Naila Fatima
201530154
Assignment 4
Speech Signal Processing

Question 2

a) Linear prediction (LP) analysis is a technique which allows us to predict current speech samples by using a linear weighted sum of previous samples. The formula which defines LP analysis has been given below.

$s_p(n) = \Sigma\ a_k s(n-k)$ where k ranges from 1 to p.

In the above notation, a_k refer to the linear prediction coefficients whereas s(n) is the actual signal and $s_p(n)$ is the predicted one. We can understand that in order to predict any sample of the speech, we require the *p* previous samples of the actual signal and the *p* LP coefficients. Note that the *p* value which is chosen is usually either 8 or 10. The error generated by LP analysis is the difference between the actual and predicted signals. The error is also known as the LP residual and our goal is to minimize this error in order to be able to predict samples more accurately. The formula to find the LP residual is given below.

$e(n) = s(n) – s_p(n)$

$= s(n) – \Sigma a_k s(n-k)$

On applying the Z-transform on both sides, we get

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

We should note that in order to get the LP residual from the speech signal, we have to pass it through an all-zeros filter (this is LP analysis). The analysis filter (which is also known as inverse filter as it is the inverse operation as the vocal tract system) is given by

$E(z)/S(z) = 1 – \Sigma a_k z^{-k}$

In order to derive the speech signal from the LP residual, we have to pass it through an all-poles filter (this is LP synthesis). The synthesis filter (which is equivalent to the vocal tract system) is given by

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

LP analysis is a segmental procedure as we need a stationary signal in order to predict future samples. We need 2-3 pitch cycles for this procedure (around 20 ms) as a smaller interval of time will not provide us with enough previous samples to predict future samples. LP analysis is a useful technique which can be used to extract characteristics like epoch locations, formants, LPCC coefficients and to separate the source and system characteristics. The various applications of LP analysis include:

1) Signal compression. This is as instead of saving the entire speech signal, we can just save the LP coefficients a_k and the LP residual e(n).

2) The spectrum of the LP coefficients a_k ends up being much smoother than the spectrum of the actual signal and we can use this for further analysis.

3) The normalized error (which is ratio between the residual energy and signal energy) tends to be lesser for voiced segments and more for unvoiced segments. This can be used to distinguish between voiced and unvoiced portions of speech thereby allowing us to use LP analysis for voice active detection (VAD).

The LP coefficients can be extracted from a given speech signal by using autocorrelation. As we want to derive LP coefficients which will allow us to accurately predict future samples, we can do this by minimising the sum of squared error.

$E(n) = \Sigma\ e^2(n)$ where e(n) is LP residual

$= \Sigma\ (\ s(n) – \Sigma a_k s(n-k))\ ^2$

To find the values of a_k where the above function is minimized, we differentiate it with respect to a_k and equate it to zero. We get

$dE/da_i = 2\ \Sigma(s(n) – \Sigma a_k s(n-k))\ (0-s(n-i)) = 0$

On rearranging, we get

$\Sigma s(n)s(n-i) = \Sigma\ \Sigma a_k s(n-i)s(n-k)$ where summation on left side is for n while that on right side is for n and k. This is equation 1.

We know that the autocorrelation sequence R(i,k) is given by the formula

$R(i,k) = \Sigma\ s(n-k)s(n-i)$ where summation is for n.

On replacing n-k with m, we get

$R(i,k) = \Sigma\ s(m)s(m-(i-k))$ where summation is for m. We can see that this is the autocorrelation between a signal and its $(i-k)^{th}$ shifted version. This means that R(i,k) = R(i-k) as they both are equivalent to finding the similarity between (i-k) shifted signals.

On putting k = 0 in the R(i,k) formula, we get

$R(i,0) = \Sigma\ s(n)s(n-i)$ where summation is for n.

We can observe that we can substitute R(i,0) on the left side and R(i,k) on the right side of equation 1. The formula becomes

$R(i,0) = \Sigma a_k R(i-k)$

If we denote R(i,0) as γ(i), then we get equations of type

$\gamma(i) = a_1 R(i-1) + a_2 R(i-2) + ... + a_p R(p-1)$ (note that autocorrelation is even function so we can neglect the negative signs). The equations for γ(i) in terms of R and a_i can be represented in terms of a matrix by the formula RA = γ, where

R is a *pxp* Toplietz matrix given by [R(0) R(1) .... R(p-1) ; R(1) R(2) .... R(p-2) ; .... ; R(p-1) R(p-2) .... R(0)] , A is a column vector given by [a_1; a_2 ; a_3 ; ...; a_p] and γ is a column vector given by [γ_1; γ_2; ...; γ_p]

Using this method, we can extract LP coefficients as we can find values of R and γ matrices by using autocorrelation and we find A = inv(R)γ, and this gives us all the LP coefficients.

b) LP residual or error signal refers to the difference between the actual signal and the one predicted by LP analysis. It tells us how well we are estimating the future samples as if the error is small, it means that we are doing a decent job whereas if it is large, it means that we need to improve our coefficients. LP residual can be extracted from the speech signal by using an LP analysis filter. The LP analysis filter maps the speech signal to its corresponding LP residual. This is as,

LP residual = $e(n) = s(n) – s_p(n) = s(n) – \Sigma a_k s(n-k)$

On applying the z-transform on both sides, we get

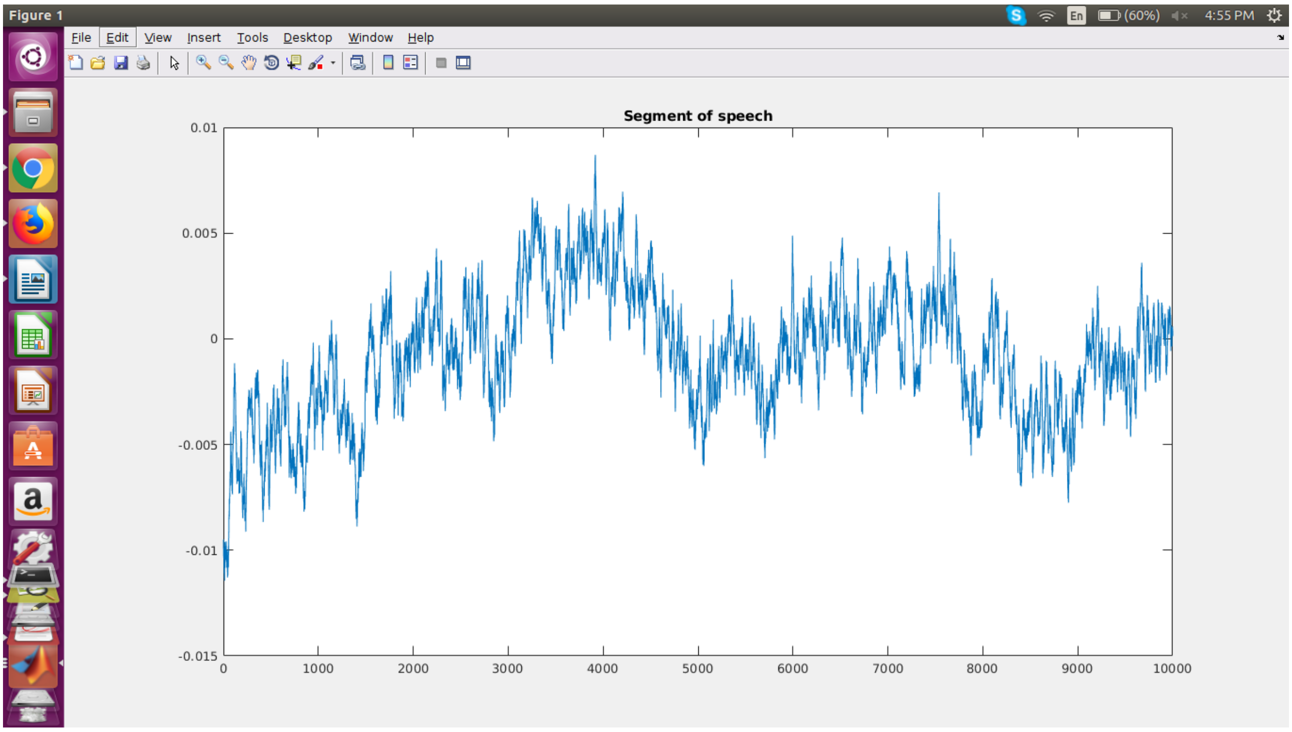
$E(z) = S(z) – \Sigma a_k z^{-k} S(z) = S(z)\ (1-\Sigma a_k z^{-k})$

The analysis filter will be

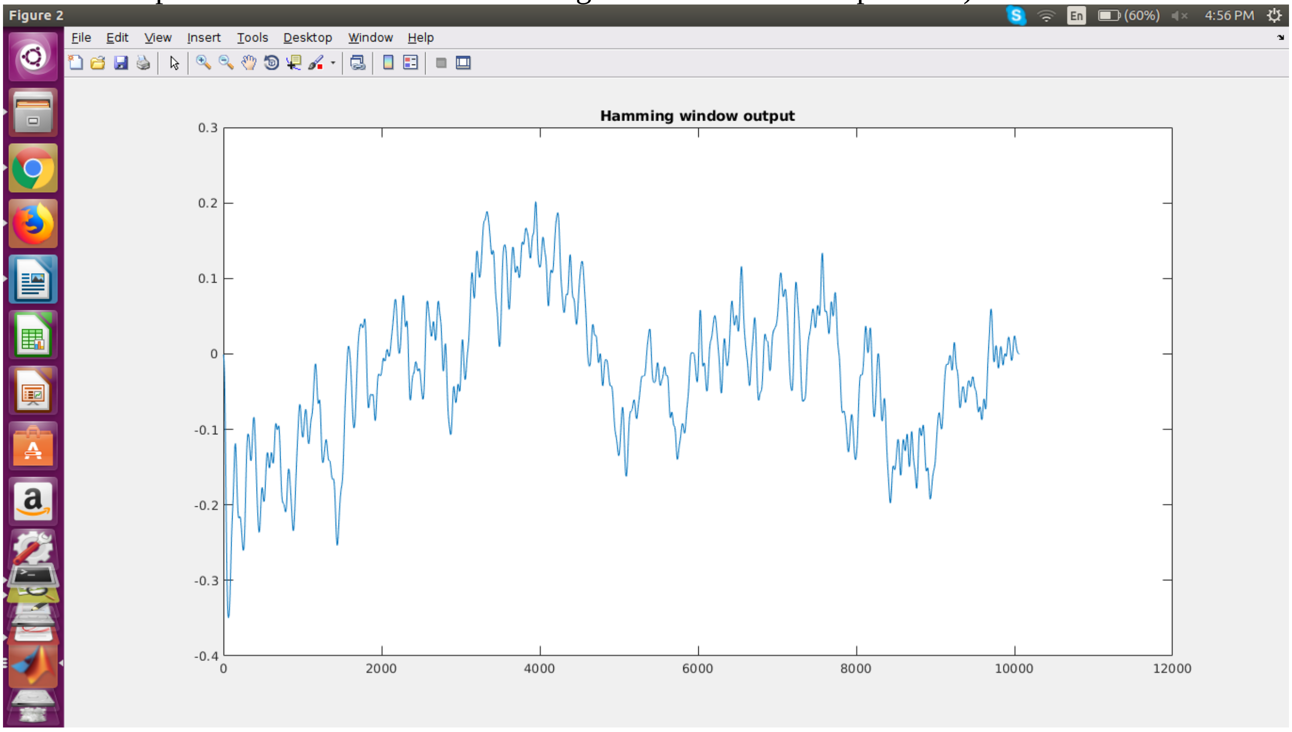
$E(z)/S(z) = 1-\Sigma a_k z^{-k}$ as it gives us the LP residual when we multiply it with the z-transform of the signal, S(z).

We can get the LP residual by taking the speech signal s(n) and finding its z-transform S(z) and multiplying it with the analysis filter E(z)/S(z). We can also do this in the time domain by convolving s(n) with [1 $-a_1$ $_-a_2$ $_-a_3$ ... $-a_p$] where a_k refer to the LP coefficients. The above two operations are equivalent as convolution in the time domain is equivalent to multiplication in the frequency domain.

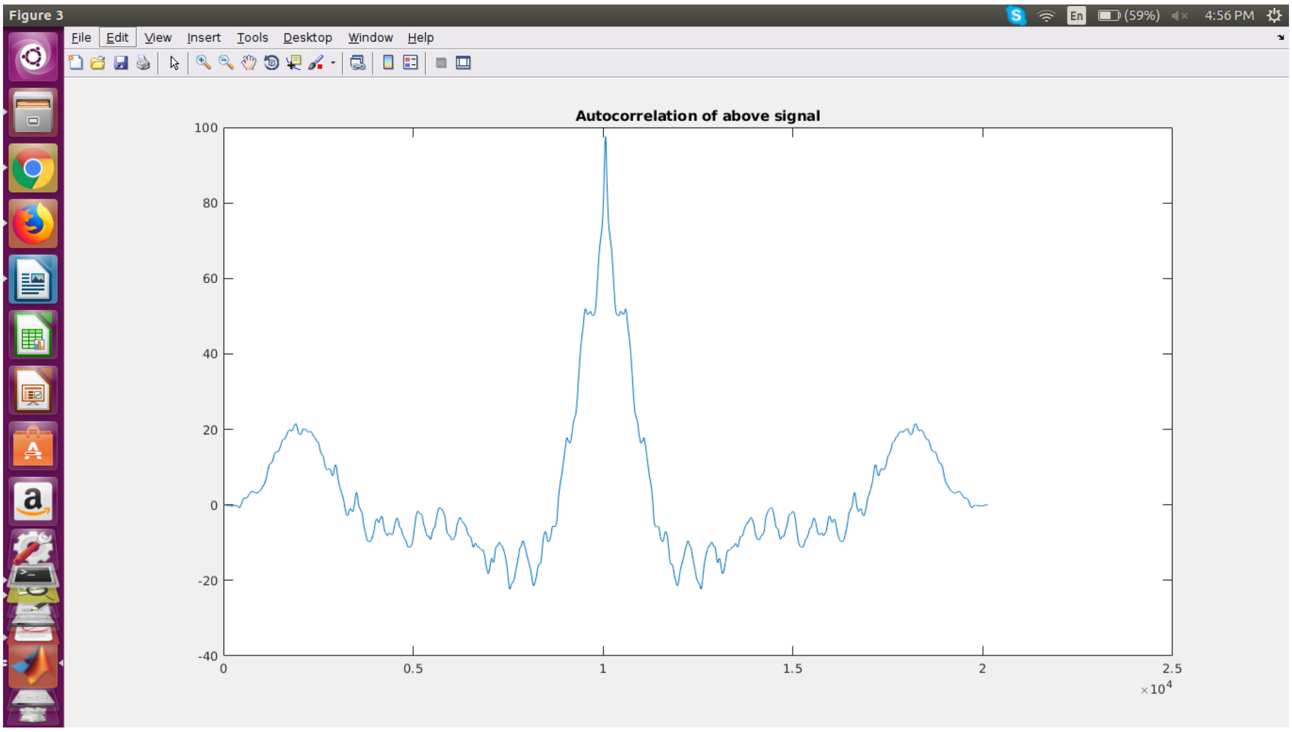
c) The wavefile *my_name.wav* contains a recording of my name. The results shown below are the outcome of the *q2c.m* file which contains the code for this section. Taking the first 10,000 samples of this speech signal, we get the following plot between the speech signal and time.



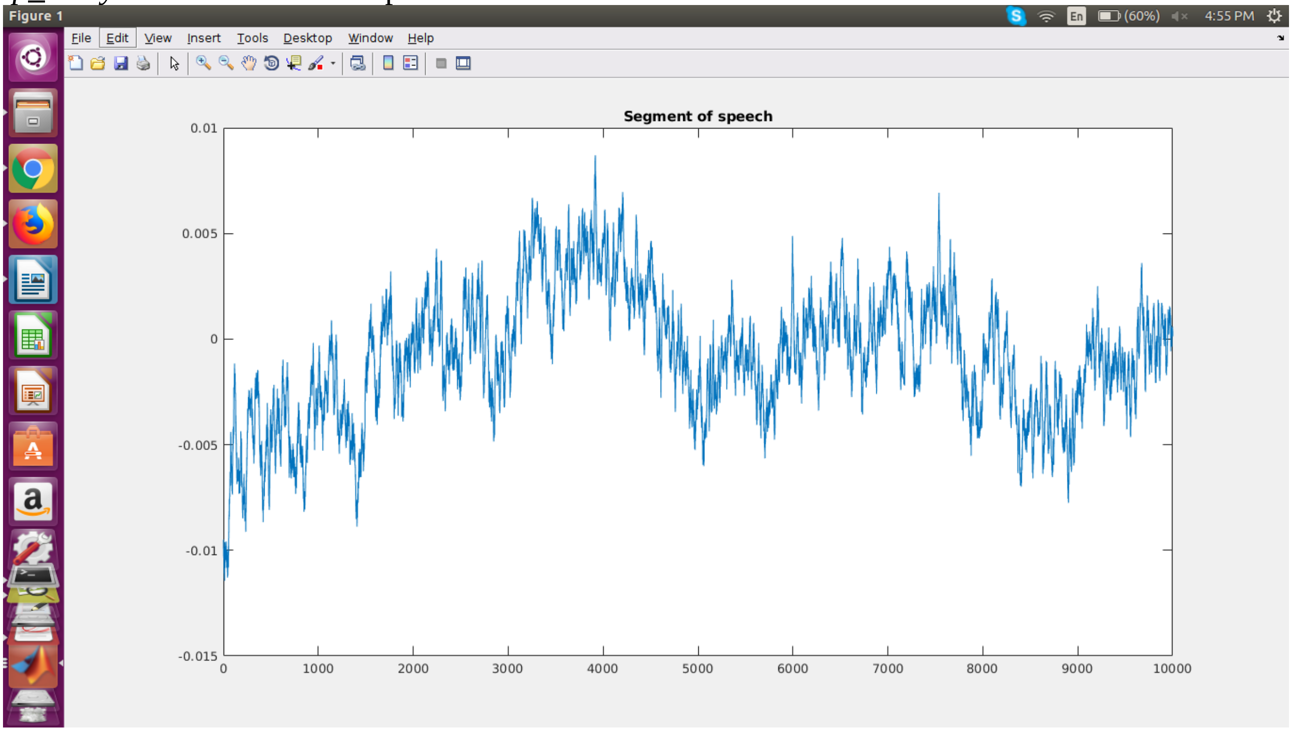
We can apply a Hamming window to the above signal in order to get the following waveform. We can observe that the output becomes less noisy in nature. (The figure for Hamming window output and LP residual are shown together in 2d with comparisons)



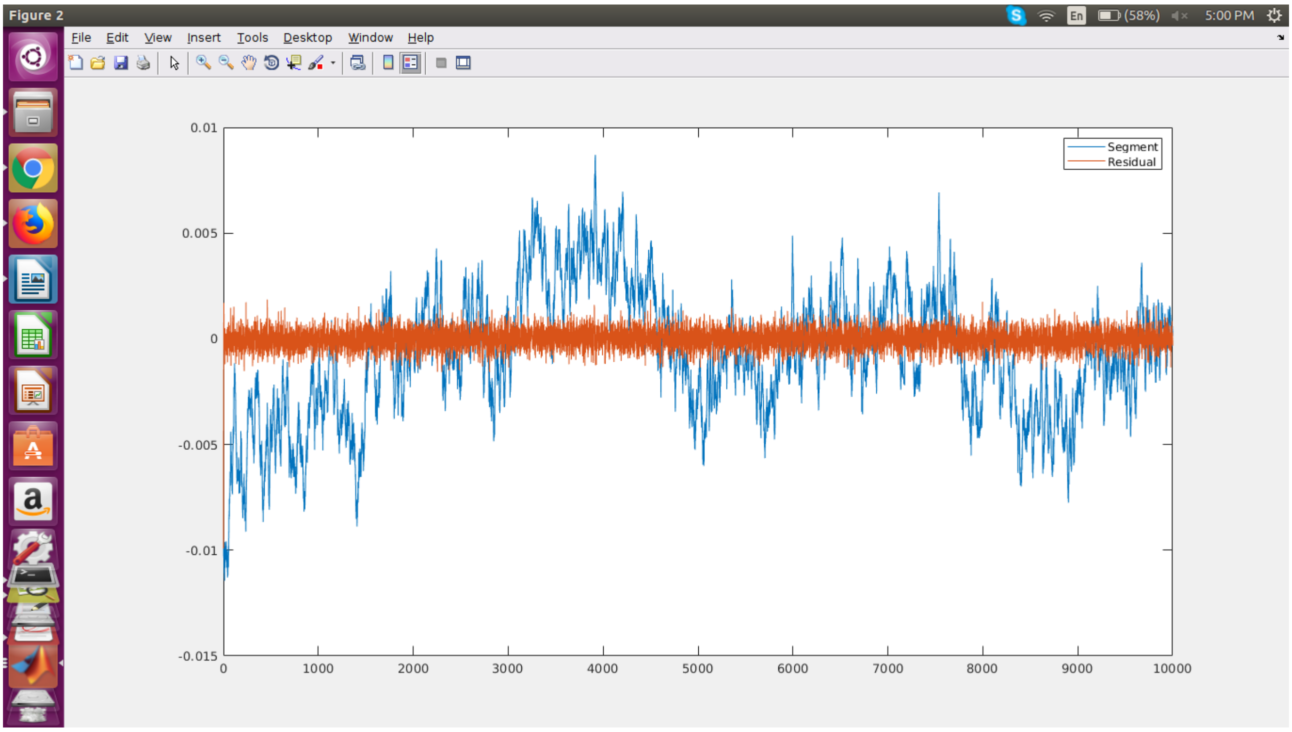
Autocorrelation is then performed on the above signal in order to get the following waveform. We can observe that this output is symmetric in nature. We can observe that there is one main peak in the plot shown below but there are several smaller peaks surrounding it. These smaller peaks can be considered to be spurious peaks as they are very close to the actual one. (The two figures are shown together in 2d with comparisons).



d) The segment of the speech signal which was used in the above subsection is plotted below against time. Note that the code used for this subsection is *q2d.m* which calls the function lp*_analysis.m* in order to compute the LP residual.



On plotting the LP residual, we get the following plot.



We can observe that the LP residual is mostly concentrated near zero, showing that the LP analysis is doing a decent job in predicting the future samples.From this plot, we can see that the error tends to be larger where the speech signal has peaks. This is as it is difficult for the system to measure and predict sudden changes in the signal. In the above plot, the red plot is for the LP residual whereas the blue plot is for the speech segment shown in the previous image.

The autocorrelation of the LP residual of the given speech segment is shown below.

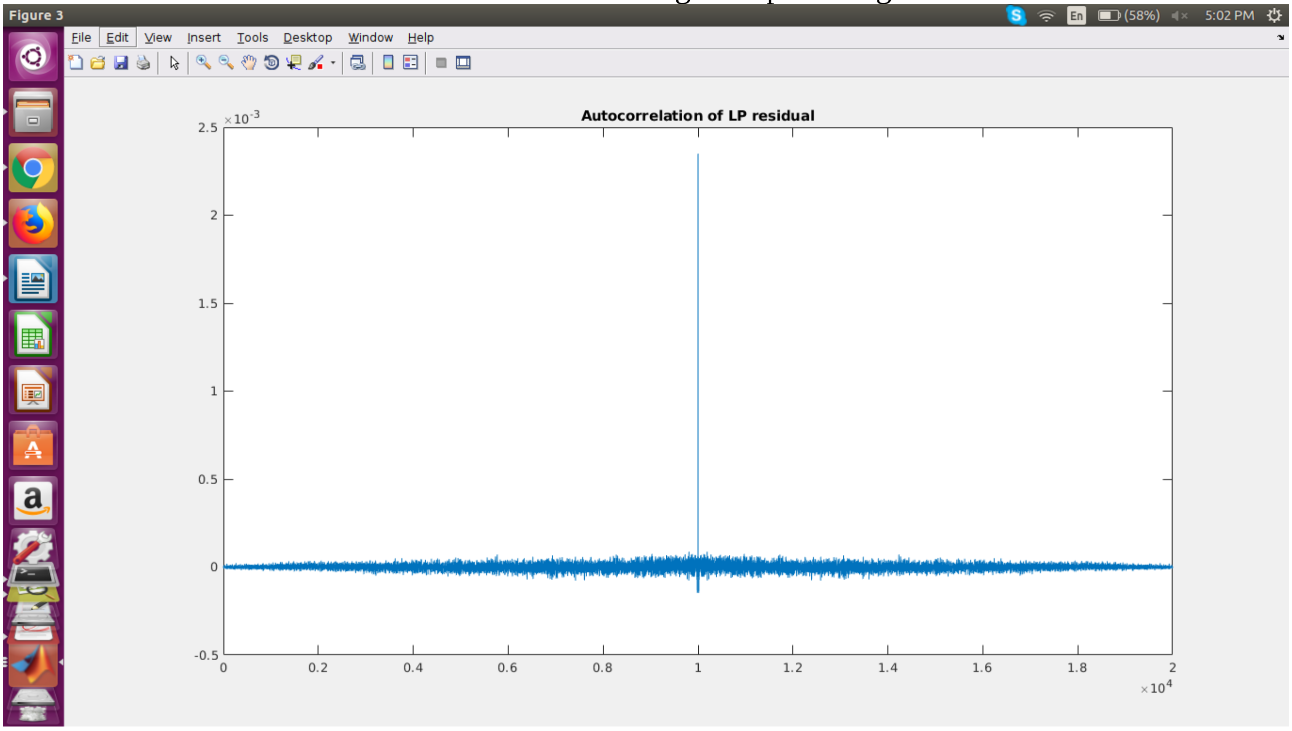
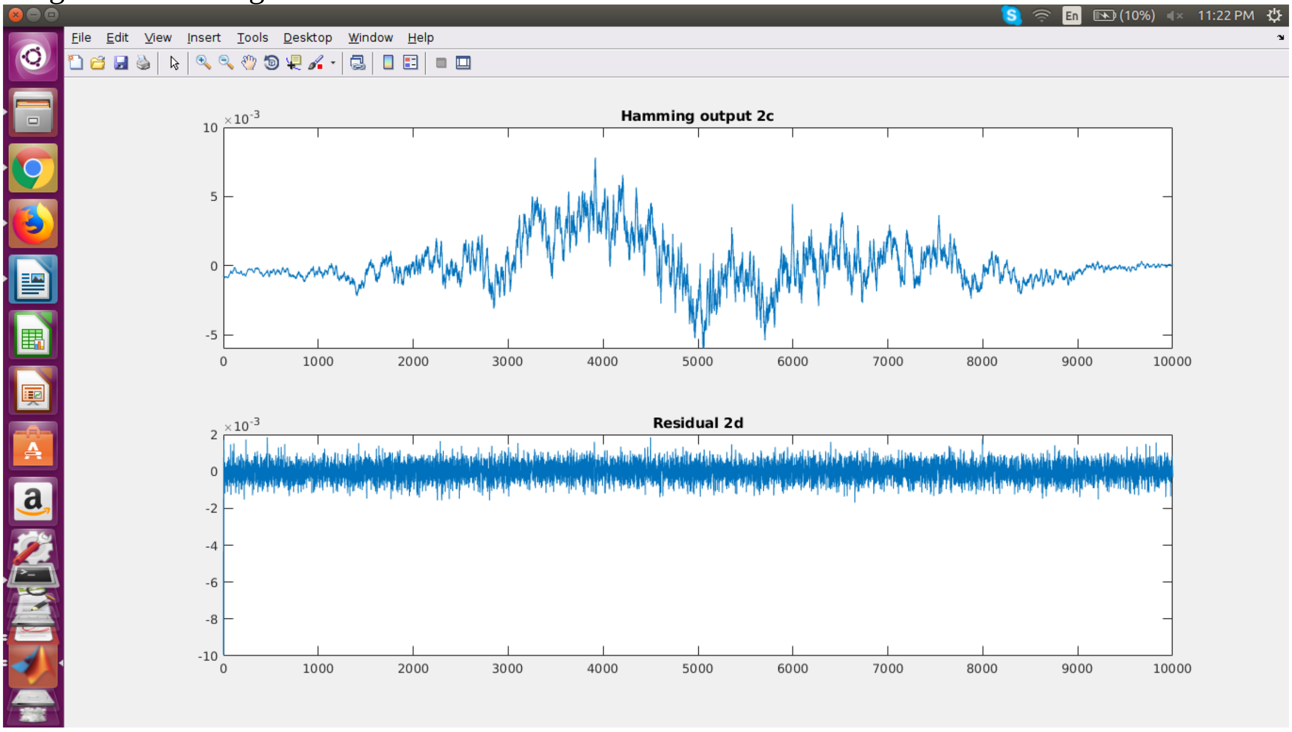


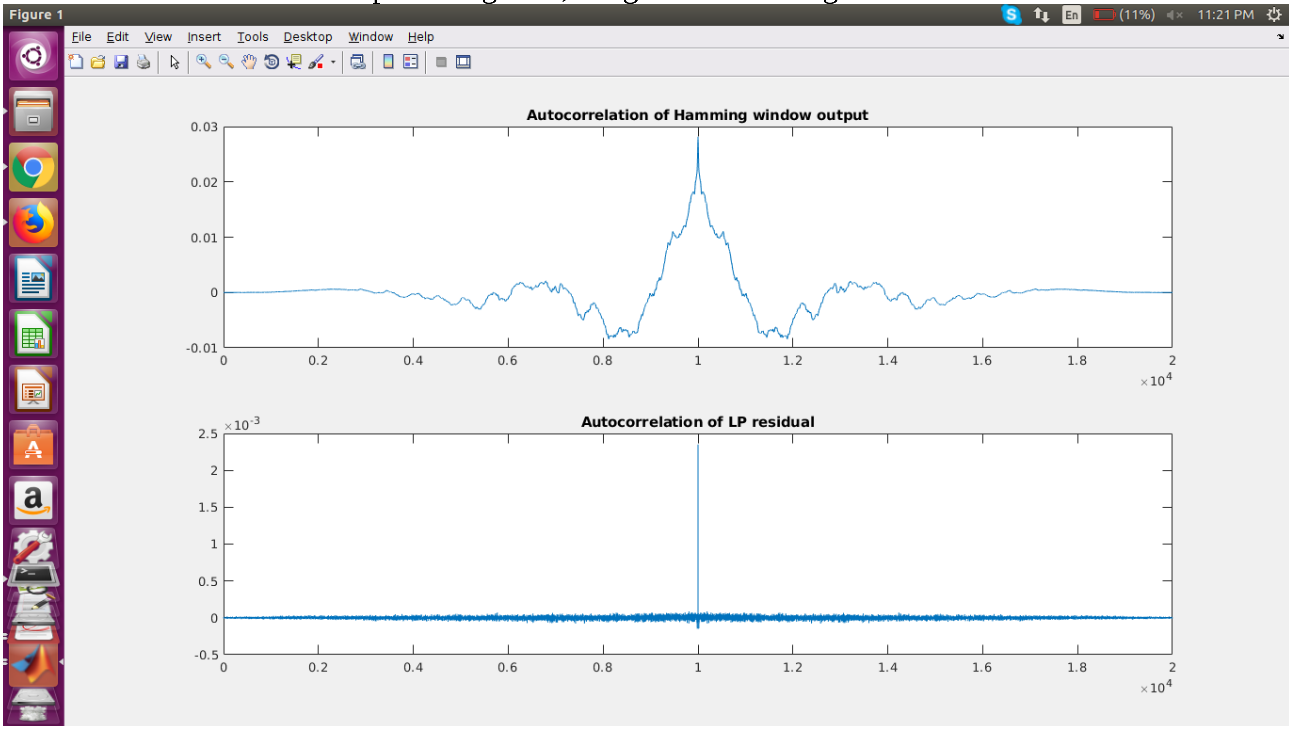
Figure 3
File Edit View Insert Tools Desktop Window Help
5:02 PM
(58%)
En
Autocorrelation of LP residual
×10-3
2.5
2
1.5
1
0.5
0
-0.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
×10^4

The autocorrelation is symmetric as it was in the previous subsection. The difference between the two autocorrelation plots is that for the LP residual, the autocorrelation seems to have a single large peak surrounded by smaller negligible values. The autocorrelation of the Hamming window does show a large peak but it is surrounded by smaller peaks which can be considered to be spurious peaks.

Comparing the Hamming window output and the LP residual of the same speech segment, we get the below figure.



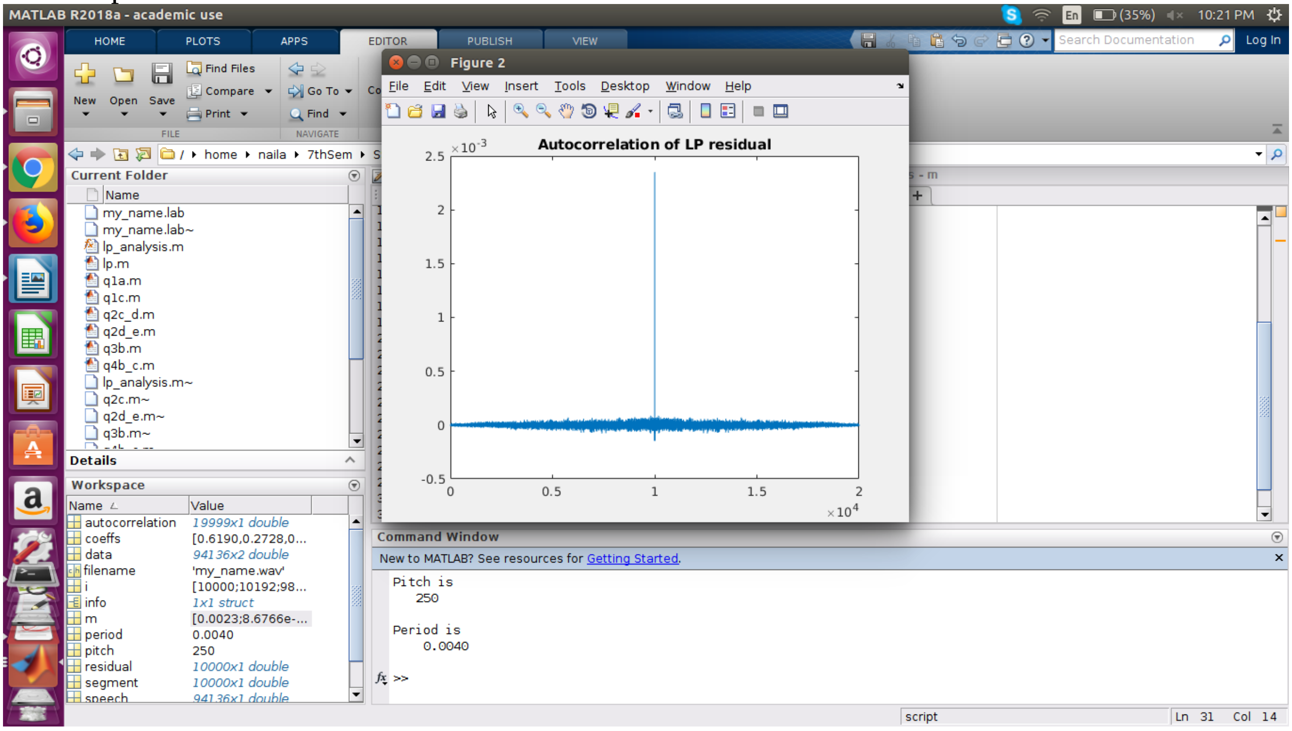
Comparing the autocorrelation of the Hamming window output and the autocorrelation of the LP residual of the same speech segment, we get the below figure.



We can observe that both have a major peak near 1. The autocorrelation of the Hamming window output shows several smaller peaks surrounding the major peak but the autocorrelation of the LP residual has very small values (around 0) in regions other than the main peak. The autocorrelation of LP residual almost resembles an impulse signal as it has no spurious peaks.

e) The pitch of a speech signal can be calculated by using its LP residual. This is as we can take the autocorrelation of the LP residual and find the difference between the two largest peaks. This is slightly better than just taking the autocorrelation of the speech signal as the spurious peaks tend to get eliminated by taking the autocorrelation of the LP residual. In section 2d, we saw how the autocorrelation of residual has a major peak but no spurious peaks whereas that of the hamming windowed projection of the signal had several spurious peaks which may lead to miscalculations.

In the code q2d_e.m, we compute the autocorrelation of the LP residual. We then find the indexes of the two largest peaks in the autocorrelation and the difference between these indices gives us the number of samples between the two largest peaks. We then find the pitch frequency by dividing the sampling rate by the difference in samples between the two largest peaks. For the audio file 'my_name.wav' this came to 250 Hz. We then find the pitch period by taking the reciprocal of pitch frequency. The pitch period is 0.004 seconds (4 ms). This makes sense as the pitch of a female speaker is between 125-280 Hz.



Workspace

Name	Value
autocorrelation	19999x1 double
coeffs	[0.6190,0.2728,0...
data	94136x2 double
filename	'my_name.wav'
i	[10000;10192;98...
info	1x1 struct
m	[0.0023;8.6766e-...
period	0.0040
pitch	250
residual	10000x1 double
segment	10000x1 double
speech	94136x1 double

New to MATLAB? See resources for Getting Started.

```
Pitch is
   250

Period is
    0.0040

>>
```

Question 4

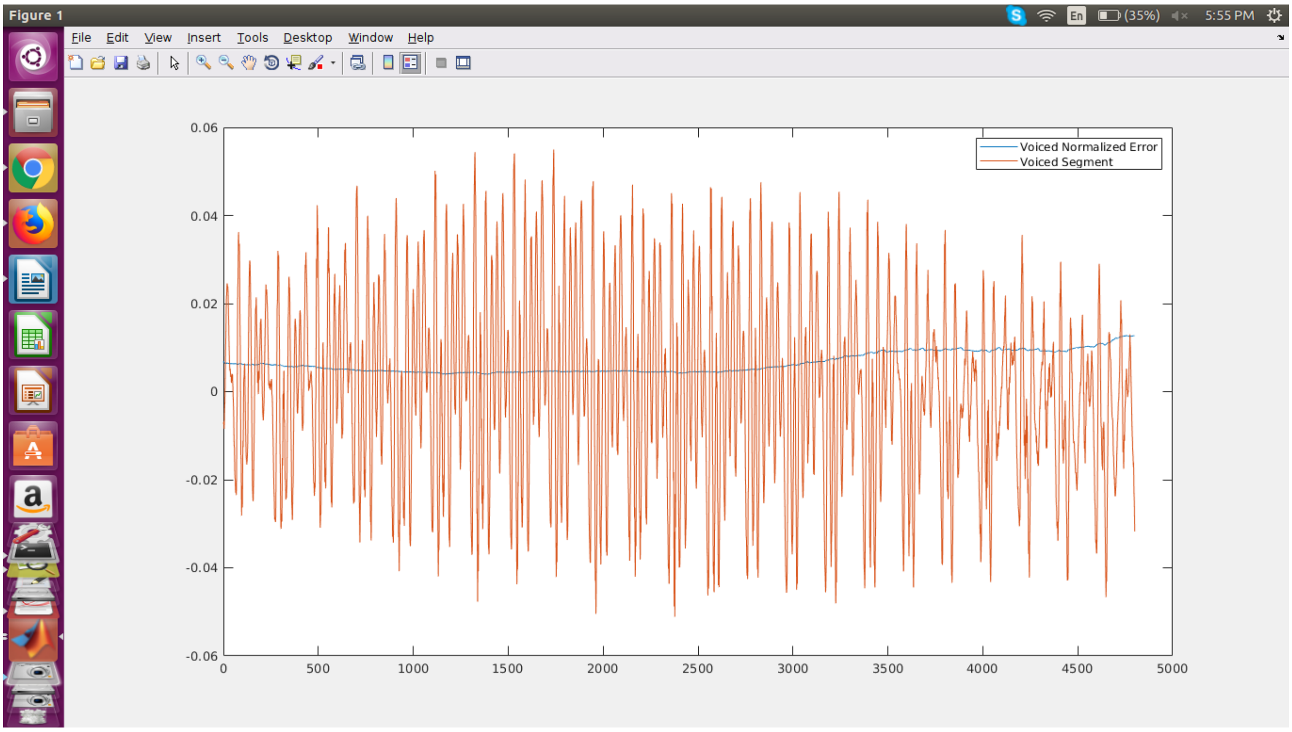
a) Normalized error refers to the ratio of the energy of LP residual to the energy of the speech signal. The energy of a signal is calculated by taking the sum of squares of the signal values. The below formula tells us how to calculate the normalized error.

Normalized error (η) = $E_{residual}/E_{signal} = \Sigma e^2(n) / \Sigma s^2(n)$

It should be noted that normalized error helps us distinguish between voiced and unvoiced segments of speech. This is as the normalized error is smaller for voiced regions and larger for unvoiced regions. The graph between normalized error and order of LP analysis (*p*) is such that the difference between voiced and unvoiced regions increases for values of *p* above and equal to 8. This is intuitive as for voiced speech, the error energy will be less (as it is more predictable and has a pattern) and the signal energy will be more whereas for unvoiced speech, the error energy will be more (as it more random) and the signal energy will be less.

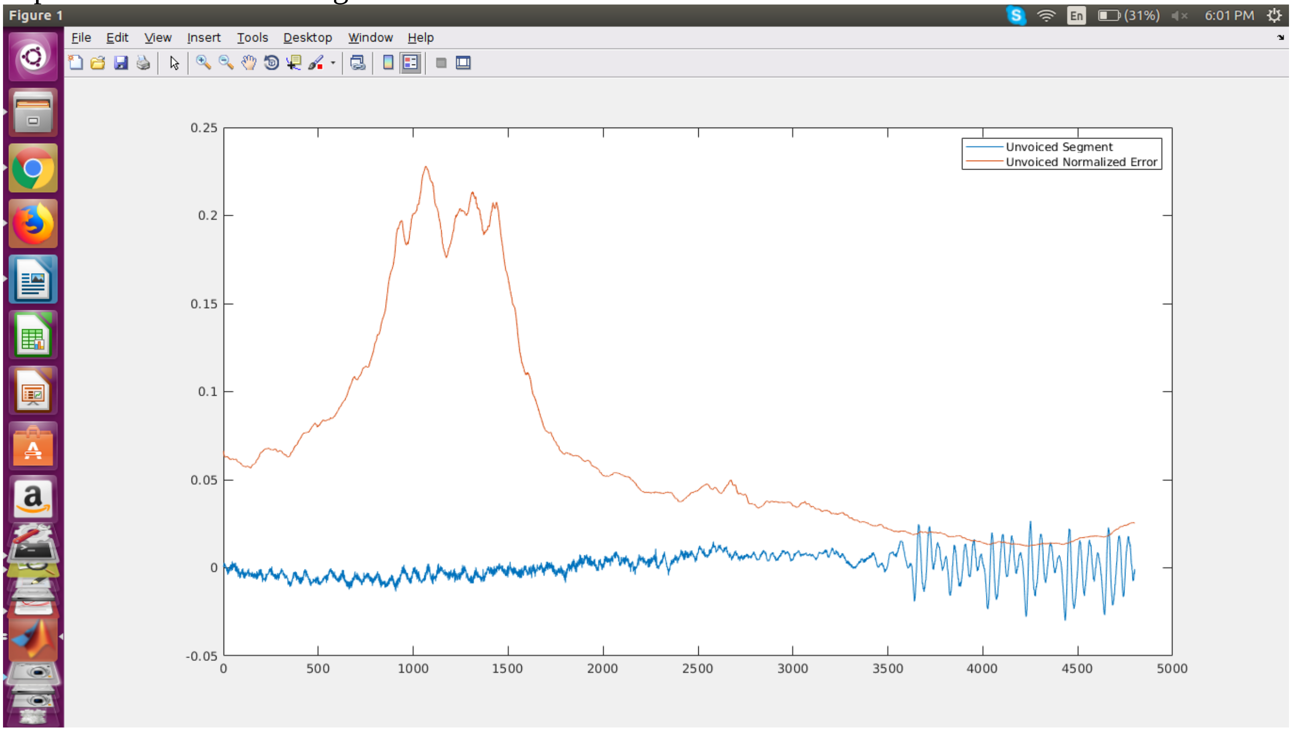
Normalized error signifies how much our predicted signal deviates from the actual signal. This is as if there is more error present, the energy of the residual will be higher. The normalized error tells us what fraction of the signal energy is present in the residual energy. Naturally, we want the fraction to be less which is why we aim for better LP coefficients.

b) The plot for voiced speech and its normalized error is shown below. Note that for the voiced segment, I extracted the part of my name which contains the voiced diphthong 'ai' and did all experiments on it. This segment is stored in the '*voiced.wav*' file.



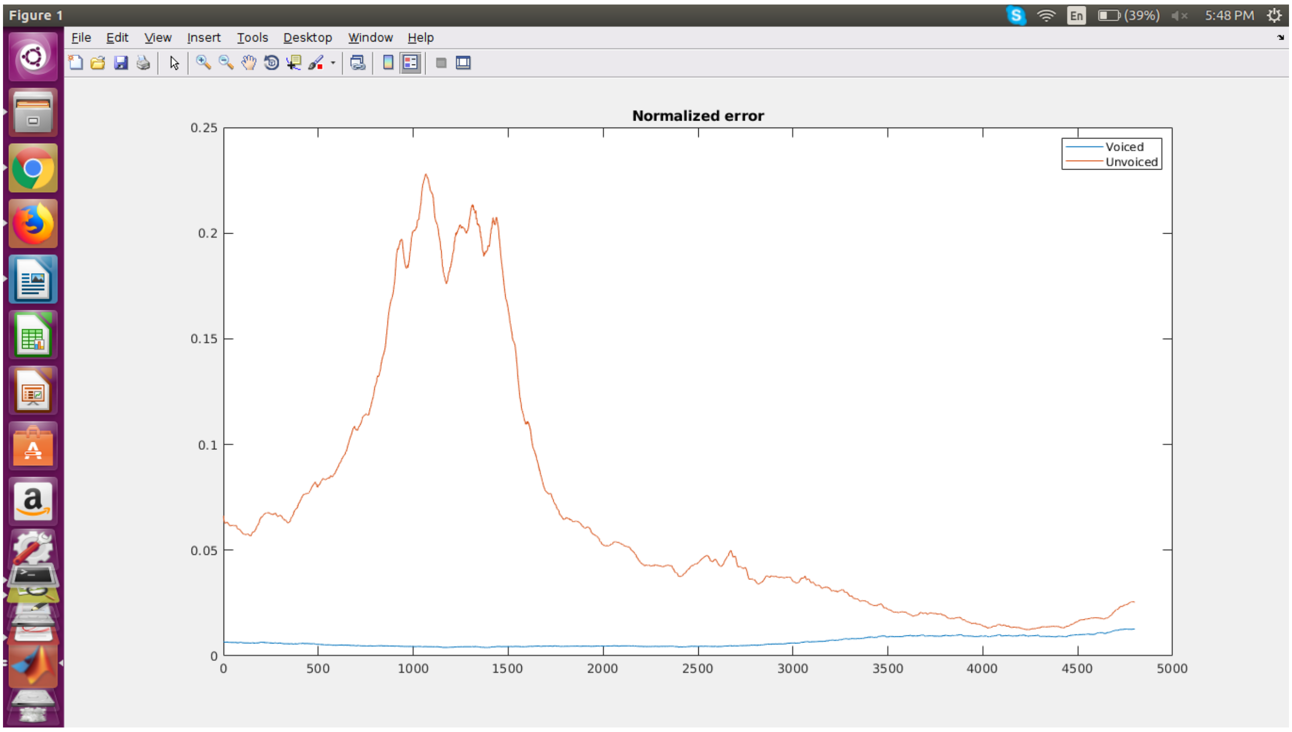
We can see above that the normalized error for voiced speech is around 0 (less than 0.02). This is expected as we did LP analysis using order 8 which gives smaller error for voiced speech and larger error for unvoiced speech.

c) The plot for unvoiced speech and its normalized error is shown below. Note that for the unvoiced segment, I extracted the part of my name which contains the unvoiced fricative ‘f’ and did all experiments on it. This segment is stored in the ‘*unvoiced.wav*’ file.



We can easily observe that the normalized error for unvoiced segment is more than that for the voiced segment. This is because unvoiced segments tend to lack quasiperiodic vibrations and are therefore random or more noise-like in nature when compared to voiced segments.

The below plot compares the normalized error for both voiced and unvoiced segments. Both have been done by using an LP order of 8.



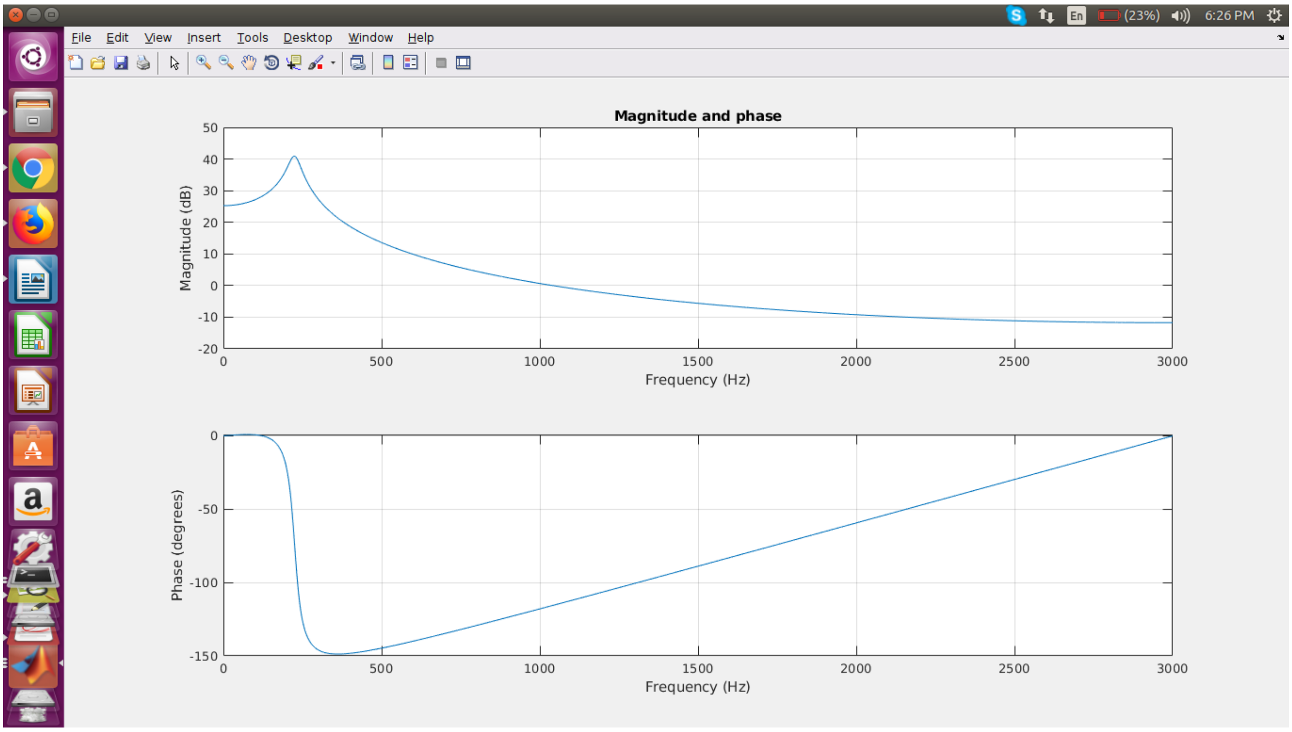
The blue line refers to the normalized error for the voiced segment whereas the red line refers to the same for the unvoiced segment. We can easily observe that the voiced segment has a normalized error which is much less than that of the unvoiced segment. This was predicted as voiced segments are more predictable in nature whereas unvoiced segments are more random or noise-like in nature.

Question 1

a) A digital resonator is a two-pole bandpass filter with a pair of complex conjugate poles located close to the unit circle. This filter is known as a resonator as it has a large magnitude response near the pole positions. The angle of the pole location determines the resonant frequency of the filter. Since it is a two-pass filter, the transfer function of the digital resonator will have a quadratic function with two roots in the denominator. The transfer function of a digital resonator will be of the form,

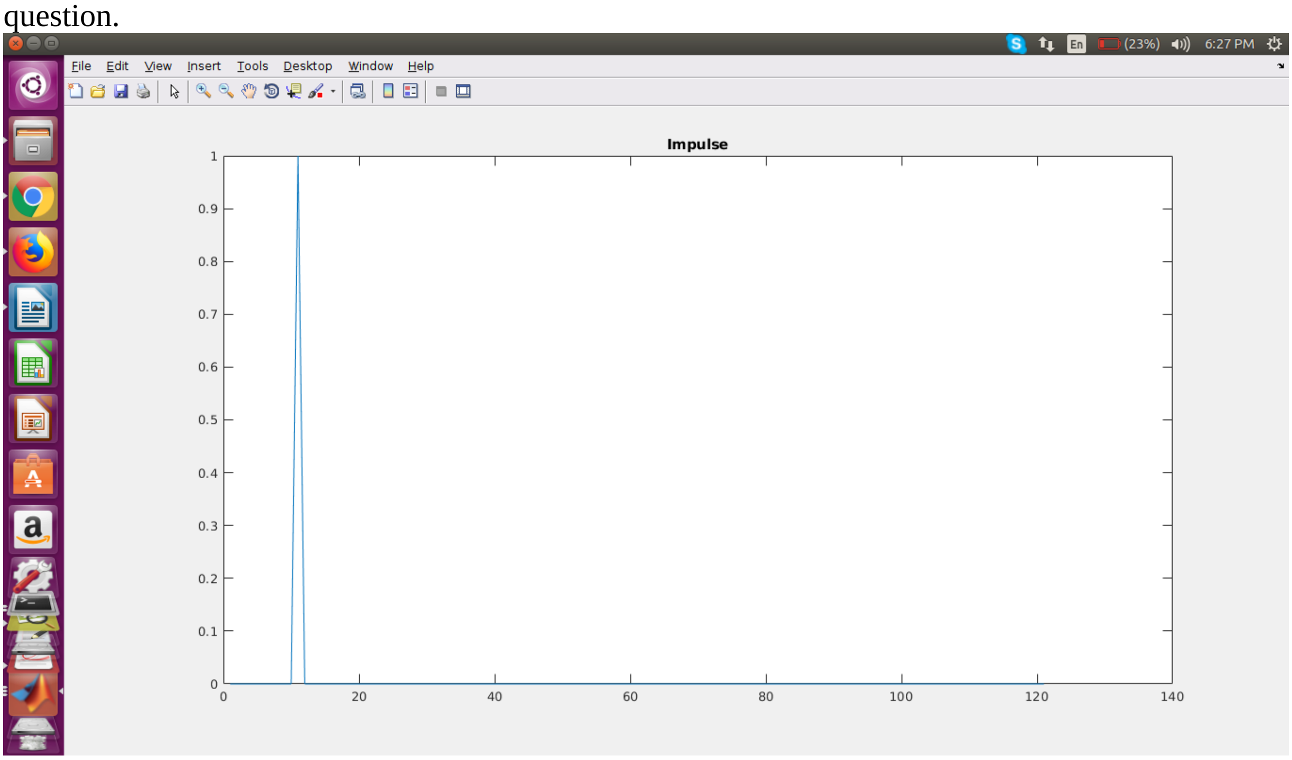
$$H(z) = b/(1+a_1z^{-1}+a_2z^{-2}) \quad \text{where } z = e^{jw}$$

The magnitude and phase response of a digital resonator with bandwidth 50 Hz and frequency 300 Hz is shown below. We can observe that the impulse response is characterized by peaks near the resonant frequencies (300 Hz in this case). We can construct this resonator by plotting the frequency response of the transfer function by using the *freqz* function. The numerator coefficient b is 1 while the denominator coefficients a_1 and a_2 are -2*e(-ΠBT)cos(2ΠfT) and e(-2ΠBT) where f = resonating frequency = 300Hz and B = bandwidth=50 Hz.

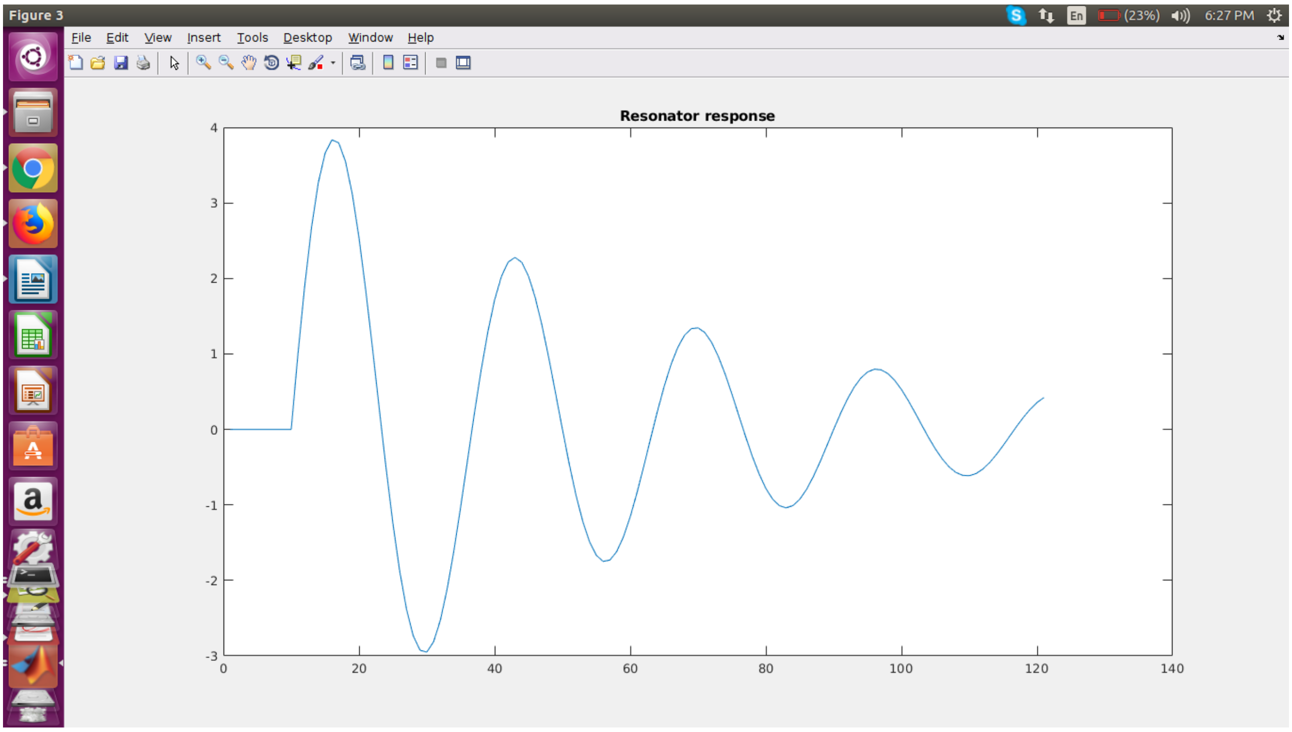


As we can notice in the above graph, the peak in the magnitude spectrum occurs at the resonating frequency 300 Hz.

b) The impulse which has been used as excitation is shown below. This impulse has been created by taking an array of zeros which has a 1 in its 11th position as this was what was shown in the question.



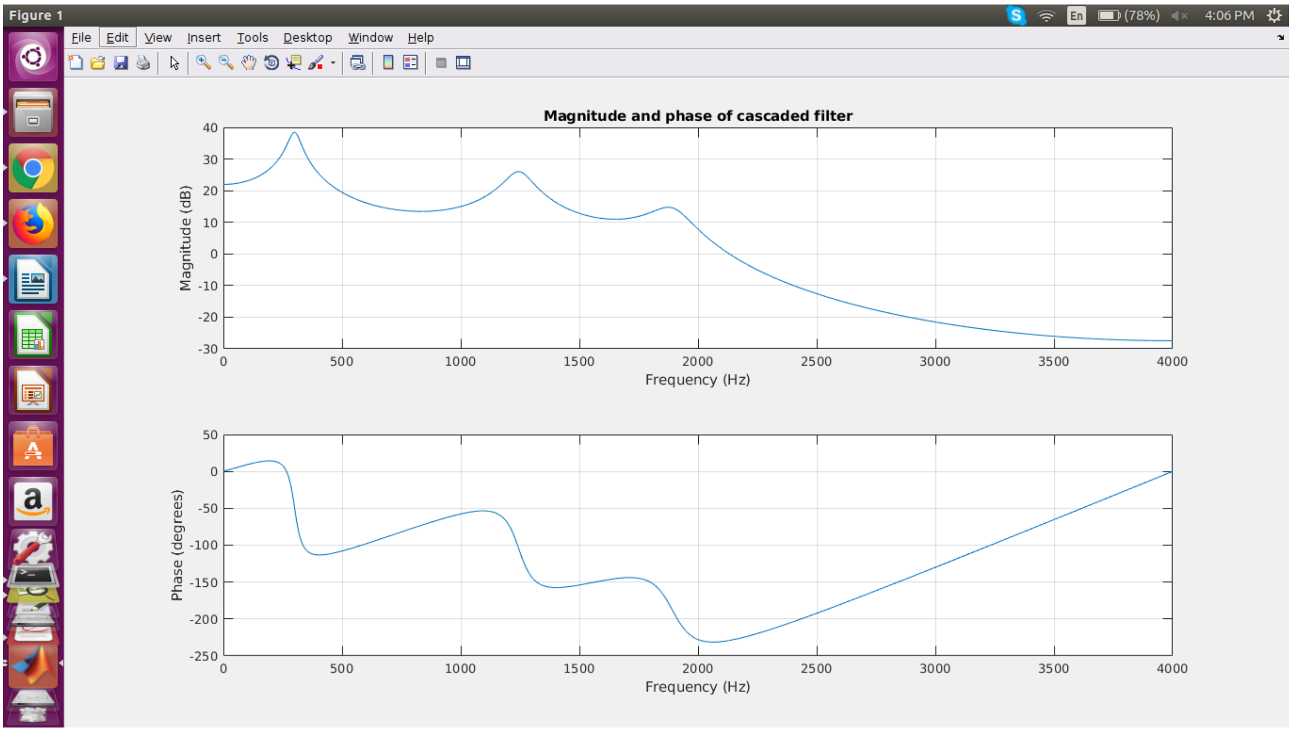
We know that on passing an impulse through a digital resonator, the output will be a damped sinusoid. The resonator response for the above impulse is shown below. As predicted, it shows a damped sinusoid (as it is sinusoidal in shape but the magnitude keeps decreasing for increasing frequencies).



c) In order to cascade three resonators which have the corresponding bandwidth and frequency characteristics:

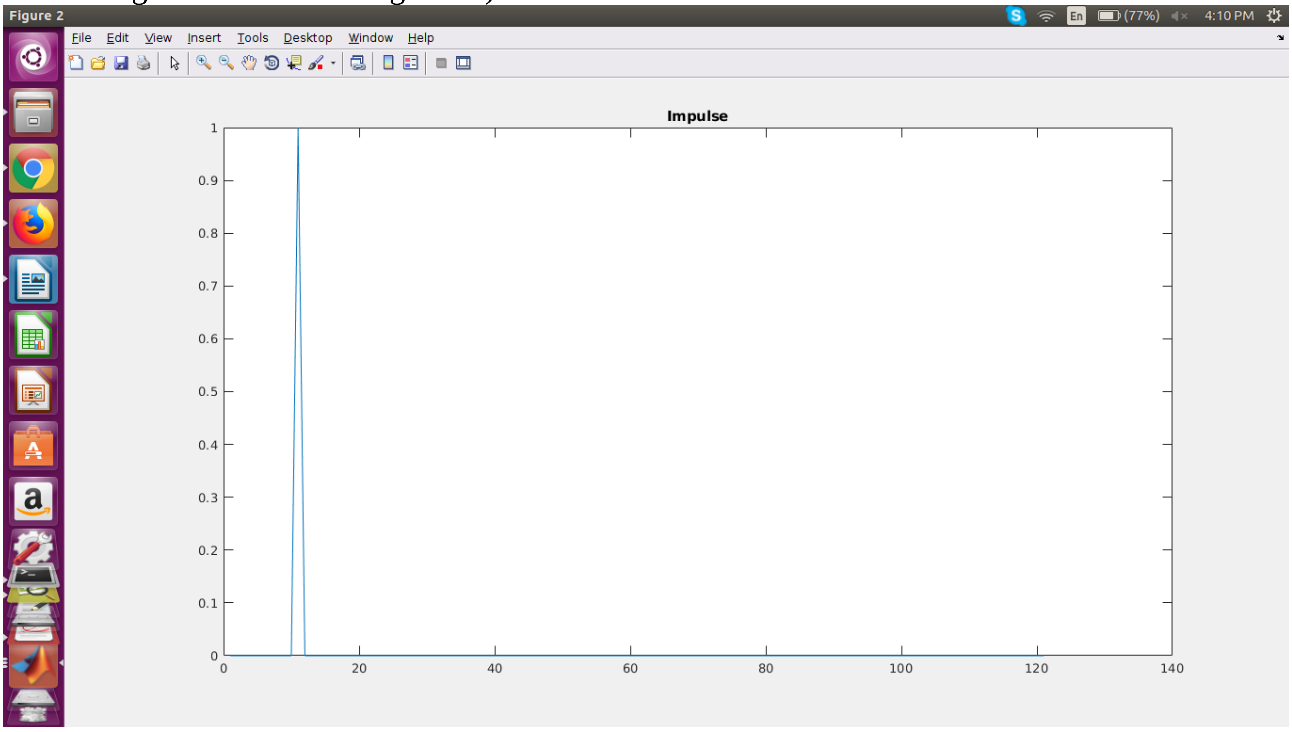
f1 = 300 Hz, b1 = 50 Hz
f2 = 1245 Hz, b2 = 110 Hz
f3 = 1892 Hz, b3 = 160 Hz

The response of the cascaded resonator is found by plotting the transfer function by using the *freqz* function. The numerator will be 1 whereas the denominator will be conv(conv(B1,B2),B3) where B1 = [1 b_1 c_1] , B2 = [1 b_2 c_2] and B3 = [1 b_3 c_3]. Note that the b_i terms will be of type -2*e(-B_i ΠT)cos(2Πf_iT) and the c_i terms will be of type e(-2ΠB_iT) where B_i and f_i are the bandwidth and frequency for the corresponding filter. The impulse magnitude and response for the cascaded filter is shown as below.

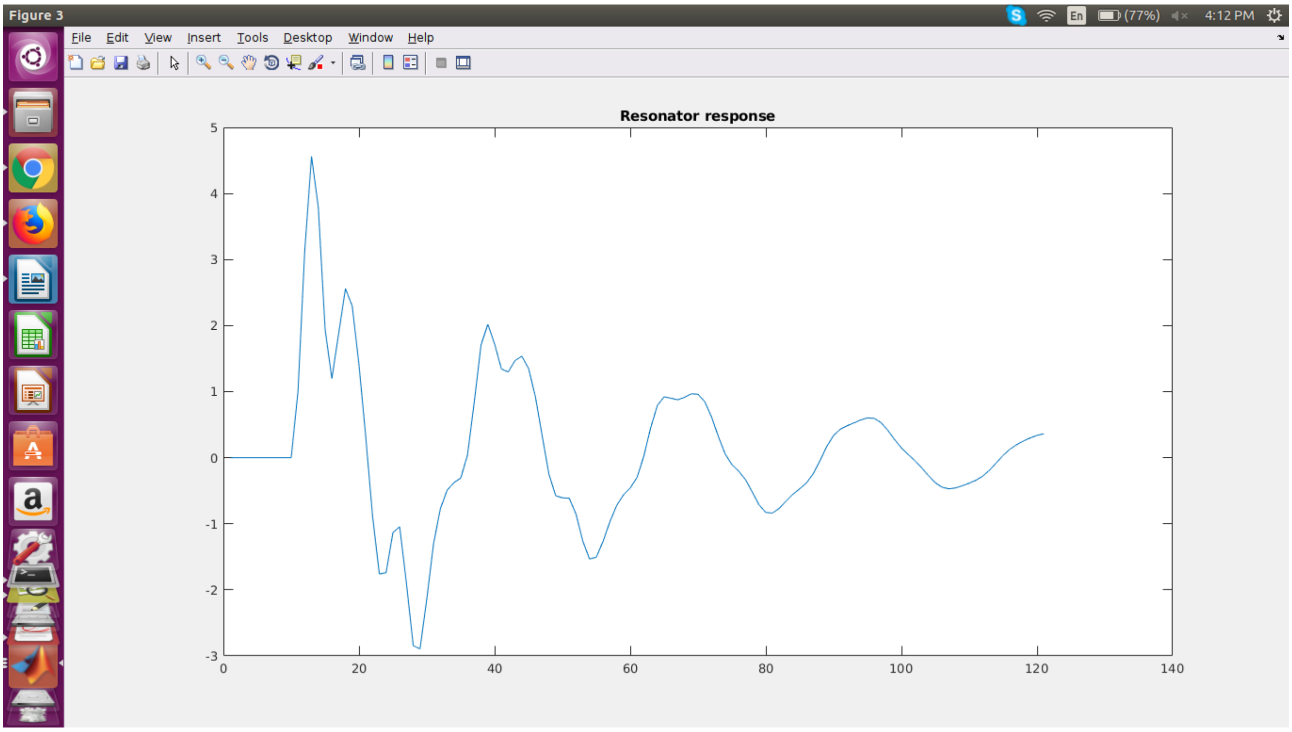


As we can see from above, there are three peaks corresponding to the three different frequencies. The first peak is at 300 Hz, the second at 1245 Hz and the third at 1892 Hz.

The impulse response that we have used is same as before and is shown below. We create this by taking an array of size 120 filled with zeros having a 1 in its 11th position (as this was similar to the image shown in the assignment).



The response of the cascaded resonator to the impulse is as shown below.



We can easily see that the response does have a shape similar to that of a damped sinusoid but it has some noise-like regions. The impulse response of the resonator in 1a was a smooth damped sinusoid but the impulse response for this shows random peaks in the sinusoidal shape.

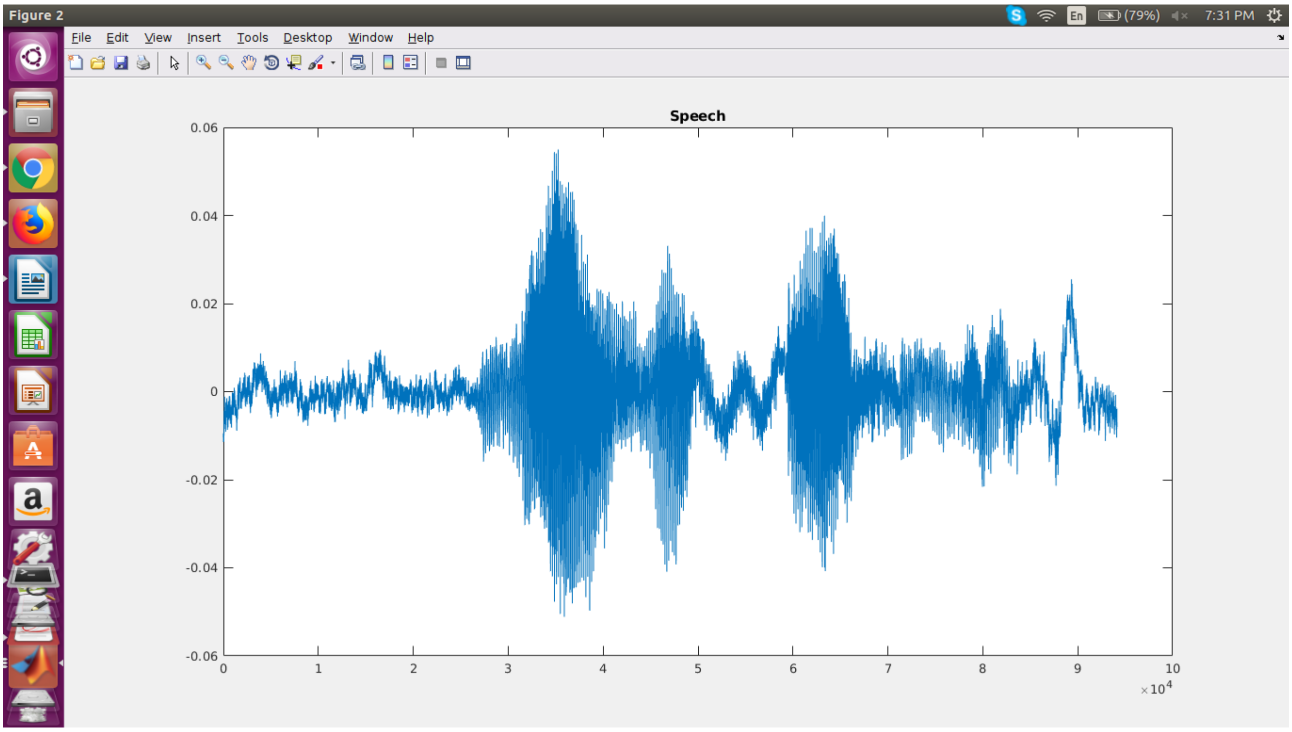
Question 3

a) The vocal tract system response can be estimated by using LP analysis. This is as LP analysis allows us to extract formants which allows us to estimate the vocal tract response as well as epochs and pitch which in turn allow us to estimate the excitation source. Formants are the resonating frequencies of the vocal tract system which are characterized by a concentration of spectral energy. We know that the vocal tract system changes with the changes in position of the articulators (tongue, lips, etc) so this in turn causes the formants to change. In order to do formant extracation using LP analysis, we have to find the vocal tract response which tells us the frequencies at which the vocal tract resonates.

As was shown in q2a, the vocal tract can be modelled as a synthesis filter. This means that we can get the vocal tract response by finding S(z)/E(z) where S(z) and E(z) are the z-transforms of the signal and LP residual, respectively. Note that we can use a synthesis filter with the residual to get the speech signal. We know that $E(z) = S(z)\ (1-\Sigma a_k z^{-k})$ as this is done by doing the z-transform on the equation, $e(n) = s(n) – \Sigma a_k s(n-k)$. We can get the synthesis filter from the above equation as it is equivalent to $S(z)/E(z) = 1/(1-\Sigma a_k z^{-k})$.

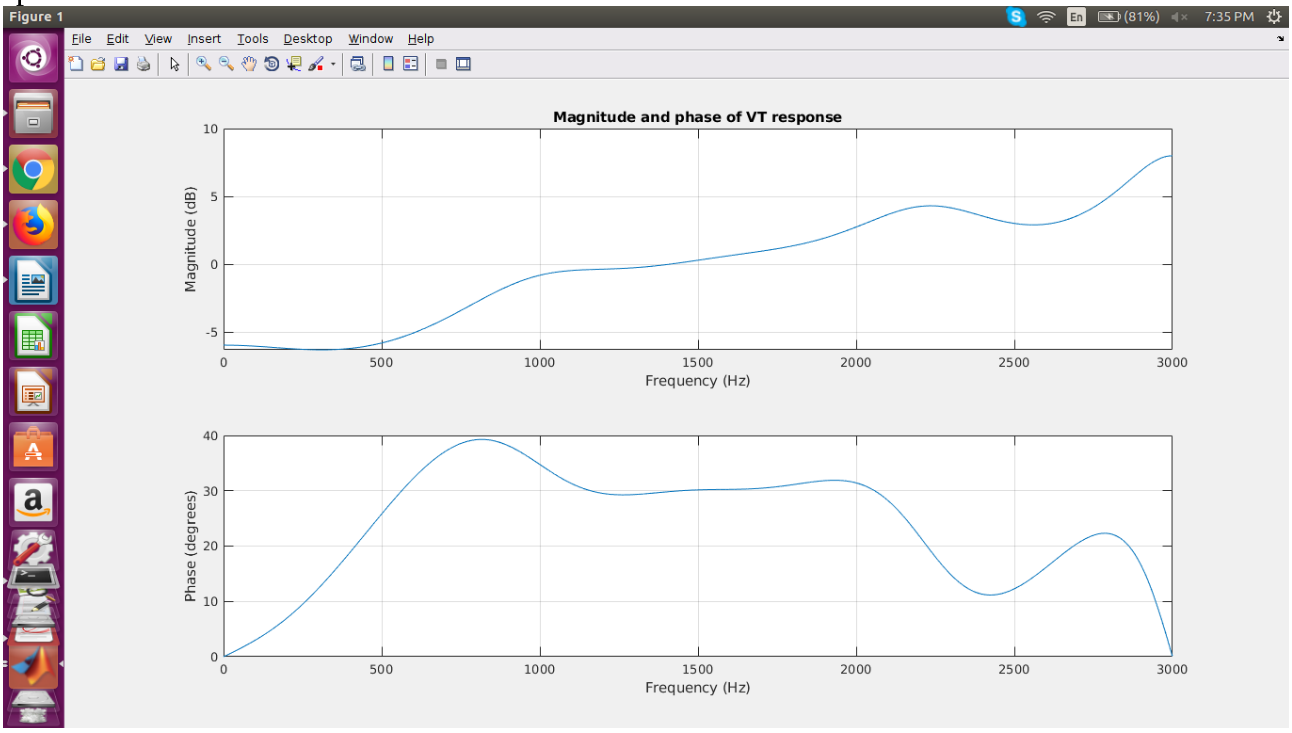
Therefore, in order to do formant extraction, we must first find out the values of the LP coefficients a_k. We should then use these values to create the synthesis filter. On plotting the synthesis filter, we get the LP spectrum. Formants are denoted by the peaks of the LP spectrum (as they have highest spectral energy) and using this we can easily extract them.

b) I have plotted the first 10000 samples of the signal ‘my_name.wav’. Below is the plot which I have gotten.



We can plot the vocal tract system response for the corresponding signal by finding the LP coefficients of the signal. When we get the LP coefficients a_k we can then construct the transfer function for the VT system by creating the following filter:

$H(z) = 1/(1-\Sigma a_k z^{-k})$ where a_k are the LP coefficients. By doing this, we get the LP spectrum. The LP spectrum was created by using the *freqz* function of matlab which had the numerator 1 and denominator $[1\ -a_1\ -a_2 \ldots -a_p]$. The following graph shows the plot for the LP spectrum.



We can see that there are three peak frequencies which correspond to the formants F1, F2 and F3 and denote a concentration of spectral energy at these frequencies.

c) Formants can be obtained from LP analysis as they are the peaks of the LP spectrum. In our above plot, we can observe that the peaks occur at around 1000 Hz, 2200 Hz and 3000 Hz.

Therefore, the formants in our speech segment will be at 1000 Hz (F1), 2200 Hz (F2) and 3000 Hz (F3). We have found the formants by finding the LP coefficients of the speech signal and then constructing the synthesis filter to go to the LP spectrum. The peaks of the LP spectrum are the formants of the signal.