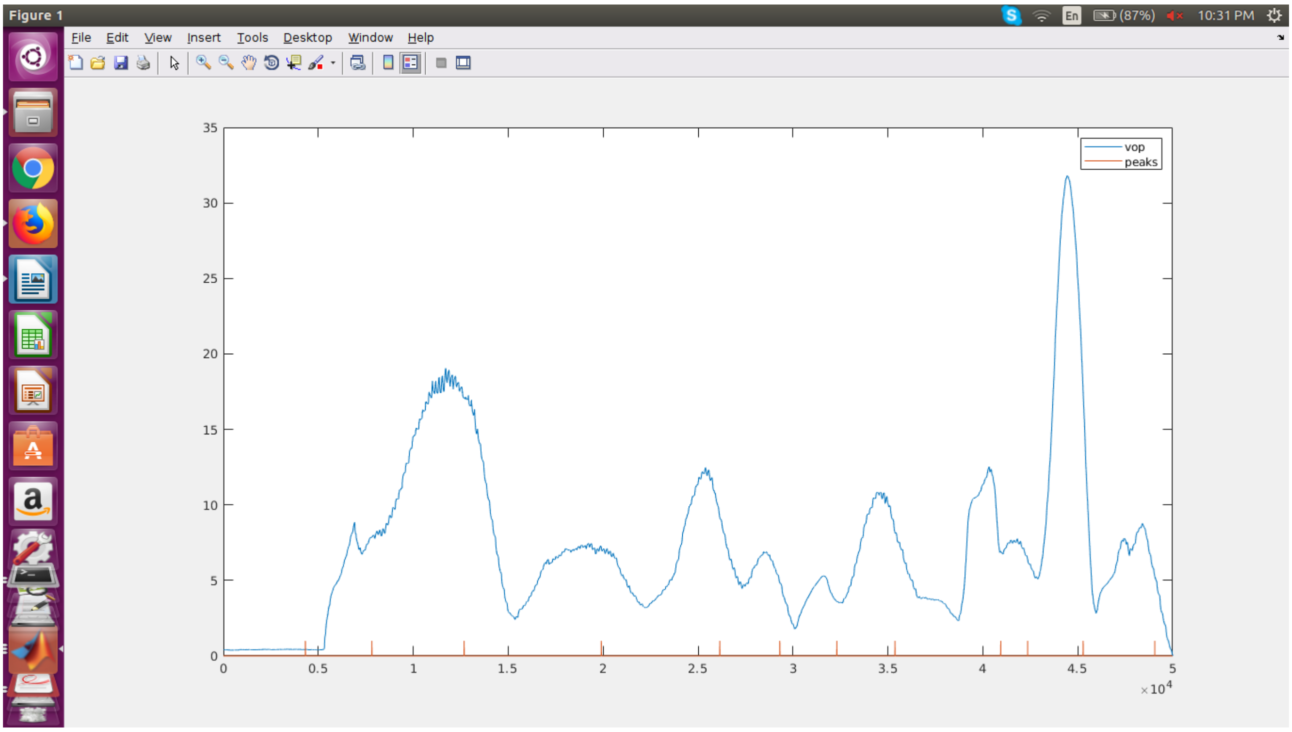
**<u>Question 3</u>**

In the paper '*Extraction and representation of prosodic features for language and speaker detection*' states a method to extract VOP based prosodic features. In this method, a vowel is taken as an anchor point to extract prosody features. The VOPs are detected by finding the Hilbert envelope of the LP residual of the speech signal and then applying the Gabor filter to this envelope and taking the sum of the product for each sample shift. This gives us the VOP evidence plot. A peak picking algorithm is applied to the VOP evidence plot to find the VOPs. The VOPs are vowel onset points which indicate the beginning of each vowel.

The speech signal waveform for the '*arctic_a0008.wav*' file is shown below.



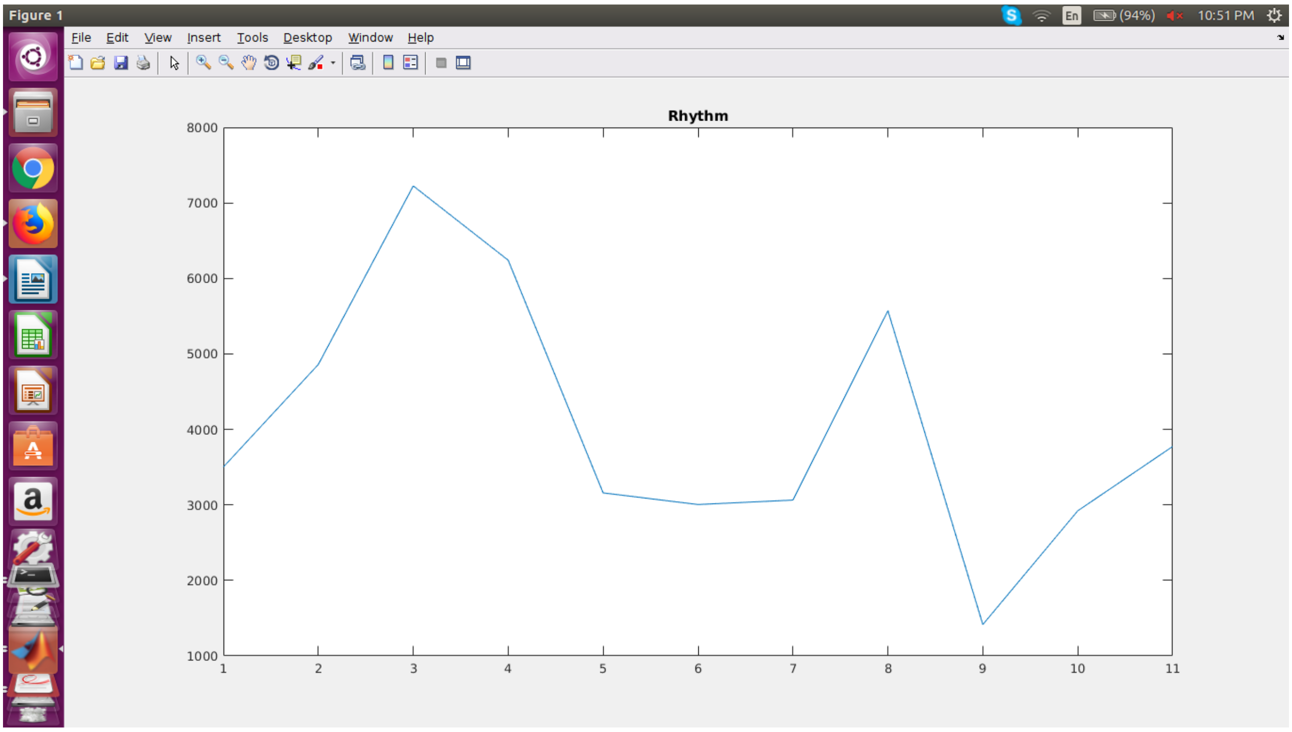
The below plot shows the VOP evidence plot which was obtained by taking the Hilbert envelope of the LP residual of the signal and applying Gabor filter and then a peak picking algorithm to it.



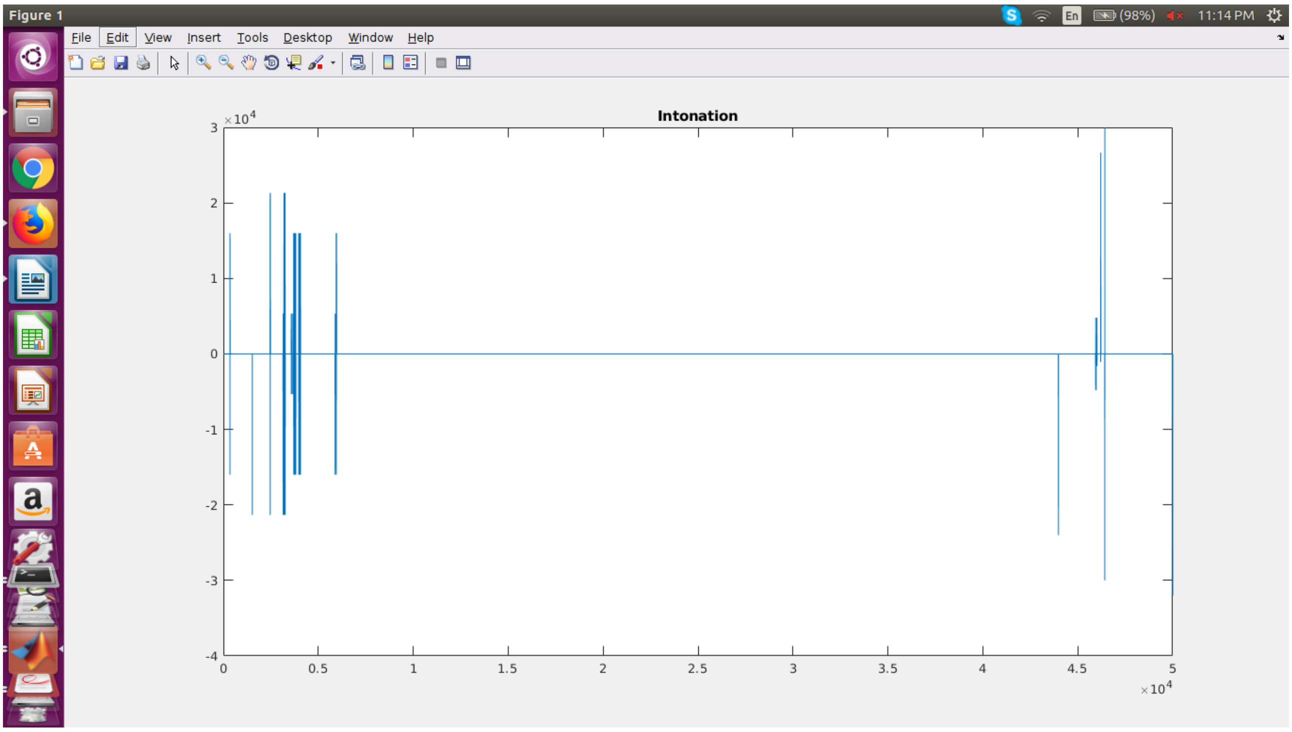
We can see that the VOP evidence plot is indicated by the blue plot. The peaks of the VOP evidence plot are the vowel onset points (VOP) and these are present wherever there is a 1 in the peaks plot (which is red). The peaks plot is one wherever there is a peak in the evidence plot and is zero otherwise.

The prosodic features- rhythm, stress and intonation- can be found out by using the VOPs as well as the F0 contour plot.

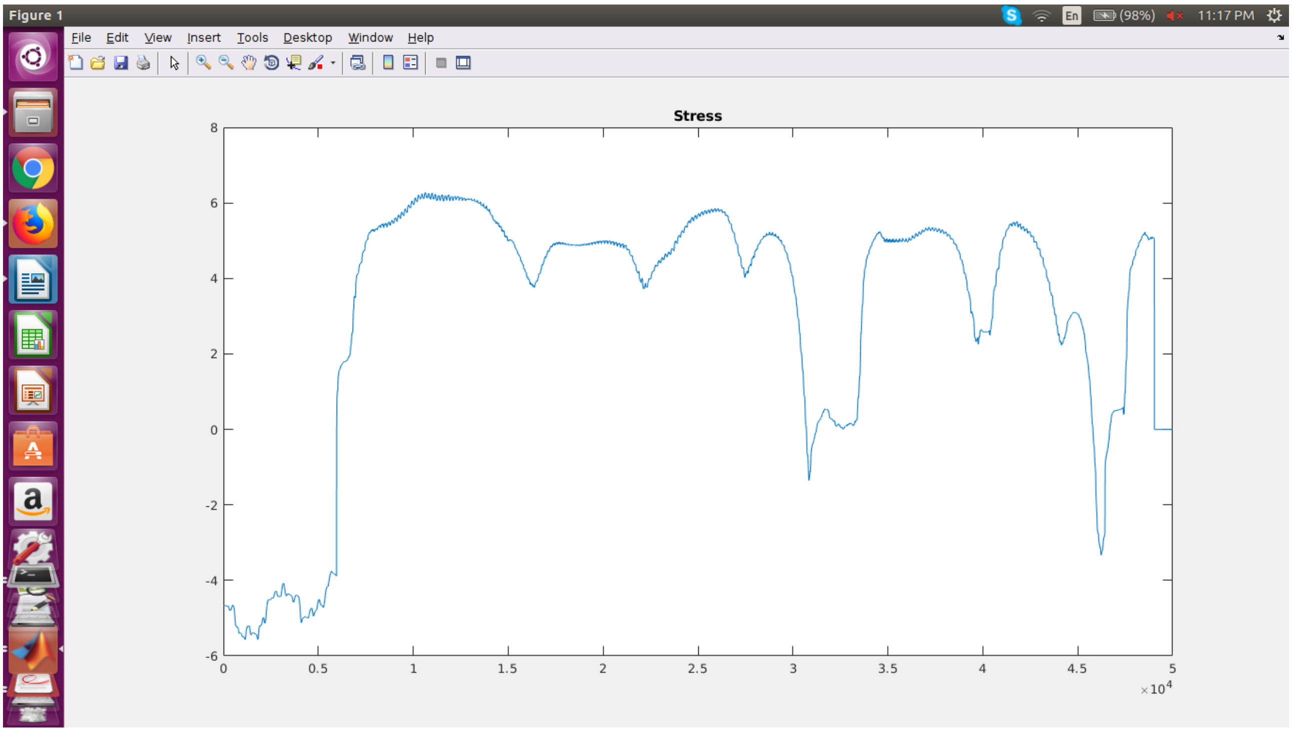
The rhythm can be found by computing the distance between successive VOPs. The plot for rhythm is shown below.



The intonation can be obtained by finding the change in the F0 contour plot. The intonation plot for the wavefile is given below.



The stress can be obtained by finding the log of the energy in a particular voiced region. The plot for the stress is shown below.



## Question 4

*a)* Zero time windowing (ZTW) is a method used to overcome the time and spectral resolution issues which occur when using windows. It is used to successfully estimate vocal tract system characteristics for short segments of speech. The ZTW method employs the use of an impulse-like window function which provides temporal resolution. In order to have spectral resolution, a group delay function is incorporated in this method. The impulse-like window is an approximation to doing integration in the frequency domain. The loss in spectral domain (which occurs because we use a highly decaying window) is overcome by using a group delay function.

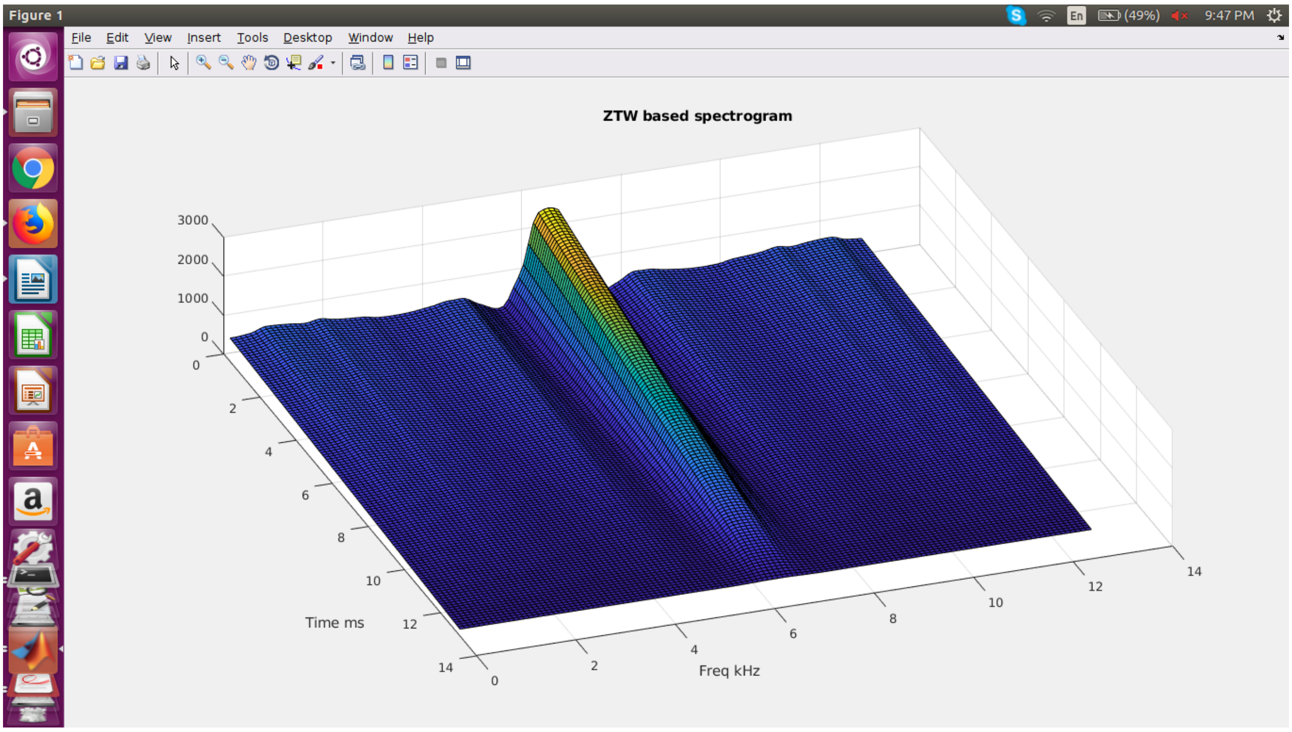
It is useful in extracting system information of speech as it minimizes the effects of the window's own spectrogram. We know that when using a window, the window size, window shape and the position at which the window is applied all matter. The main problem with using a window is that it has its own spectrogram which affects the spectrogram of the signal when it is applied to it. We know that with a rectangular window, the sinc lobes (which are not wanted) as well as the overlap are the main problems. In order to minimize overlap, we have to use a window with a larger window length. But this causes us to lose temporal resolution. In order to overcome this, we use ZTW which can be applied to the signal without losing much of the temporal and spectral resolutions. Since ZTW allows us to use the signal without much loss of information, we can find parameters such as formants and pitch (which are vocal tract characteristics) with greater accuracy as there is less information which is lost/changed in the signal. ZTW is essential as it allows us to work with a signal in either domain without losing much information.

*b)* In order to obtain a Zero Time Windowing based spectrogram, I have done the following steps. A differenced signal s(n) is computed from the speech signal. We take M samples of the signal and I have used M = 50.  The DFT length N is chosen such that N is not less than M. I have chosen N to be 128. A window w(n) is computed such that w(n) is 0 for n = 0 and is $1/4\sin^2(pi*n/2N)$ for values of n between 1 and N-1. The windowed signal x(n) is computed by multiplying s(n) with w(n). The NGD function of x(n) is then computed as shown below.

$$NGD[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k]$$

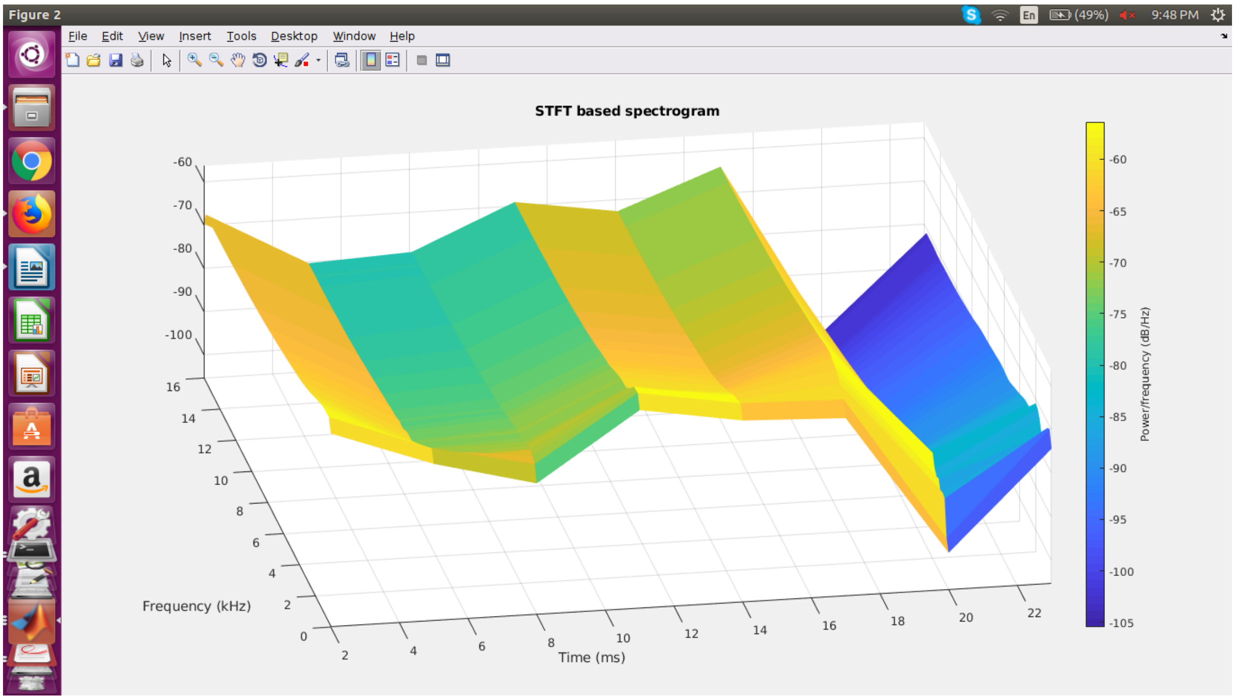
where $X[k] = X_R[k]+jX_I[k]$ is the N-point DFT of x(n) and $Y[k] = Y_R[k]+jY_I[k]$ is the N-point DFT of y(n) = nx(n)

The ZTW based spectrogram is shown below for the wavefile '*arctic_a0008.wav*'.



We can observe that there are no sharp discontinuities in this plot. All changes are gradual and smooth. We can also observe that as the time increases, the magnitude of the spectrogram reduces.

The STFT based spectrogram is shown below.



We can observe that when this is compared to the ZTW spectrogram, there are several sharp discontinuities (which are shown by sharp edges). This case probably occurs as we lose spectral information when using a narrow window in the time domain. In the ZTW spectrogram, the values were smoother and more gradual. There is less loss of information in the ZTW case.

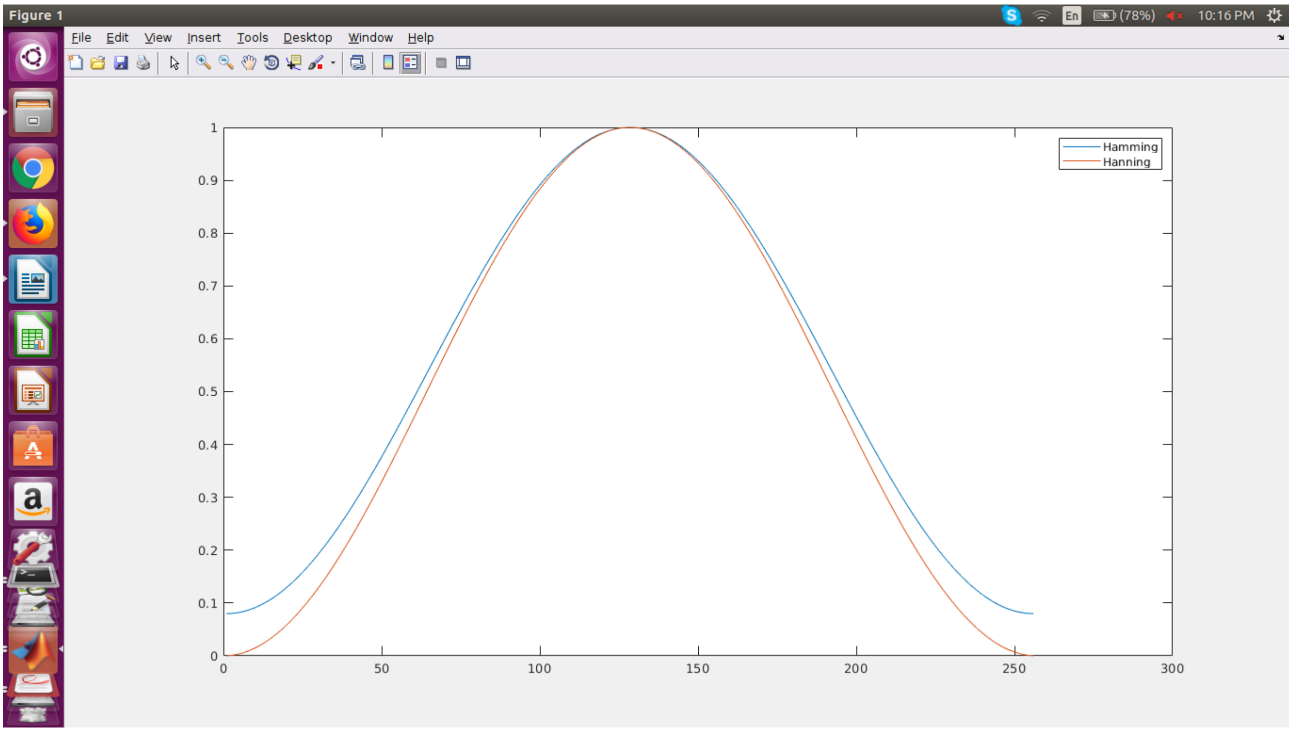
The two are similar in the sense that as time increases, the magnitude of the spectrogram tends to decrease.

*c*) Windows effect the spectral information of a signal. Ideally, we want a window to have minimum effect on the spectral content of the signal. In order to do this, the frequency response of the window should be close to impulse as convolution with it will give the same signal.

The Hamming and Hanning windows are both sinusoidal in shape. Both windows have a wide peak but low side lobes. The Hanning window is a better option as it touches zero at both ends thereby eliminating discontinuity. Since the Hamming window does not touch zero, it displays a little discontinuity. Due to this, the Hamming window cancels the nearest side lobe better than the Hanning but cancels the other side lobes in a worse manner. For the Hamming window, we observe that there is some cancellation of the side lobes and that the width of the main lobe is doubled.

The Hanning window is an improvement to the rectangular window (which suffers from sinc lobes and overlap) as it has lobes with smaller amplitudes thereby having a smaller effect on the spectral information of a signal. However, the width of the primary lobe is larger and the primary lobe has a significant ripple (equiripple error) but the benefits outweigh the disadvantages in this case. The Hanning window is used for analyzing transients longer than the time duration of the window and for general-purpose applications. It achieves side lobe reduction by superposition.

The Hamming and Hanning windows are shown below in the time domain.



The Gaussian window is the bell-shaped window. The side lobes for this window are caused by the truncation of the window and are very small compared to the main lobe. The side lobes for this window are small and the main lobe can be estimated as a parabola.

High decaying windows are similar to those which have a transform H(z) = 1/(1-az$^{-1}$) where a is low. We should observe that when a is low, we lose the frequency resolution, so this is not preferred. When a is high (close to 1), we lose the time resolution, and this is another advantage. High decaying windows cause us to lose spectral information as a is low.