Assignment 2 of Speech Signal Processing
Naila Fatima
201530154

Question 1:

a) Voice active detection (VAD) or voice/unvoice detection refers to the process which allows one to distinguish between voiced and unvoiced regions of speech by using certain characteristics. Certain parameters which can be used for this detection include energy, zero crossing rate, normalized autocorrelation coefficients and pre-emphasized energy ratio.

Energy: Energy can be used to distinguish between voiced and unvoiced regions in speech as voiced speech tends to have more energy in the time domain as well as the spectral domain (in the form of darker spectrograms). This parameter measures the summation of the square of the signal values over an interval of time. The formula for this parameter is *e(i) = sum(signal(j)^2)* where *j* takes all values in the interval *[i, i+N-1*]. Another way of computing energy is to take the logarithm of the sum of squared values (take log(e)) as this gives a smaller dynamic range. We can observe that this parameter tends to capture the amount of energy present in an interval of the speech signal. As voiced speech tends to have higher energy and unvoiced speech has lower energy, this parameter can be used to do VAD as voiced speech will show higher values in *e*, whereas unvoiced speech will show lower values in *e*. A simple threshold can be used to distinguish between voiced and unvoiced speech when using this parameter.

This parameter captures the energy of the signal in the interval taken. The shortcoming of this parameter is that the presence of noise in the signal affects the accuracy with which the voice/unvoice detection is done. Also, putting a threshold is slightly difficult as thresholds vary across signals and especially when the signals have noise.

Zero Crossing Rate (ZCR): The ZCR refers to how many times the signal crosses the zero level over an interval of time. This parameter captures the “noisy-ness” or the randomness present in a signal by measuring how many times it crosses the zero level. We know that voiced speech tends to have some sort of predictable pattern whereas unvoiced speech tends to resemble a noisy, random waveform. As unvoiced speech is more noisy, it will tend to have more fluctuations and a higher ZCR whereas voiced speech, which has a more predictable pattern, will have a lower ZCR. Note that silence will have a zero ZCR provided that the signal is not corrupted by noise.

The ZCR captures the noisyness or the randomness of the signal in a given interval. The shortcoming of this parameter is that it is easily affected by noise thereby rendering it nearly useless for real life applications.

Normalized autocorrelation coefficients: Autocorrelation is a method which measures how similar a signal is to itself. As voiced speech has a predictable pattern, it is obvious that autocorrelation will be higher for voiced speech when compared to unvoiced speech, which is more random in nature. This parameter measures how similar a signal is to itself thereby indicating whether it has some sort of a pattern or not. The formula to calculate these coefficients is shown below:

$$C_1 = \frac{\Sigma\ s(n)s(n-1)}{\text{root}(\ \Sigma s^2(n)\Sigma s^2(n))}$$

Note that in the above formula, the summation in the numerator is from 1 to N, where N is the size of the interval taken. One summation in the denominator is from 1 to N whereas another is from 0 to N-1. Since there is a difference of 1 between the ranges of summation taken in the denominator, we say that the coefficients are one sample shifted. As the coefficients are normalized, they range from 0 to 1. Voiced speech tends to have coefficients which are closer to 1 as it has a more predictable pattern which is why the signal is similar to itself. Unvoiced speech tends to have coefficients which are low as it is more random and has very little similarity to itself.

This parameter captures the similarity/predictability of a signal to itself. The shortcoming of this parameter is that it tends to give a lot of details as for a very small sample, we will get a large number of peaks. We should note that autocorrelation works well even in the presence of noise.

Pre-emphasized energy ratio: Pre-emphasized energy ratio is another parameter using which we can distinguish between voiced and unvoiced parts of a speech. Pre-emphasizing tends to emphasize the high frequency components of a signal. We know that our speech signals are produced in a manner that low frequencies tend to have higher energy and pre-emphasis is done to emphasize the higher frequencies. This parameter basically measures how different adjacent parts of a signal are, and by using this it is able to distinguish between voiced and unvoiced speech. The formula for this parameter is given below:

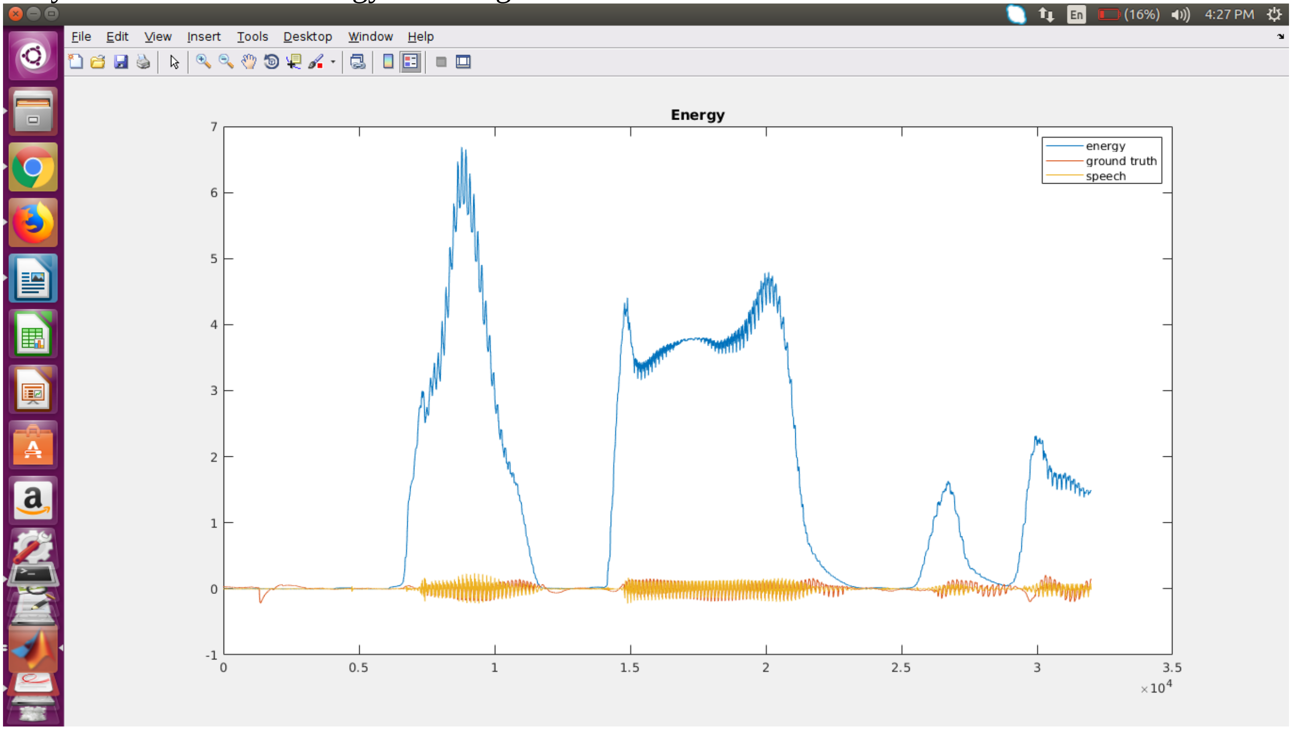
$$\frac{\Sigma \text{ abs}(s(i) - s(i-1))}{\Sigma s^2(i)}$$

In the above formula, abs() refers to the absolute value of the parameters inside the parantheses. We should note that this parameter will be less for voiced speech and more for unvoiced speech. This is intuitive as the numerator measures the difference between adjacent parts of a signal and this will be less for voiced speech (as it follows a more predictive pattern) and more for unvoiced speech (as it is more random in nature). Also, the denominator will be lesser for unvoiced speech as compared to voiced speech.

The shortcoming of this parameter is that it will not work that well in the presence of noise.

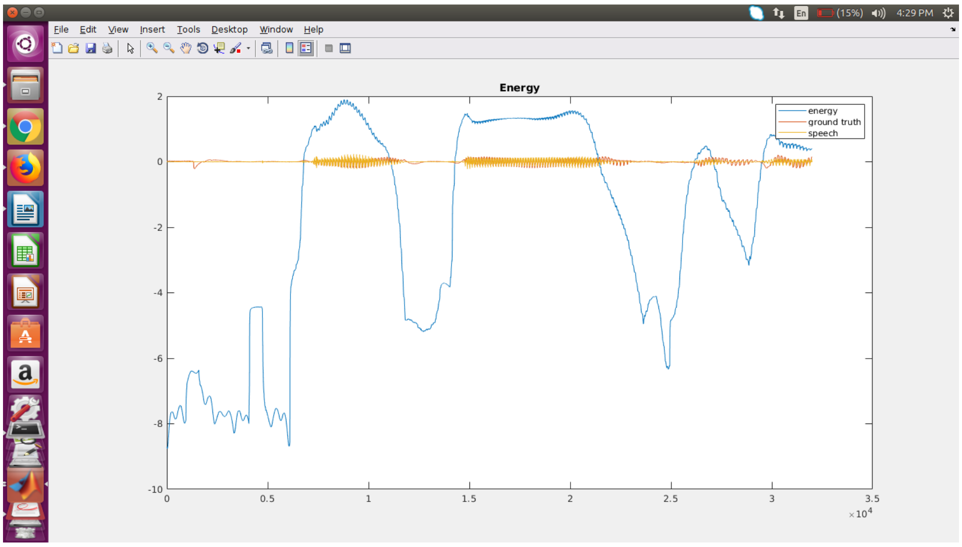
b) The plots for the various parameters along with the speech signal and the ground truth are given below. All the plots have used the '*arctic_a0001.wav*' file.

The plot for the energy parameter is shown below. The x-axis refers to the samples whereas the y-axis refers to the energy of the signal.



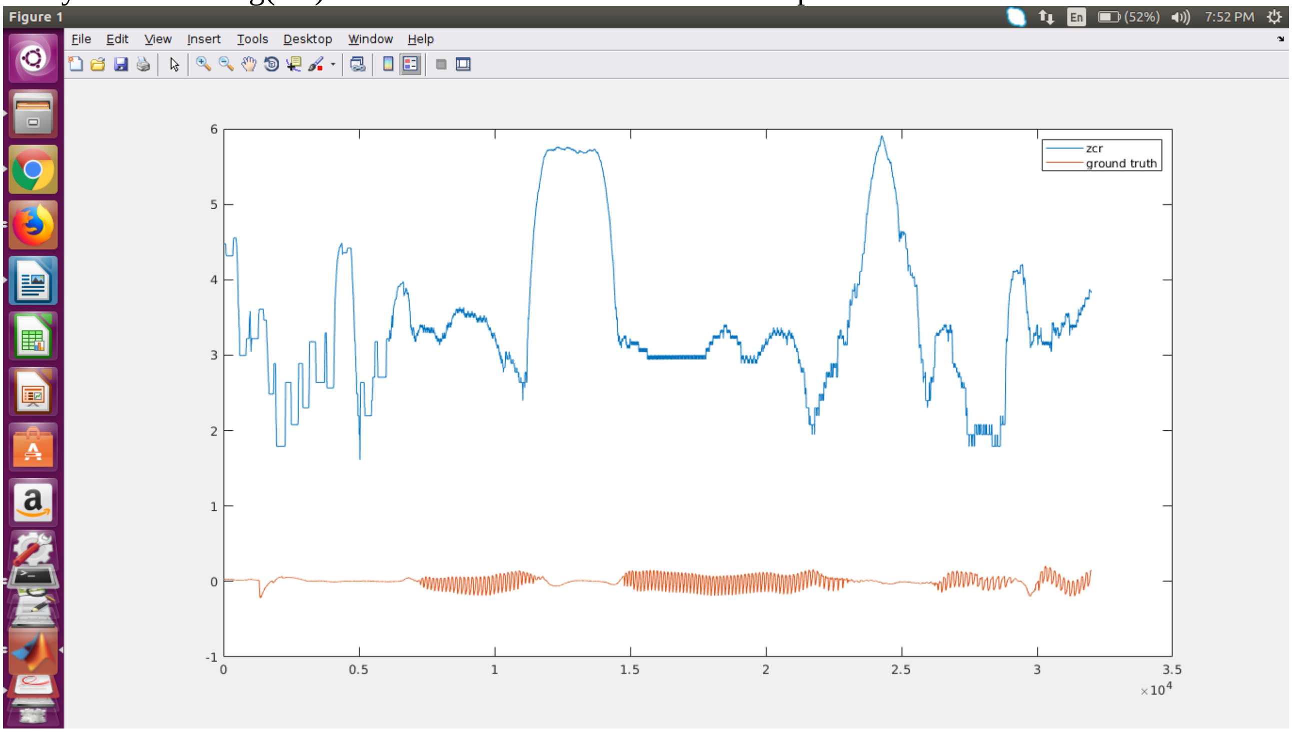
We can see that the energy values show peaks only when there is a voiced region and this is verified by checking the corresponding EGG values which were used as the ground truth. We can observe that the detection is almost accurate as the energy plot corresponds with the ground truth.

For the above graph, we can use 0.2 as the threshold to distinguish between voiced and unvoiced regions. Anything greater than the threshold is voiced and anything less than it is unvoiced. The plot for the log(energy) parameter is shown below.



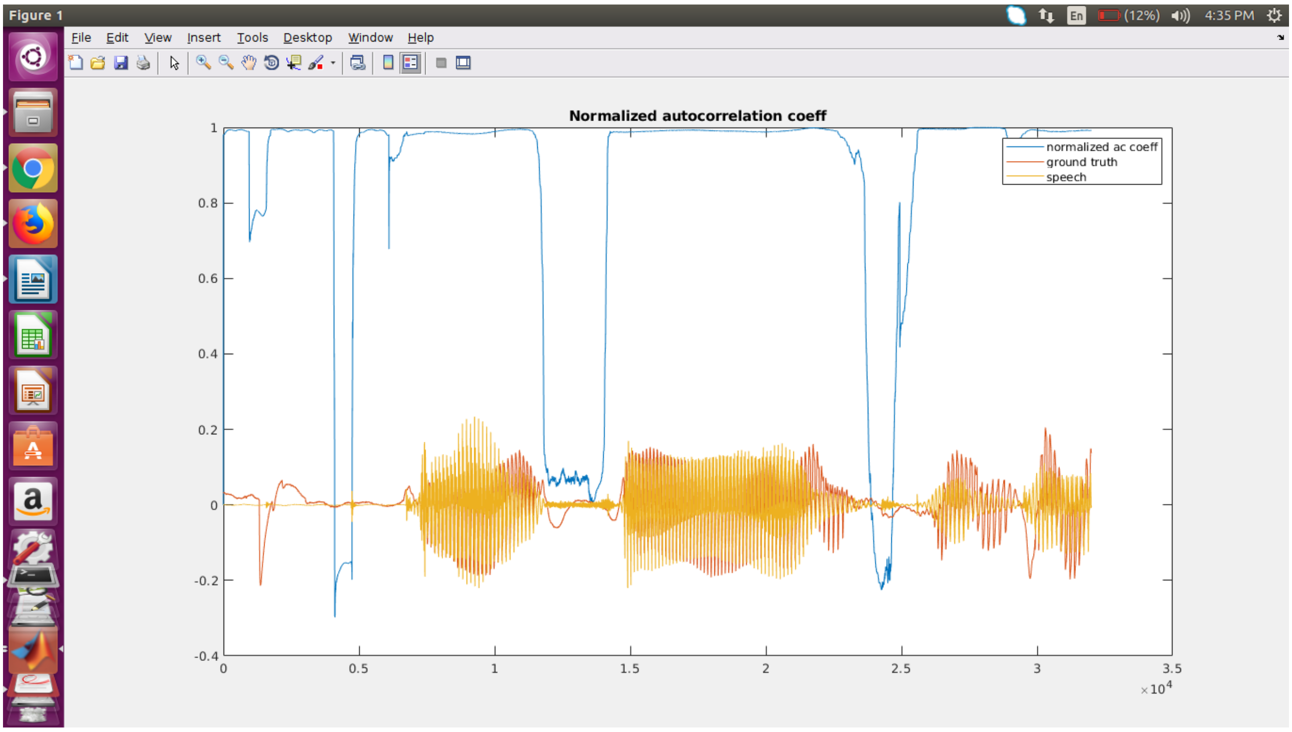
For the above graph we can observe that the log(energy) function crosses zero only when there are voiced regions present. We can therefore use 0 to be the threshold in this case.

The plot for the zero crossing rate is shown below. Note that I have plotted log(zcr) so as to be able to see the ground truth EGG values which are negligible compared to the actual zcr values. The y-axis is the log(zcr) values whereas the x-axis is the samples.



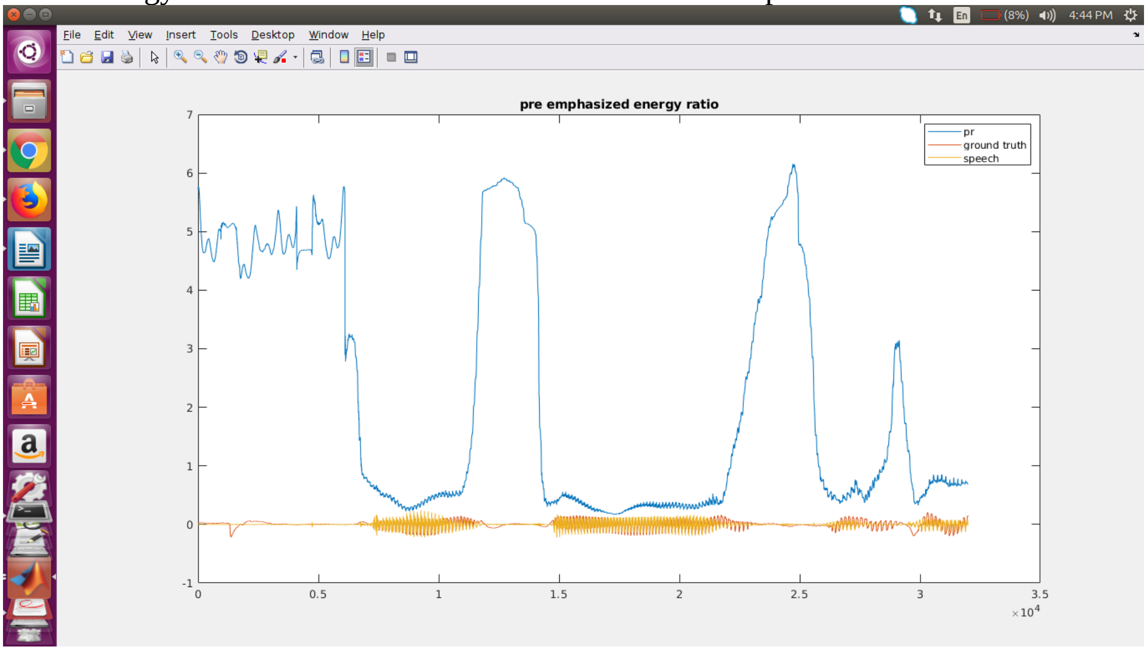
We can observe that the threshold can be kept 4 in this case. The regions which have a value less than 4 are voiced speech whereas those with values greater than 4 are unvoiced speech.

The plot for the normalized autocorrelation coefficients is shown below. The y-axis is the normalized autocorrelation coefficients whereas the x-axis is the samples.



The autocorrelation that I have plotted uses signals which differ by one sample shift. In the above plot, we can observe that the results are almost accurate. We can keep 0.85 to be the threshold. Anything above 0.85 is voiced region and anything below it is the unvoiced region.

The plot for the pre-emphasized energy ratio is shown below. The y-axis is the pre-emphasized energy ratio whereas the x-axis is the number of samples.



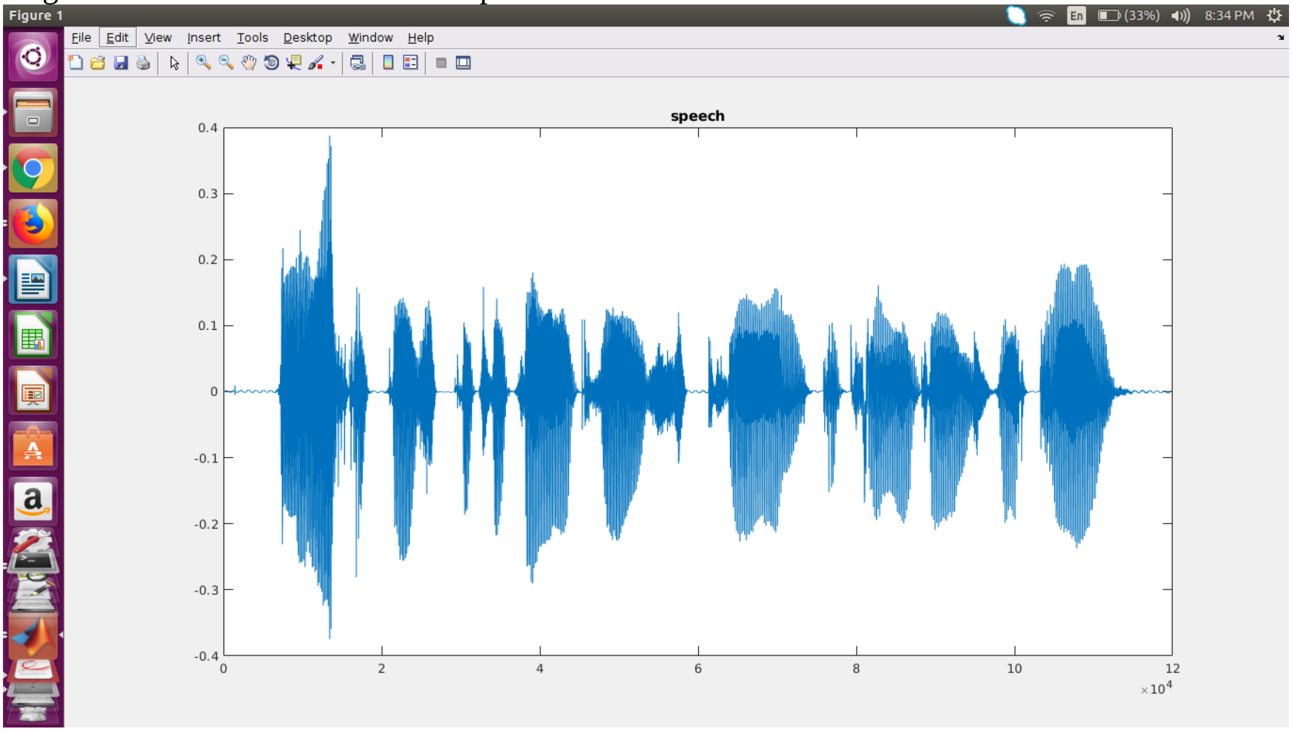
We can observe that in the above plot, the voiced regions are present where the energy ratio plot is less whereas the unvoiced regions are present where the energy ratio plot is high. We can clearly see that we can keep 0.8 as the threshold as all the voiced regions have a pre-emphasized energy ratio less than this and unvoiced regions have values greater than this.

c) Among the parameters, I feel that the normalized autocorrelation coefficients is doing the best job. This is as the normalized autocorrelation coefficients are unaffected by noise and this property is not inherent to the other parameters considered. If we take energy as our parameter, we know that it is not immune to noise and that it is also difficult to set the threshold for it. The same problem occurs for the zero crossing rate. However, as normalized autocorrelation coefficients are unaffected by noise and since we know that the voiced regions will have a coefficient close to 1 and unvoiced regions will have coefficients less than 1 (the problem of threshold is not as pronounced in this case), this parameter becomes suitable for our use. The problem with the pre-emphasized energy ratio is that we have to set a threshold for it.

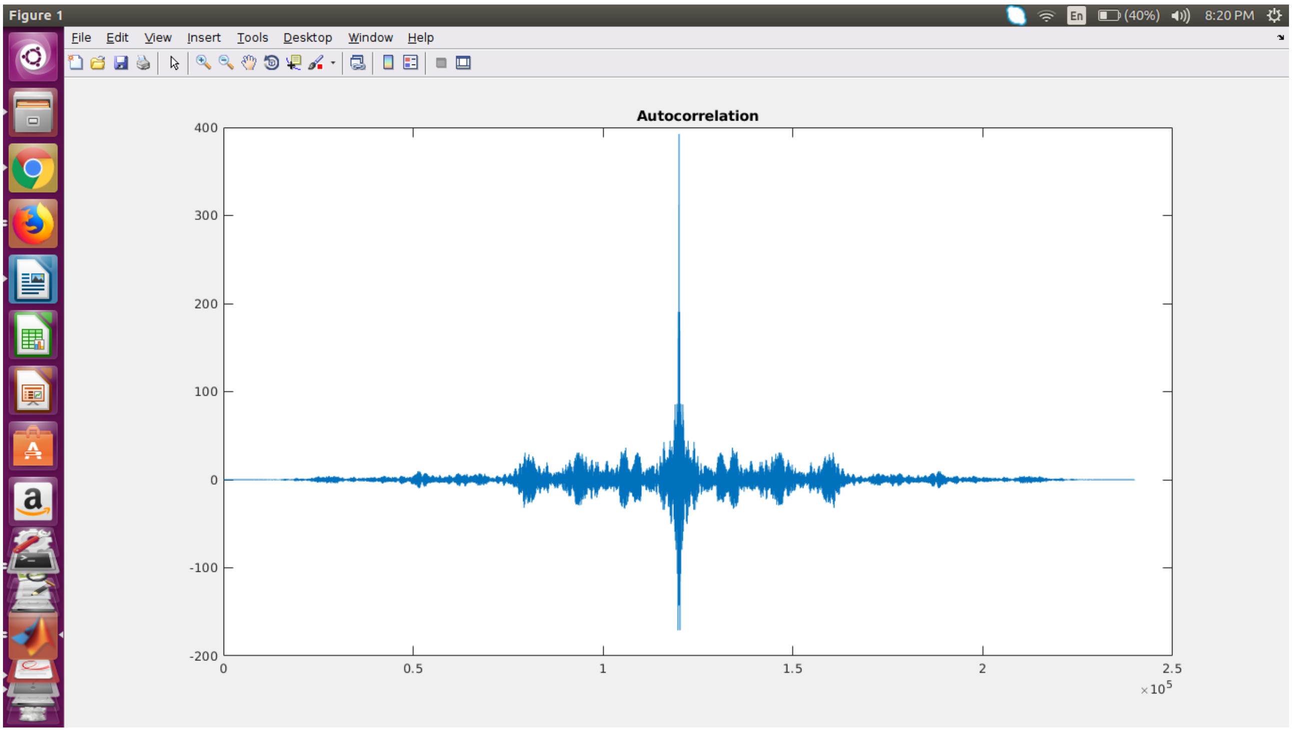
Question 2

The pitch period and fundamental frequency can be calculated by using the autocorrelation method. We know that the autocorrelation method measures how similar a signal is to itself. When we do autocorrelation to a signal, we get peaks in the autocorrelation graph which come after a period. The intituition for this method is that the time period between successive peaks will be the pitch period T and the fundamental frequency will be the inverse of the pitch period (= 1/T). The only shortcoming with this method is that there may be some spurious peaks which lead us to calculate the pitch period and hence the fundamental frequency incorrectly.

For the file '*arctic_0002.wav*', the speech signal is shown below. The y-axis is the signal magnitude and the x-axis are the samples.



The autocorrelation function for the corresponding signal has been computed and plotted. It is as shown below. The y-axis is the magnitude of the autocorrelation and the x-axis are the samples.



After computing the autocorrelation of the signal, we find the peaks or the local maximas present in the autocorrelation by using the *findpeaks* function in MATLAB. We then find the maximum of these peaks and compute the time difference between the maximum of the peaks and the peak right after it. The difference between these peaks will give us the number of samples between the peaks which was 111. We can get the pitch frequency by using the below formula:

pitch frequency = sampling rate/ number of samples between consecutive peaks

The sampling rate for this audio file was 32000 Hz so the pitch frequency will be 32000/111 = 288.29 Hz. We can get the pitch period from this value by taking the inverse of the pitch frequency.

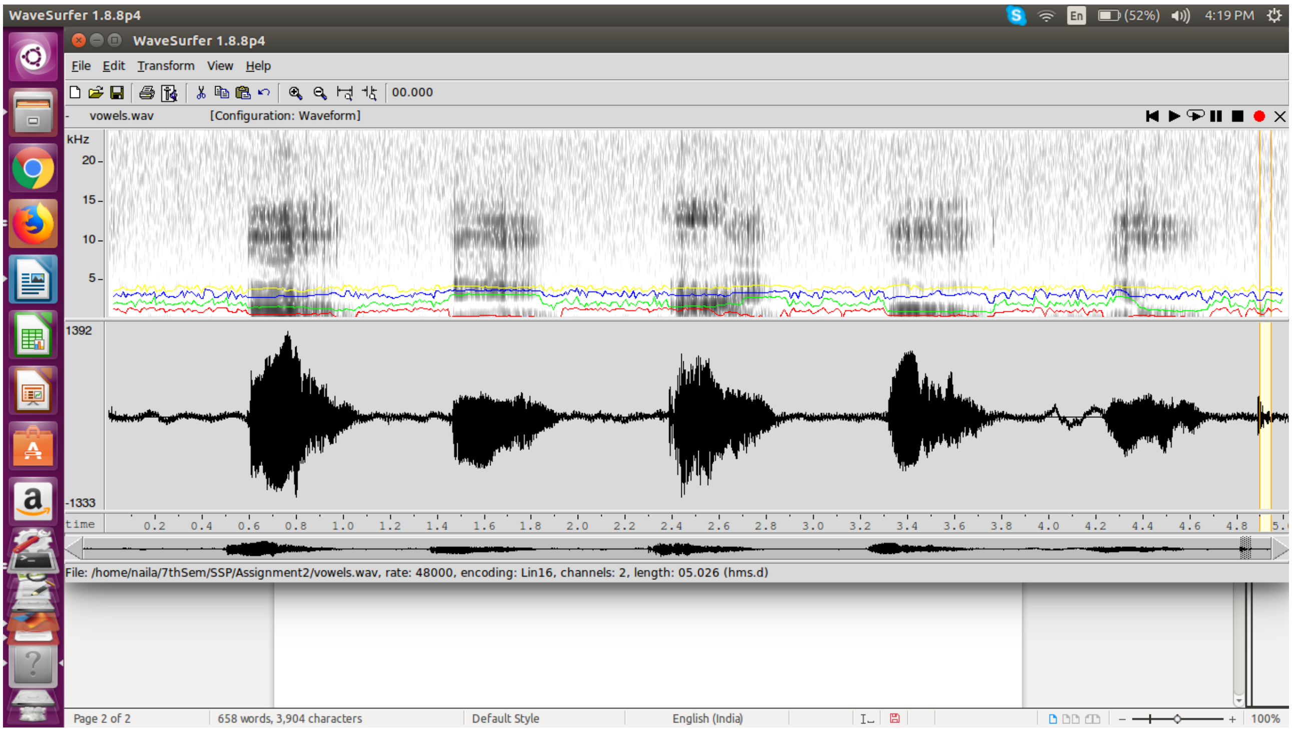
Pitch period = 1/pitch frequency

Using the above formula, we get the pitch period to be 0.0035 seconds or 3.5 ms.

Similarly, if we run the code for ‘*arctic_0003.wav*’ file we get the pitch frequency to be 209.15 Hz and pitch period to be 0.0048 seconds or 4.8 ms.

Question 3

I have recorded the vowels ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’ (in this order) in the *vowels.wav* file. The formant plot for the vowels is shown below:



File: /home/naila/7thSem/SSP/Assignment2/vowels.wav, rate: 48000, encoding: Lin16, channels: 2, length: 05.026 (hms.d)

The formant frequencies f1, f2, f3 and f4 for the vowel ‘a’ are 657 Hz, 1587 Hz, 2907 Hz and 3711 Hz. The corresponding formant frequencies for the vowel ‘e’ are 435 Hz, 3165 Hz, 3676 Hz and 4093 Hz. The corresponding formant frequencies for the vowel ‘i’ are 334 Hz, 2786 Hz, 3372 Hz and 4308 Hz. The corresponding formant frequencies for the vowel ‘o’ are 460 Hz, 927 Hz, 3296 Hz and 3929 Hz. The corresponding formant frequencies for the vowel ‘u’ are 382 Hz, 1132 Hz, 3100 Hz and 4122 Hz.

	F1 (in Hz)	F2 (in Hz)	F3 (in Hz)	F4 (in Hz)
a	657	1587	2907	3711
e	435	3165	3673	4093
i	334	2786	3372	4308
o	460	927	3296	3929
u	382	1132	3100	4122

If we arrange the vowels based on the ascending order of f1 frequencies, we will get the order: ‘i’, ‘u’, ‘e’, ‘o’ and ‘a’.

If we arrange the vowels based on the ascending order of f2 frequencies, we will get the order: ‘o’, ‘u’, ‘a’, ‘i’, ‘e’.

If we arrange the vowels based on the ascending order of f3 frequencies, we will get the order: ‘a’, ‘o’, ‘u’, ‘i’ , ‘e’

We know that the formants are the resonant frequencies of the vocal tract and that they can be changed by changing the positions of the tongue and lips as this changes the frequencies at which the vocal tract vibrates at.

The first formant F1 is inversely related to the vowel height. Therefore, the higher the position of the tongue when saying the vowel, the lower is the first formant. Because of this, high vowels such as ‘i’ have a lower first formant whereas low vowels such as ‘a’ and ‘ae’ have a higher first formant. We can see this information reflected in the chart above as ‘i’ has a low first formant, ‘a’ has a high first formant and mid-high vowels like ‘u’ have a slightly low first formant and mid-low vowls like ‘e’ have a slightly higher first formant.

The second formant F2 is related to the tongue backness. This is as vowels which require the tongue to be more front tend to have a higher second formant. This fact indicates why front vowels such as ‘i’, ‘e’ and ‘ae’ have a higher second formant and why back vowels such as ‘u’ have a lower second formant. We should also observe that the difference between F1 and F2 indicates the degree of backness of the vowel. The closer F1 and F2 are to each other, the more back a vowel is. For example for both ‘o’ and ‘u’, the difference between F1 and F2 is less than 1000 Hz indicating that they are back vowels.

The third formant F3 is somewhat related to lip rounding. This is as the more rounded the lips become in order to articulate a vowel, the lower the third formant frequency will be. This explains why ‘o’ and ‘u’ have a low F3 since we need to round our lips in order to articulate these vowel. We should note that lip rounding tends to decrease F2 also, but to a small extent.

Question 4

The consonant pairs ‘p,b’, ‘t,d’ and ‘k,g’ have been recorded in the VCV format in the corresponding files: ‘*p_b.wav*’, ‘*t_d.wav*’ and ‘*k_g.wav*’. The corresponding transcriptions are in the files ‘*p_b.lab*’, ‘*t_d.lab*’ and ‘*k_g.wav*’. The formants in the consonant regions were observed for all the pairs.

P-b pair:

	‘p’	‘b’
F1	836 Hz	778 Hz
F2	1397 Hz	1463 Hz
F3	2690 Hz	2766 Hz
F4	4336 Hz	4332 Hz

We can observe that while F1 and F4 are slightly more for ‘p’, F2 and F3 are slightly more for ‘b’.

T-d pair:

	‘t’	‘d’
F1	968 Hz	978 Hz
F2	1654 Hz	1771 Hz
F3	2971 Hz	3028 Hz
F4	4265 Hz	4025 Hz

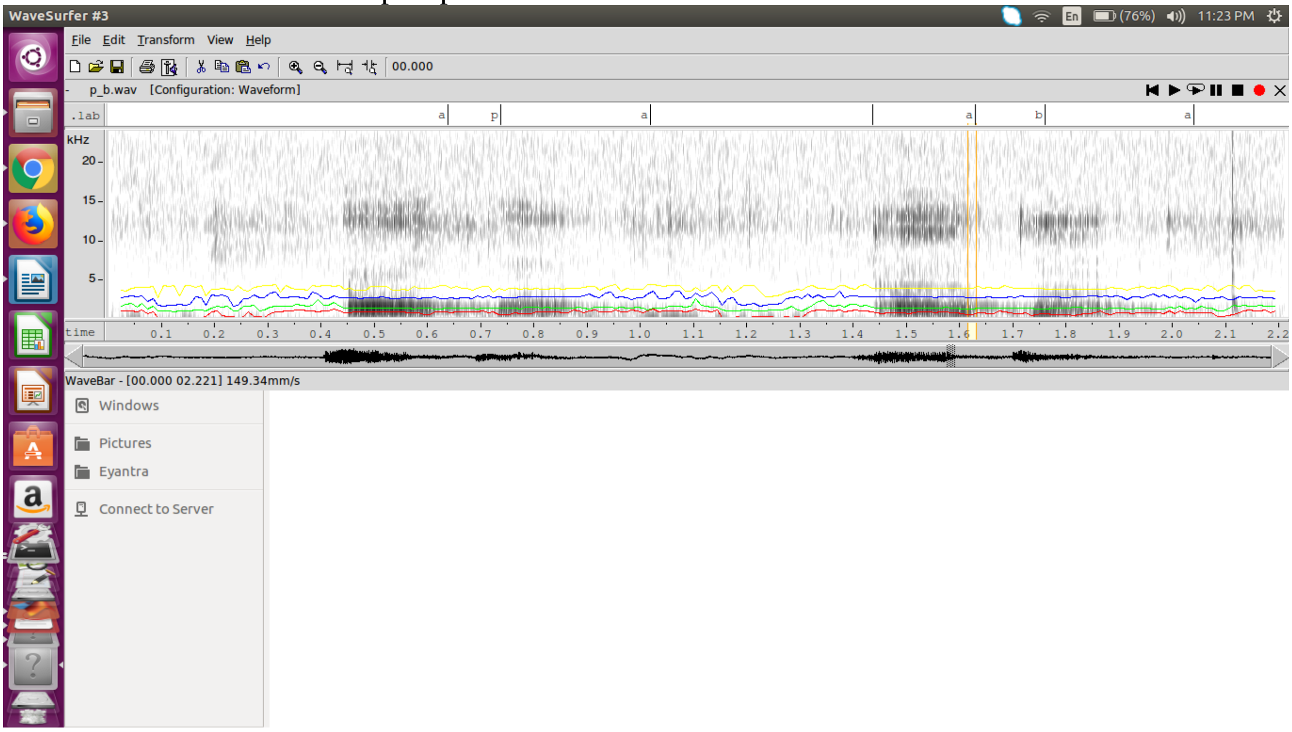
b We can observe that F1,F2 and F3 are higher for ‘d’ but F4 is higher for ‘t’.

K-g pair:

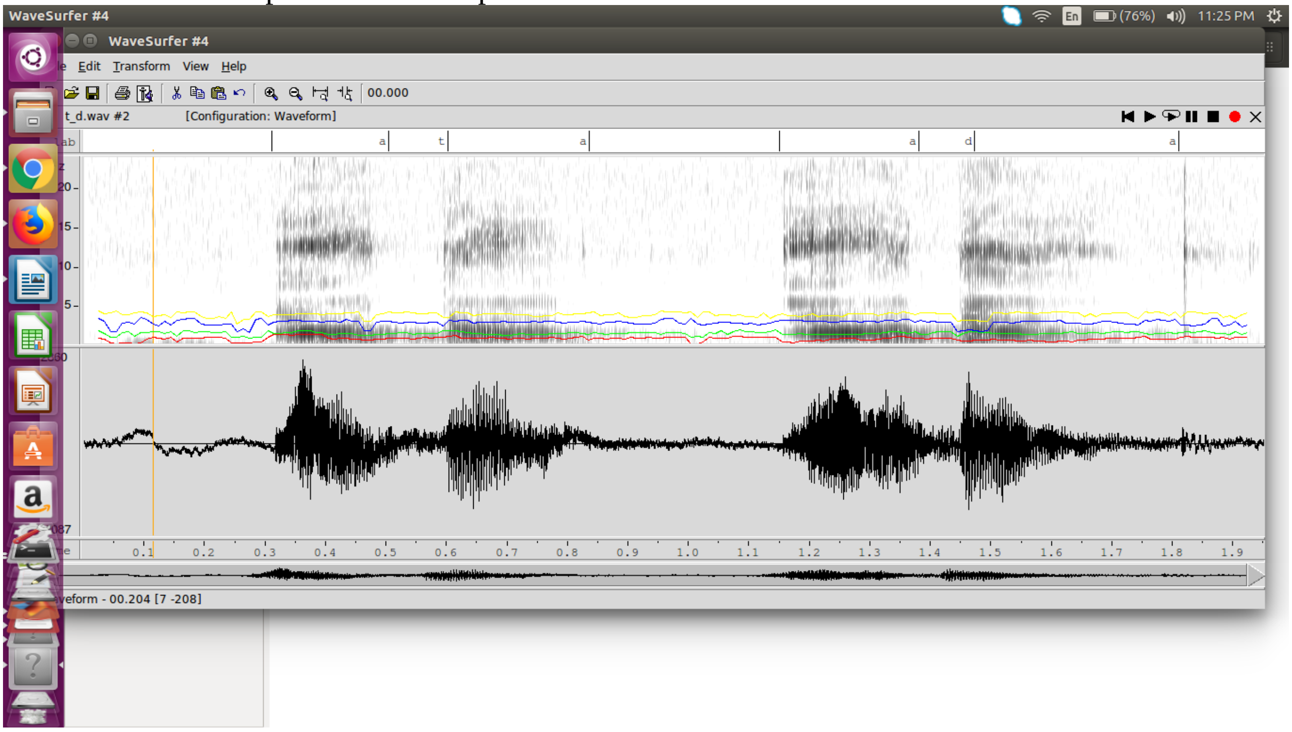
	‘k’	‘g’
F1	1000 Hz	933 Hz
F2	1523 Hz	1522 Hz
F3	2881 Hz	2919 Hz

F4	3998 Hz	3987 Hz

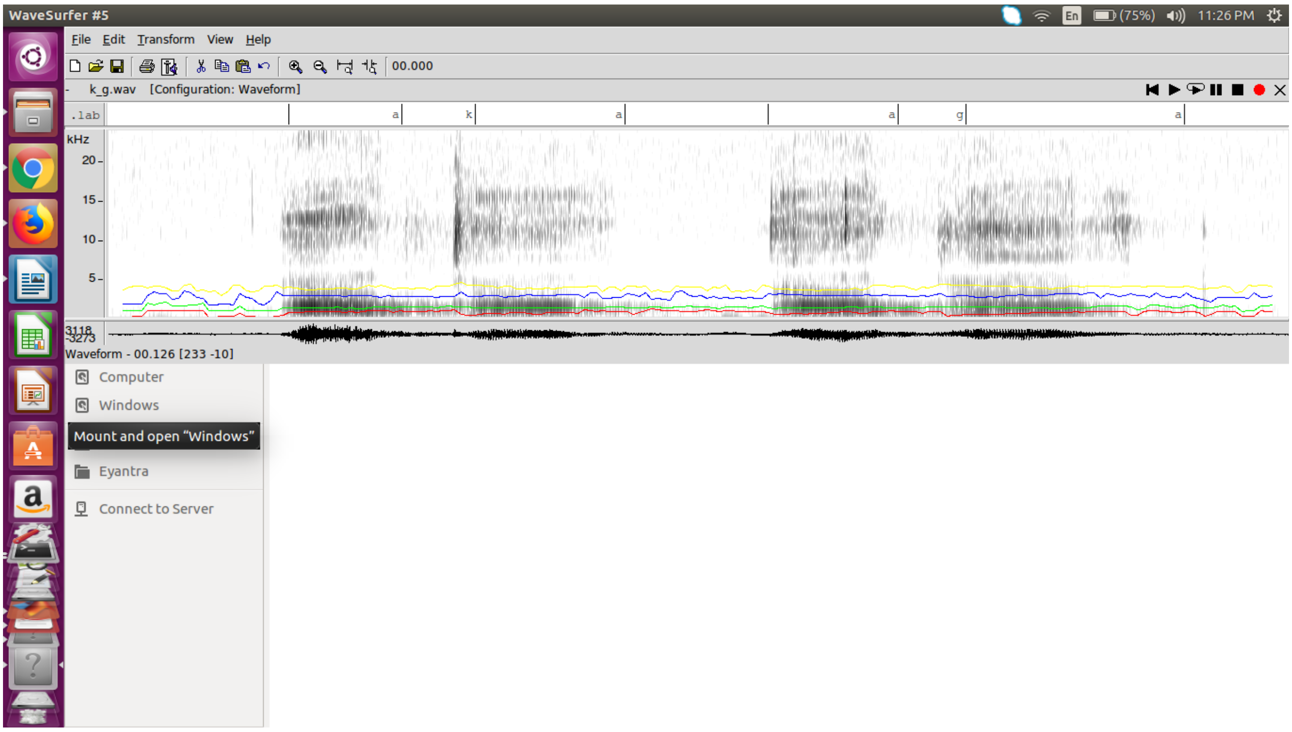
We can observe that F1, F2 and F4 are higher for ‘k’ but F3 is higher for ‘g’.
The formants for the ‘p-b’ pair are shown below:



The formant plot for the ‘t-d’ pair is shown below:



The formant plot for the ‘k-g’ pair is shown below:

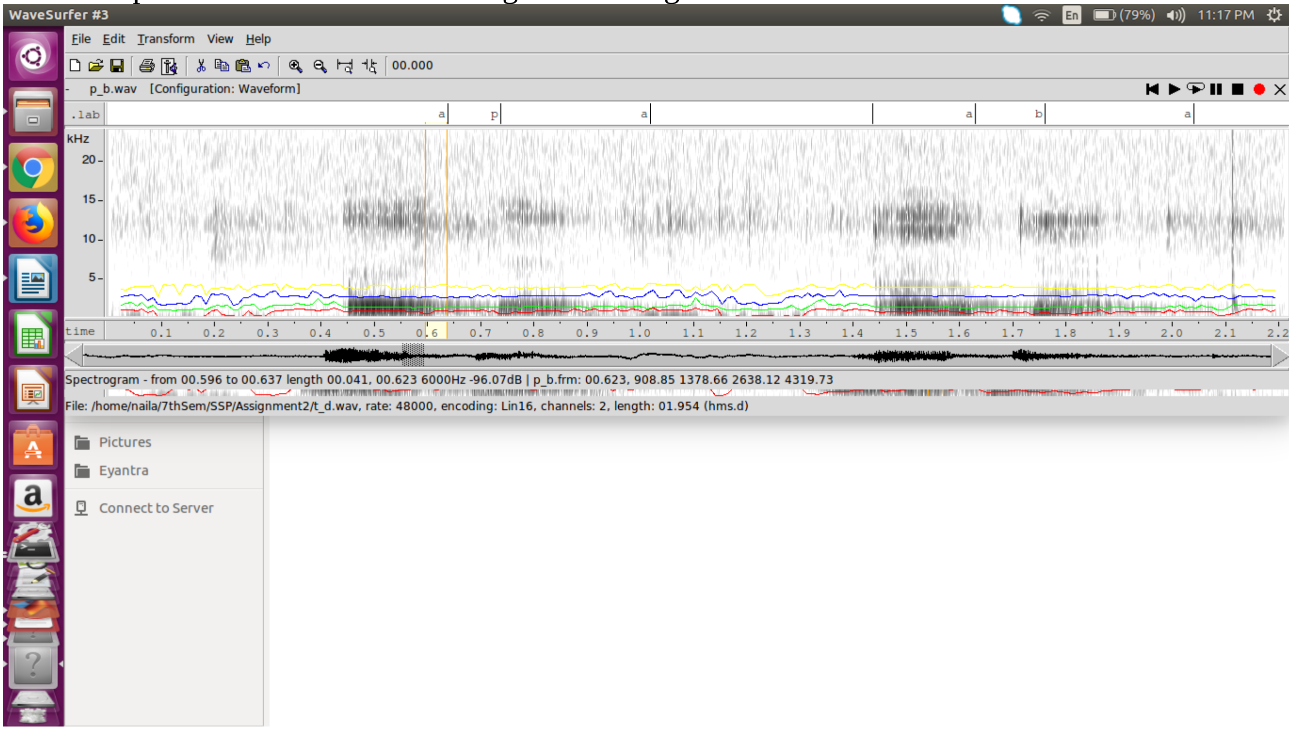


Waveform - 00.126 [233 -10]

Voice onset time or VOT is a feature of the production of stops like 't'- which are a type of consonants. It is defined as the length of time that passes between the release of a stop consonant and the onset of voicing or the vibration of the vocal folds. The 'p' and 'b' pairs are bilabial, the 't' and 'd' pairs are alveolar and the 'k' and 'g' pairs are velar. We should observe that these sounds are made when there is a closure in the mouth so that air is blocked for a small interval of time after which it is suddenly released causing that sound. The voice onset times for the following pairs are given below.

<u>P-b pair:</u> 'p' has a VOT of 41 ms whereas 'b' has a VOT of 16 ms (close to 20ms).

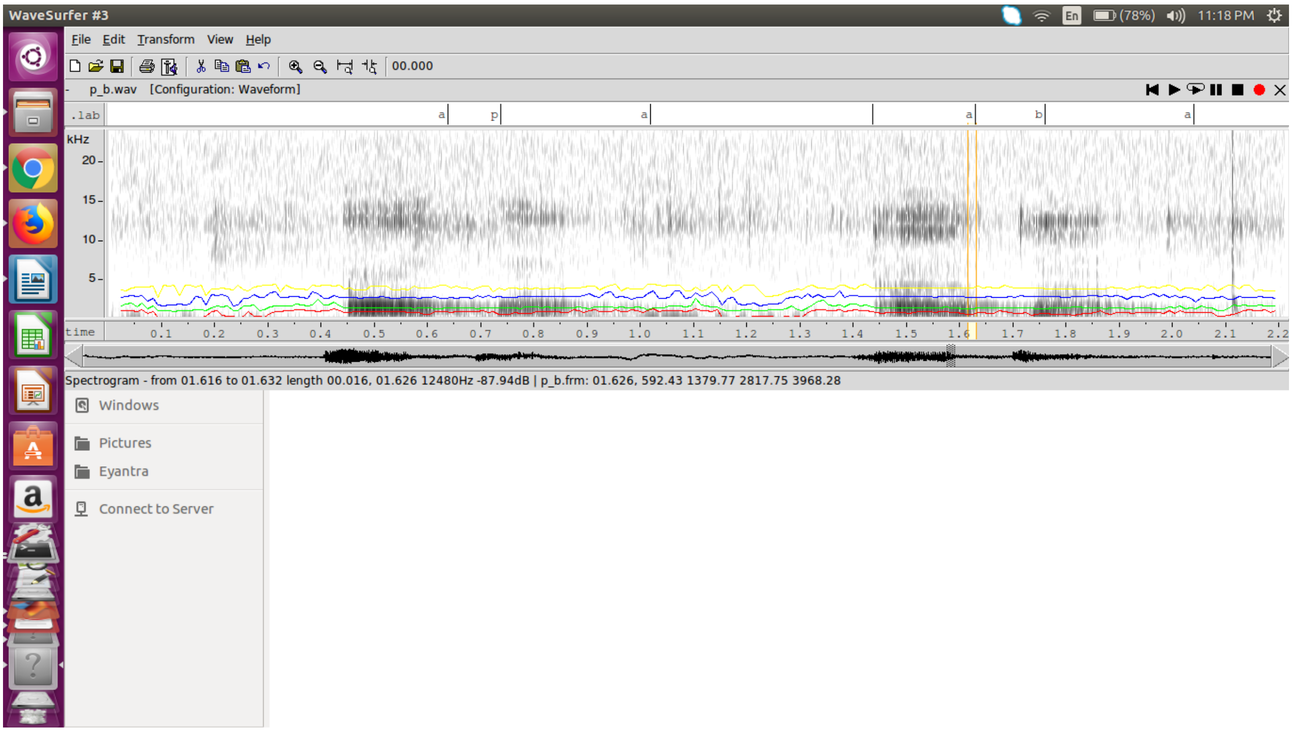
VOT for 'p': Note that time interval is given in 'length' attribute.



Spectrogram - from 00.596 to 00.637 length 00.041, 00.623 6000Hz -96.07dB | p_b.frm: 00.623, 908.85 1378.66 2638.12 4319.73

File: /home/naila/7thSem/SSP/Assignment2/t_d.wav, rate: 48000, encoding: Lin16, channels: 2, length: 01.954 (hms.d)

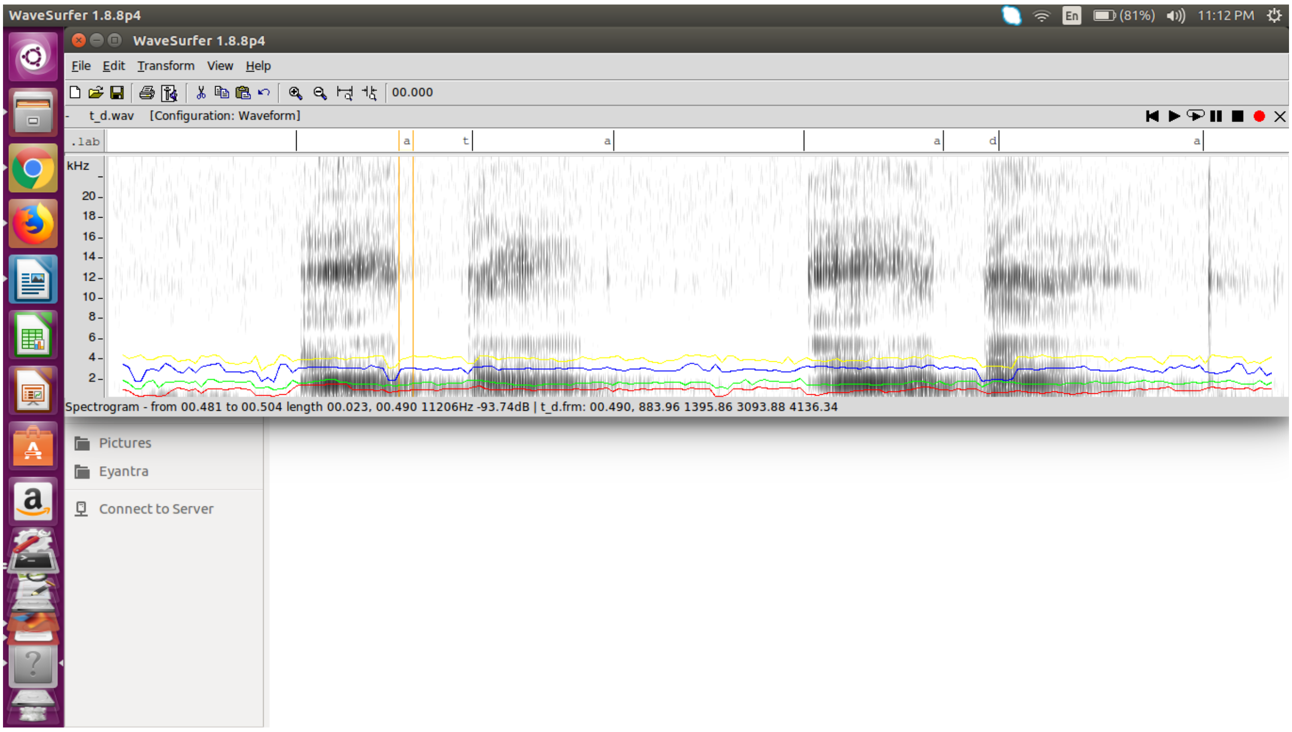
VOT for ‘b’:



```
Spectrogram - from 01.616 to 01.632 length 00.016, 01.626 12480Hz -87.94dB | p_b.frm: 01.626, 592.43 1379.77 2817.75 3968.28
```

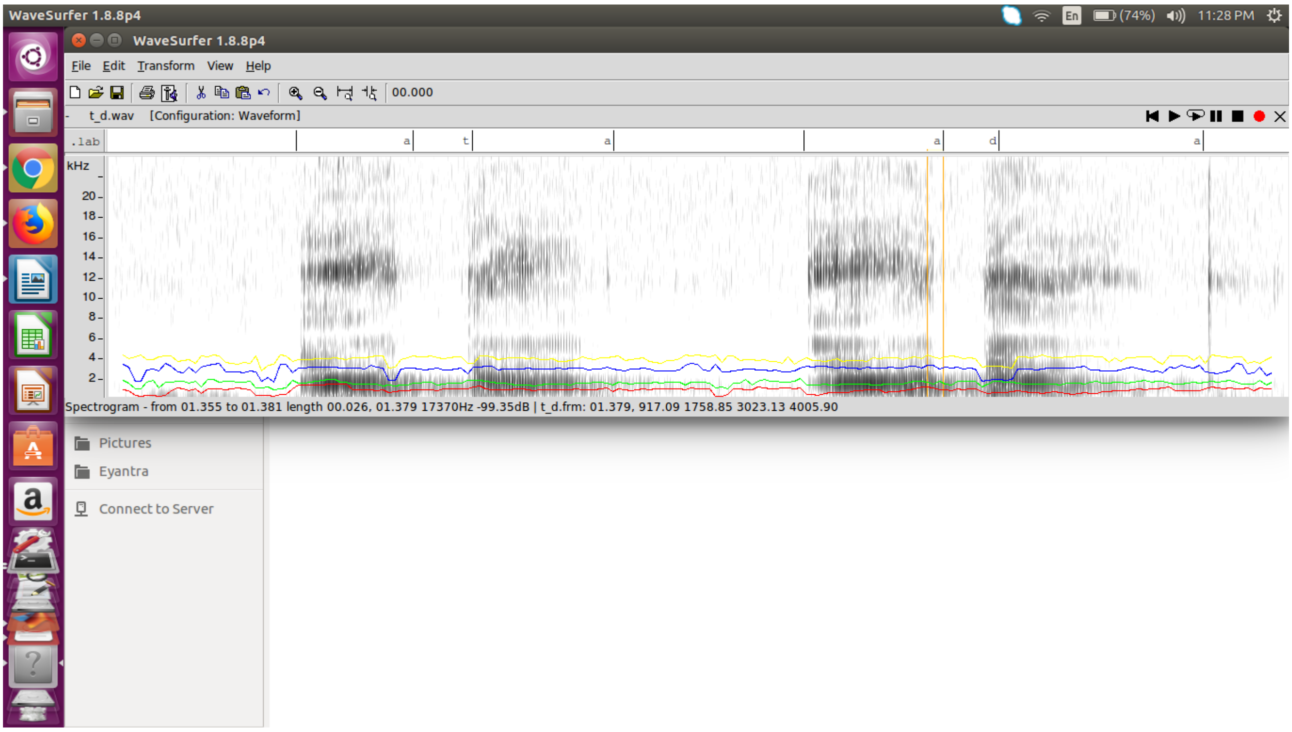
<u>T-d pair:</u> ‘t’ has a VOT of 23 ms whereas ‘d’ has a VOT of 26 ms

VOT for ‘t’:



```
Spectrogram - from 00.481 to 00.504 length 00.023, 00.490 11206Hz -93.74dB | t_d.frm: 00.490, 883.96 1395.86 3093.88 4136.34
```

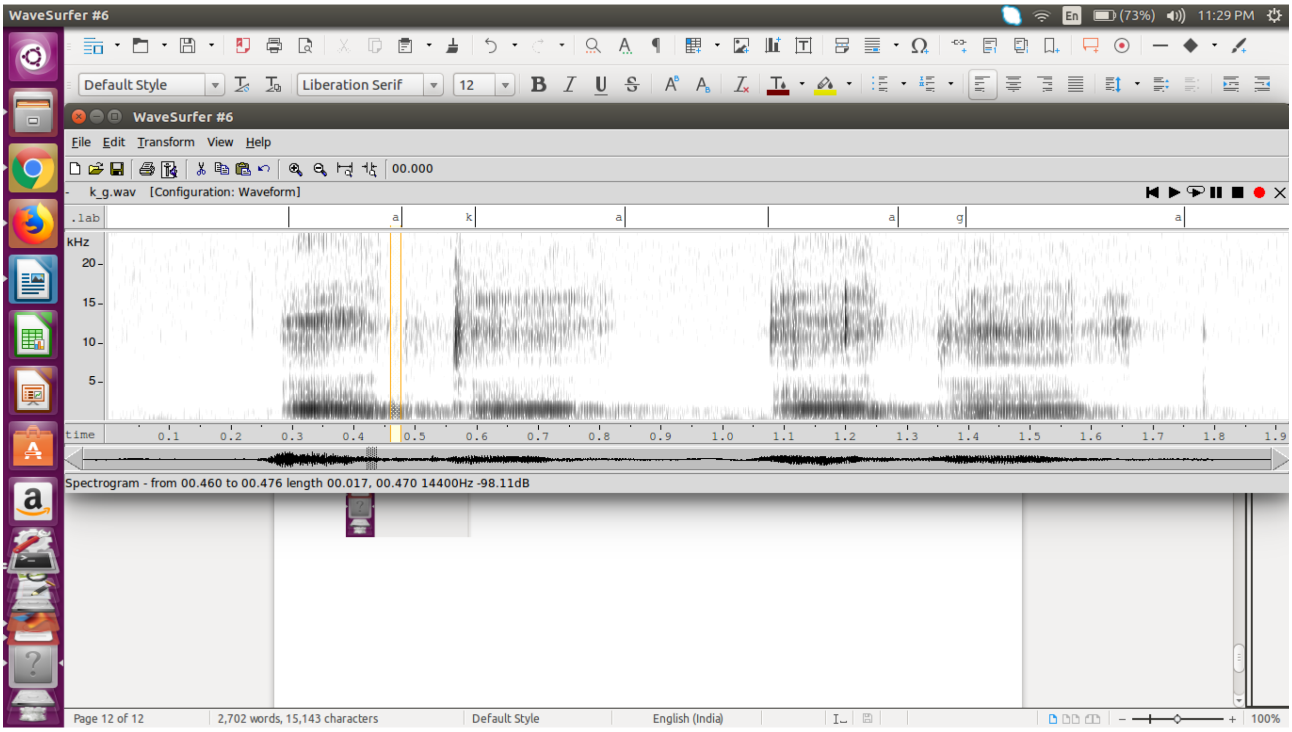
VOT for ‘d’:



Spectrogram - from 01.355 to 01.381 length 00.026, 01.379 17370Hz -99.35dB | t_d.frm: 01.379, 917.09 1758.85 3023.13 4005.90

K-t pair: The VOT for 'k' is 17 ms whereas the VOT for 'g' is 21 ms

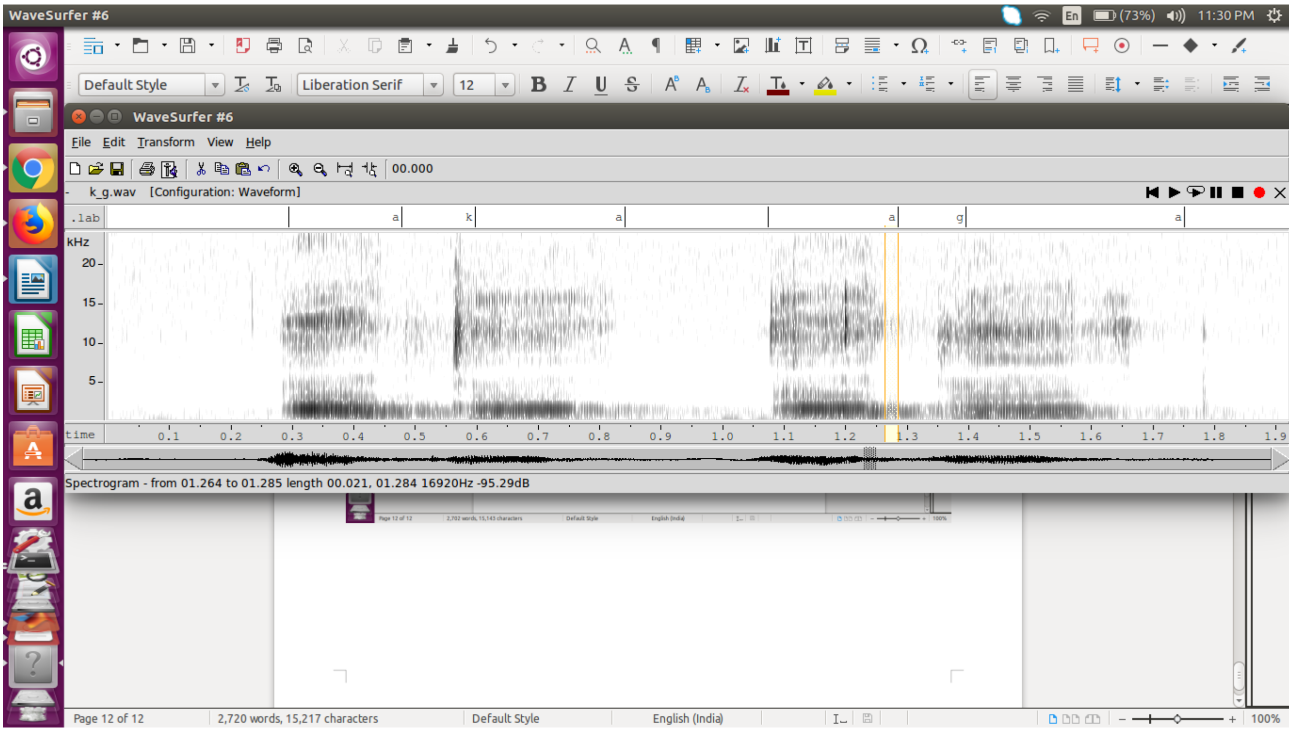
VOT for 'k':



Spectrogram - from 00.460 to 00.476 length 00.017, 00.470 14400Hz -98.11dB

Page 12 of 12 2,702 words, 15,143 characters Default Style English (India) 100%

VOT for 'g':



Spectrogram - from 01.264 to 01.285 length 00.021, 01.284 16920Hz -95.29dB