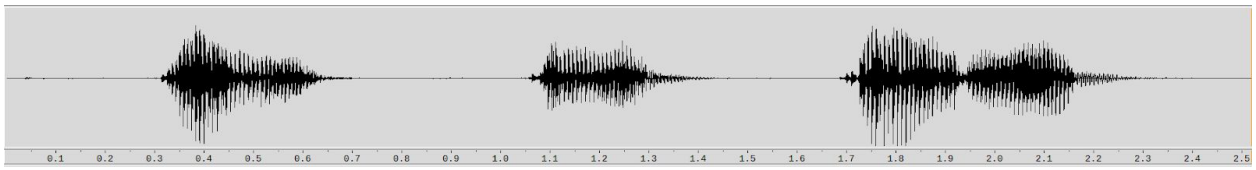
`

Question 1



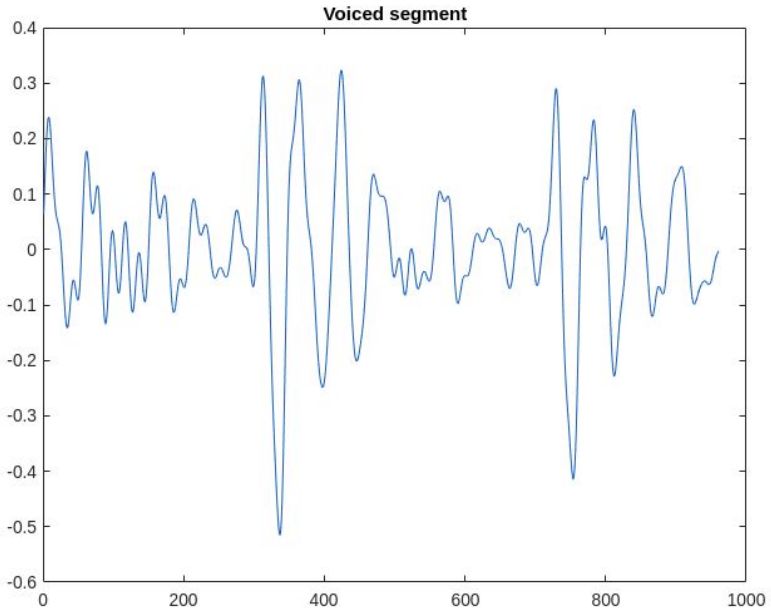
Part a

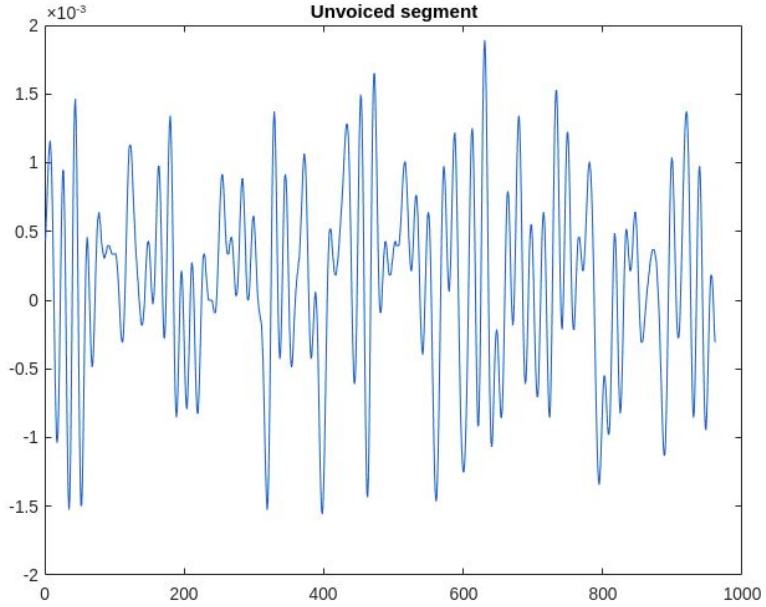
I have marked voiced and unvoiced regions through wavesurfer and I have attached the annotated transcriptions in the zip folder.

Part b

I selected two 20ms segments from the recorded audio (mentioned in matlab) and did all 3 sub parts for those 2 segments. The sampling frequency (Fs) is 48000. For voiced part I selected samples [85,000 to 85960] which correspond to time [1.77s to 1.79s]. For the unvoiced part I selected samples [80,000 to 80,960] which correspond to time [1.66s to 1.68s].

`





Part 1

Zero Crossing of Voiced Part came out to be 37 and unvoiced part came out to be 78. So as expected due to random behaviour in the unvoiced part (non-periodic), the number of times signal crosses 0 is more than that in the voiced part (periodic). This is due to quasi-stationary

`

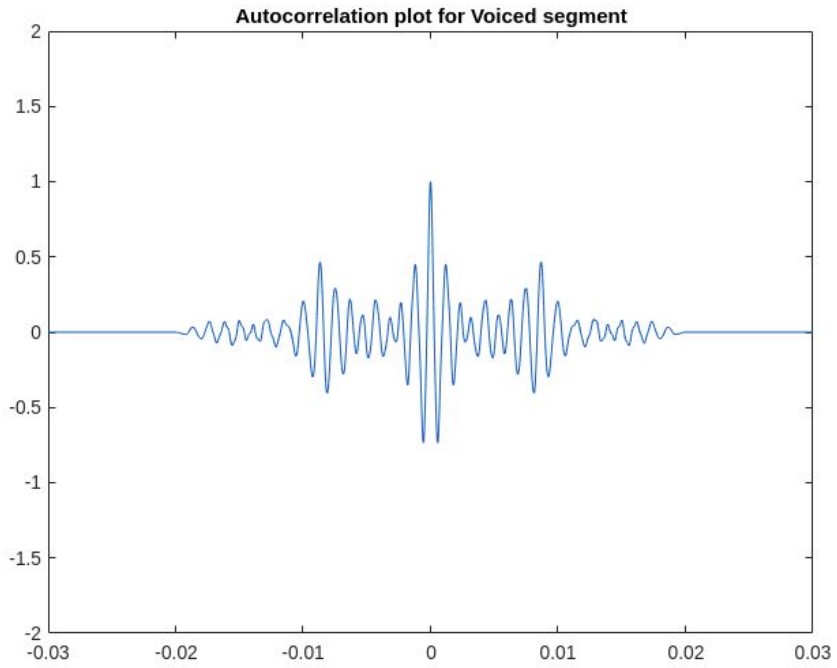
behaviour of speech in voiced regions. We can use this criteria to distinguish between voiced and unvoiced regions but if the audio has noise, this method is not so efficient.

Part 2

Energy of the Voiced part is approximately 16.4 whereas energy of unvoiced part is 0.00043. We see the energy of the Unvoiced part is very less as they are regions which are not intentionally spoken by a person. Thus these regions don't have energy. . We can use this criteria to distinguish between voiced and unvoiced regions but if the audio has noise, this method is not so efficient. Also some vowels or semi-vowels have low energy and thus deciding threshold is difficult.

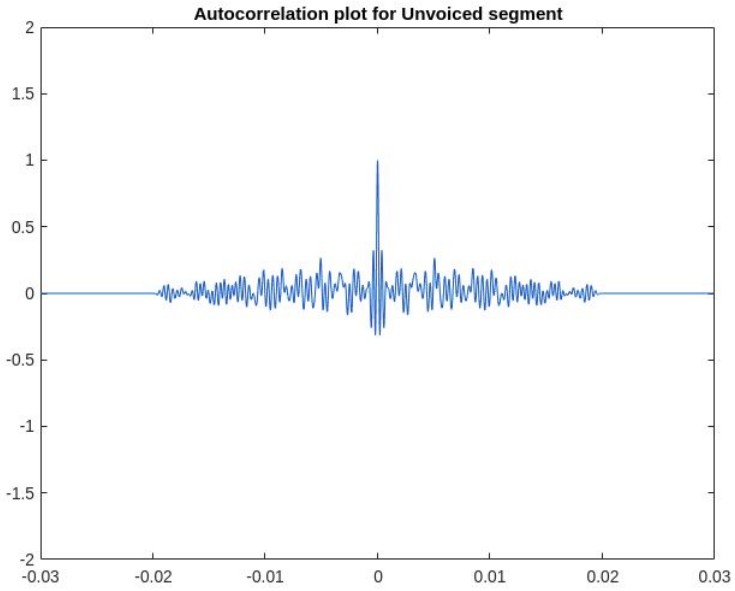
Part 3

Autocorrelation is used to determine Pitch of audio signal. The reciprocal of distance between two regular (high-strength) peaks in an autocorrelation plot is Pitch Frequency.



In the above figure we can clearly see there are more than one significant peak and thus we can get some defined pitch in the voiced part. This happens due to periodicity or repetitive nature of signal in the voiced part.

`



In the above figure we can see, there is only one significant peak showing non-periodic nature in an unvoiced region.
This approach can be used to separate voiced and unvoiced regions.

Question 2

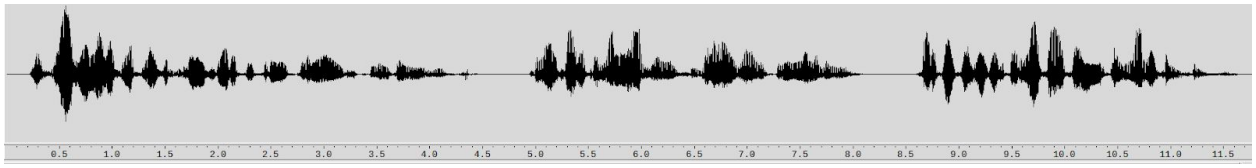
Epochs are instants of significant excitation or glottal closure instants or location of impulses of air entering in the glottal cavity. They are used to model the Vocal tract system as an LTI system where the air entering the glottal cavity is considered as impulses or epochs. Air goes to vocal folds as impulse trains. Electroglottography (EGG) is used to record vocal fold vibration which gives information about the vocal tract system of a person. The negative peaks in a differenced EGG graph are epoch locations.

Epoch based analysis of speech helps in segmenting speech on the basis of speech production characteristics, glottal vibrations, formants, etc. We can manipulate prosody features, characterize voice quality features and enhance speech using epochs. We can determine fundamental frequency using epoch locations.

The example or application of epochs is EGG. Epoch-based speech analysis is used for glottal activity detection. The strength of impulses during glottal activity is determined by rate of closure of vocal folds in each glottal cycle.

`

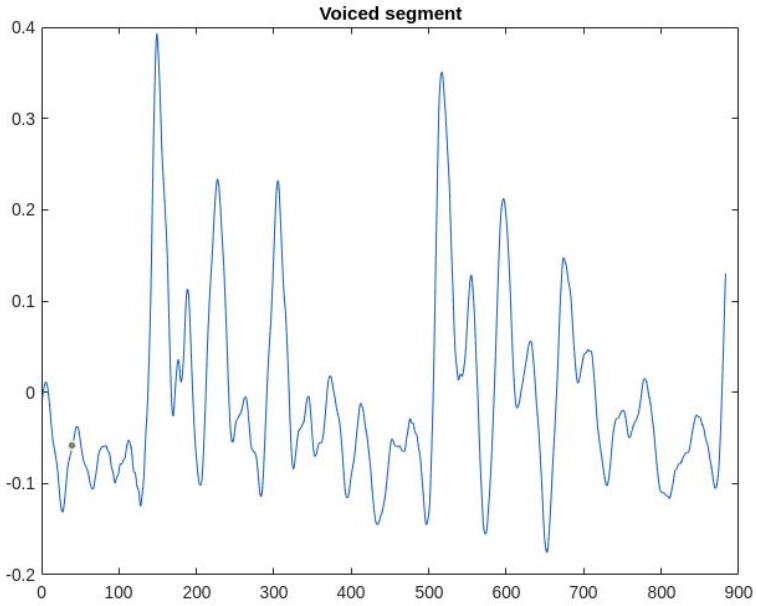
Question 3



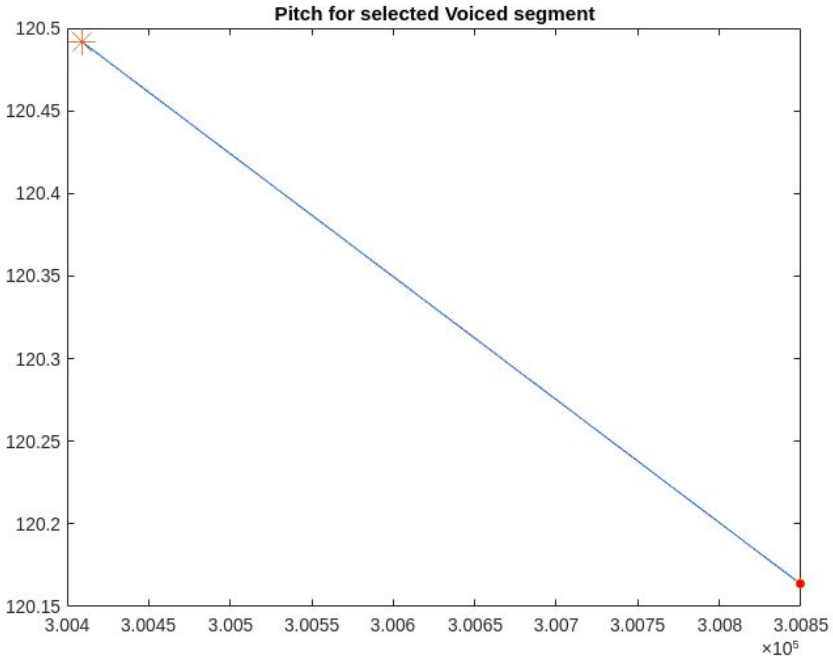
Part a

I have marked voiced and unvoiced regions through wavesurfer and I have attached the annotated transcriptions in the zip folder.

Part b

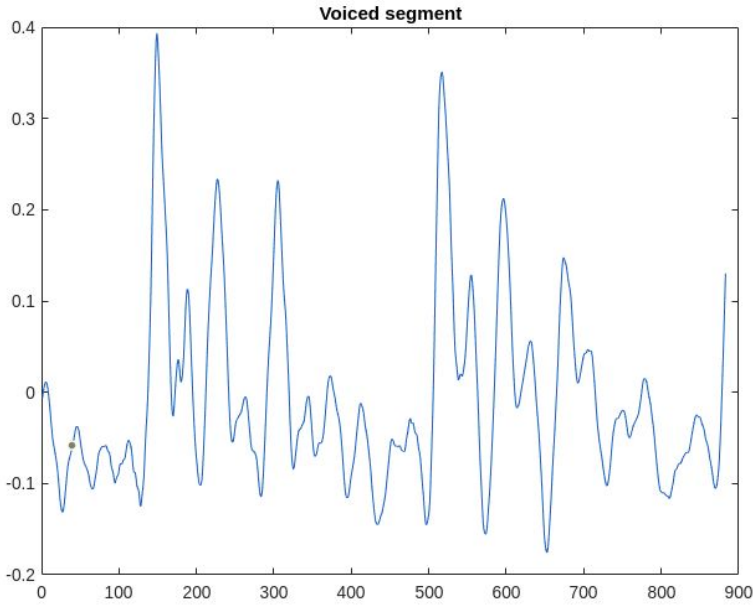


`

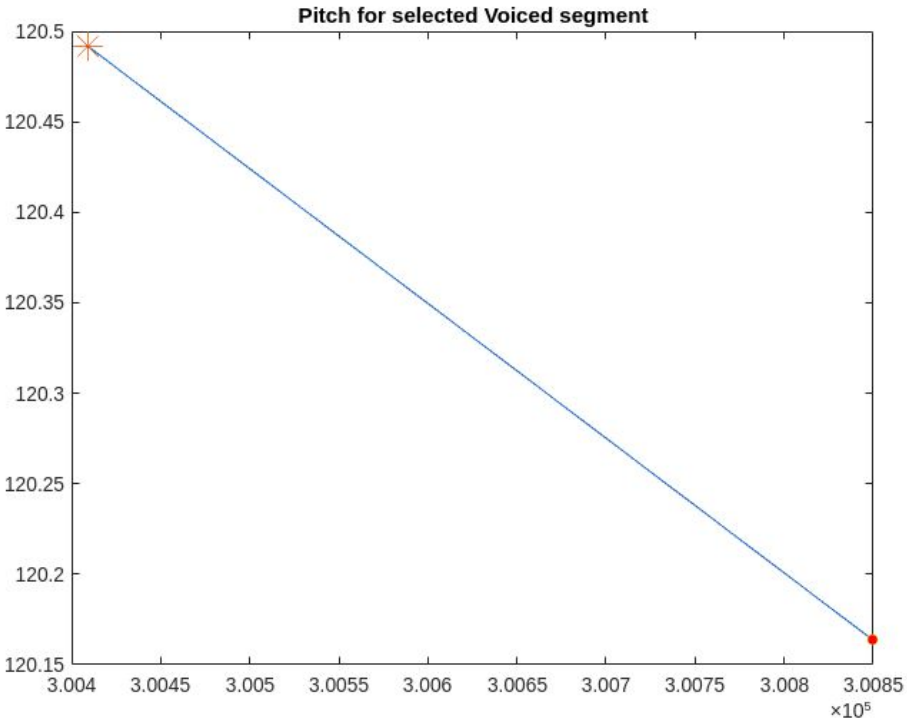


We can clearly see in the above figure that ‘ * ’ shows a maximum Pitch of 120.49 in the selected Voiced segment. The minimum Pitch of 120.16 marked with ‘ o ‘ in the selected Voiced segment.

Part c



`



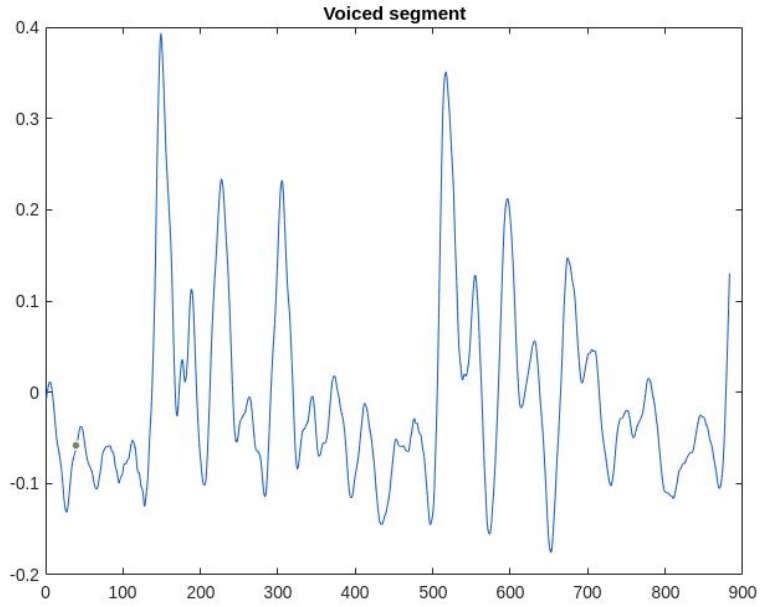
Above figure shows Pitch distribution for the selected Voiced frame. Average Pitch comes out to be nearly 120.32. As the variation is not much in this frame we consider our Fundamental frequency to be approximately 120 Hz in the selected voiced frame.

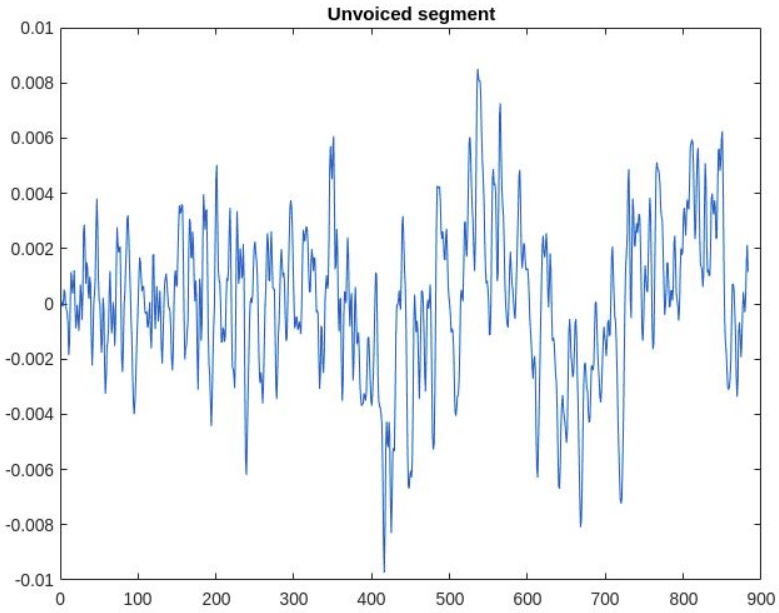
Part d

I have marked voiced and unvoiced regions through wavesurfer and I have attached the annotated transcriptions in the zip folder.

I selected two 20ms segments from the recorded audio (mentioned in matlab) and did the rest of the parts for those 2 segments. The sampling frequency (Fs) is 44100. For the voiced part I selected samples [3,00,000 to 3,00,882] which correspond to time [6.8s to 6.82s]. For the unvoiced part I selected samples [1,49,000 to 1,49,882] which correspond to time [3.378s to 3.398s].

`





Energy of the Voiced part is approximately 9.34 whereas energy of the unvoiced part is 0.0068. We see the energy of the Unvoiced part is very less as they are regions which are not intentionally spoken by a person. Thus these regions don't have energy. . We can use this criteria to distinguish between voiced and unvoiced regions but if the audio has noise, this method is not so efficient. Also some vowels or semi-vowels have low energy and thus deciding threshold is difficult.

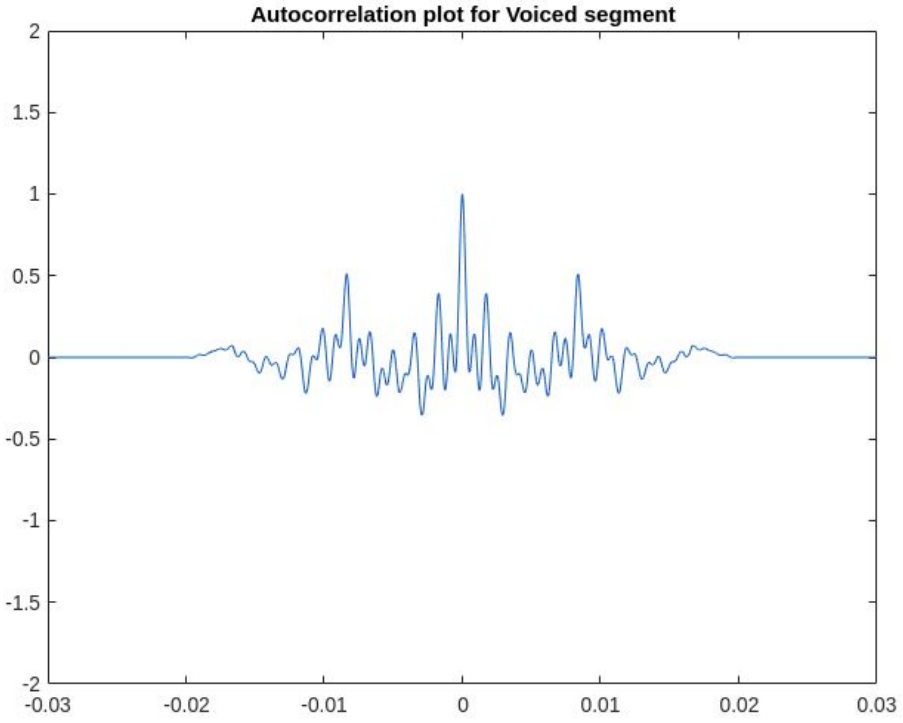
`

Part e

Zero Crossing of Voiced Part came out to be 23 and unvoiced part came out to be 118. So as expected due to random behaviour in the unvoiced part (non-periodic), the number of times signal crosses 0 is more than that in the voiced part (periodic). This is due to quasi-stationary behaviour of speech in voiced regions. We can use this criteria to distinguish between voiced and unvoiced regions but if the audio has noise, this method is not so efficient.

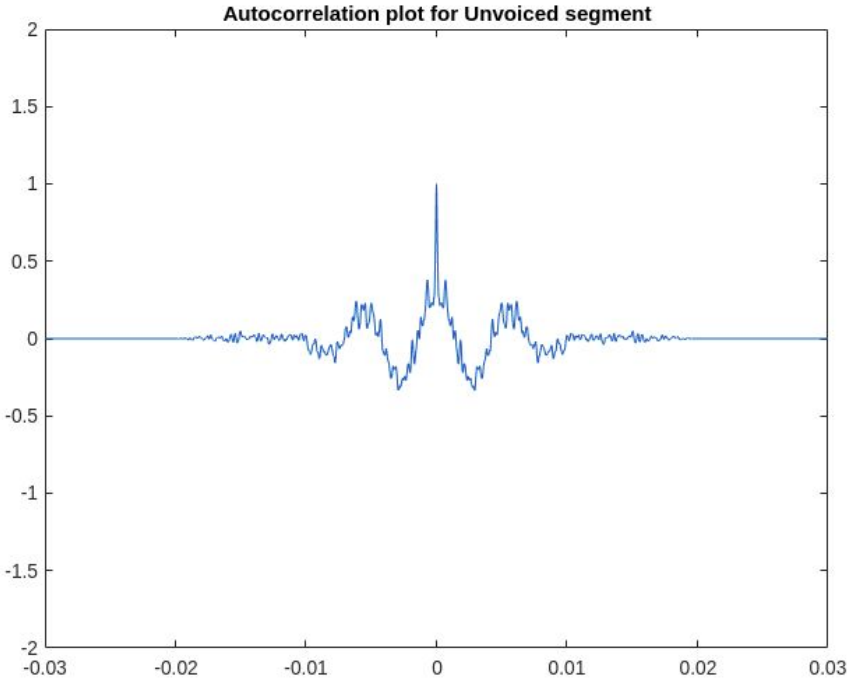
Part f

Autocorrelation is used to determine Pitch of audio signal. The reciprocal of distance between two regular (high-strength) peaks in an autocorrelation plot is Pitch Frequency.



In the above figure we can clearly see there are more than one significant peak and thus we can get some defined pitch in the voiced part. This happens due to periodicity or repetitive nature of signal in the voiced part.
We can see the difference between the central and significant peak at 0.008 is 0.008 and thus reciprocal of this gives the fundamental frequency that is approximately 120 Hz.

`



In the above figure we can see, there is only one significant peak showing non-periodic nature in an unvoiced region.
This approach can be used to separate voiced and unvoiced regions.