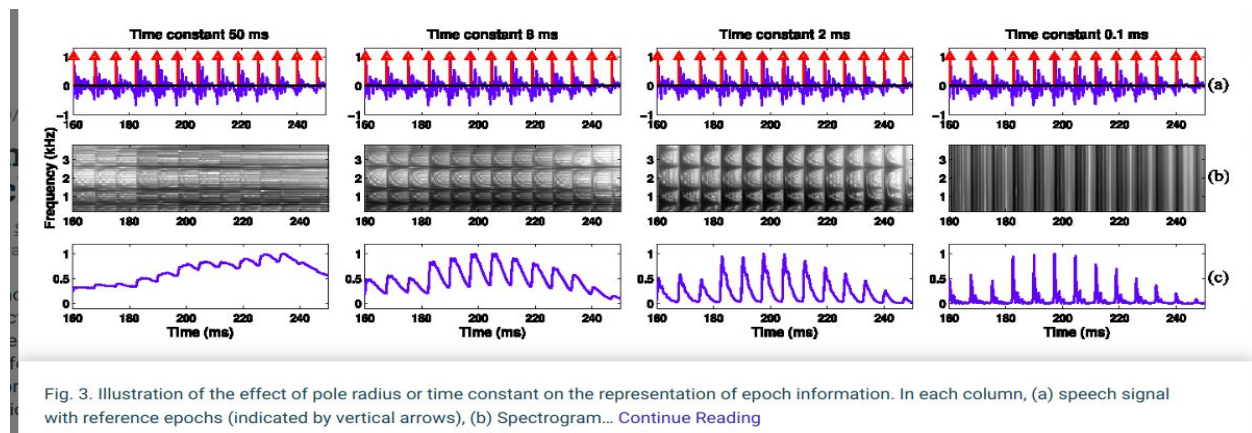


Speech Signal Processing

Assignment 2

Q.2 What are epochs in speech production? Why are they significant in speech signal processing? Illustrate with an example.

Definition of Epochs: Epoch is the instant of significant excitation of the vocal-tract system during production of speech. For most voiced speech, the most significant excitation takes place around the instant of glottal closure. In the following diagram, each part signifies an epoch.



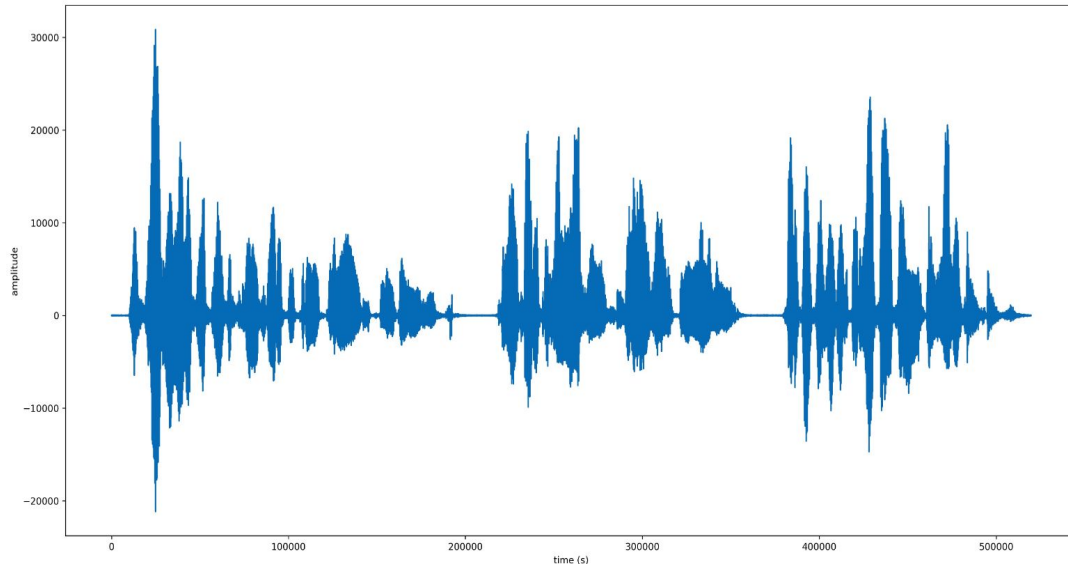
Usage of Epochs : Voiced speech analysis consists of determining the frequency response of the vocal-tract system and the glottal pulses representing the excitation source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal-tract system is within a glottal pulse. For example, knowledge of the epoch locations is useful for accurate estimation of the fundamental frequency.

- Often the glottal airflow is zero soon after the glottal closure. As a result the supralaryngeal vocal-tract is acoustically decoupled from the trachea. Hence, the speech signal in the closed phase region represents the free resonances of the supralaryngeal vocal-tract system. Analysis of the speech signal in the closed phase regions provides an accurate estimate of the frequency response of the supralaryngeal vocal-tract system . With the knowledge of the epochs, it is possible to determine the characteristics of the voice source by a careful analysis of the signal within a glottal pulse.
- The epochs can be used as pitch markers for prosody manipulation, which is useful in applications like text-to-speech synthesis, voice conversion and speech rate conversion.
- Knowledge of the epoch locations may be used for estimating the time-delay between speech signals collected over a pair of spatially distributed microphones .
- The segmental signal-to-noise ratio (SNR) of the speech signal is high in the regions around epochs, and hence, it is possible to enhance the speech by exploiting the characteristics of speech signals around the epochs . It has been shown that the excitation features derived from the regions around the epoch locations provide complementary speaker-specific information to the existing spectral features.

Q.2 Load the files H MKB.wav into MATLAB/Python using the function wavread/audioread. Audio file is shared along with the assignment.

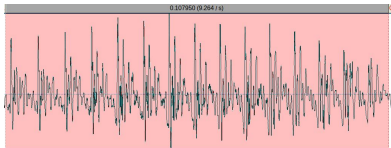
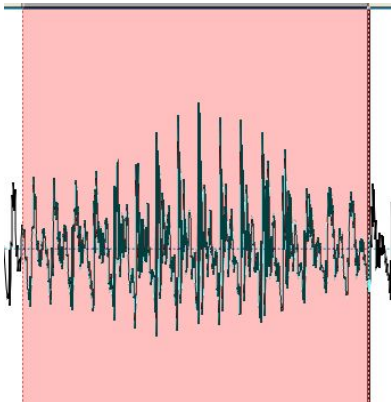
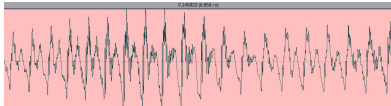
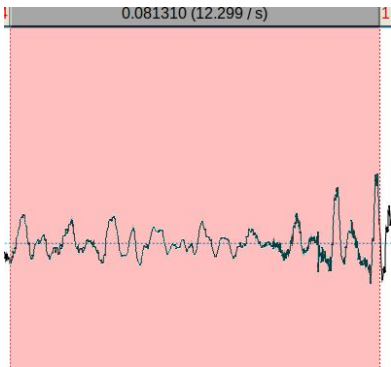
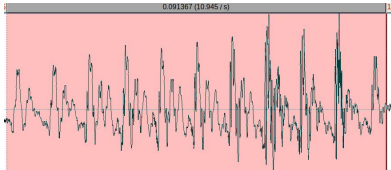
a) Time Plot Axis :

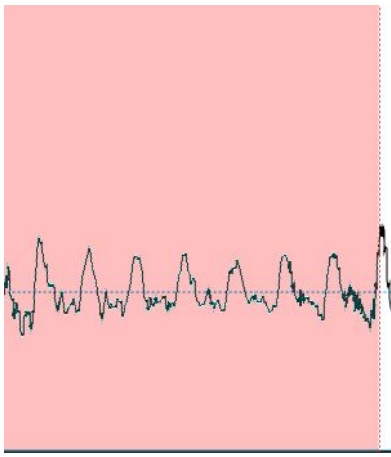
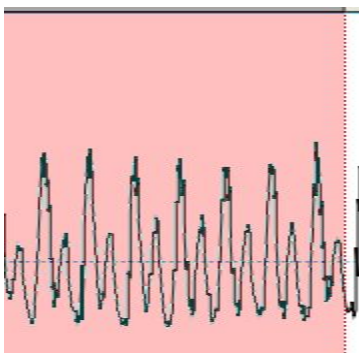
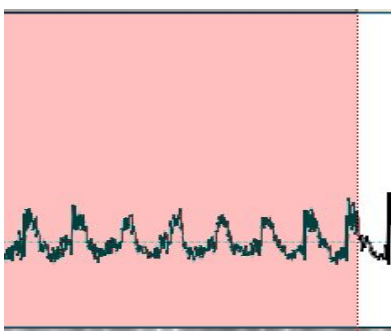
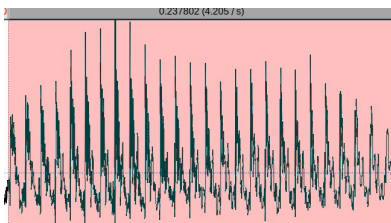
File Name : H_MKB
 Number of Channels : 1
 Sample Rate : 44100 Hz
 Time Duration: 11.784126984126985 secs


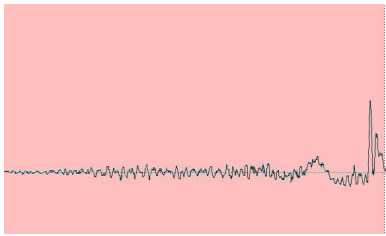
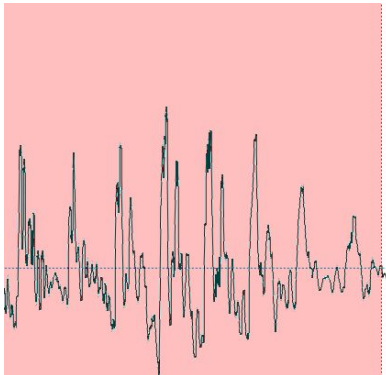
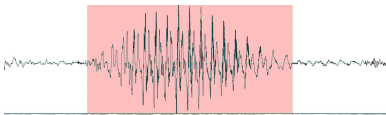
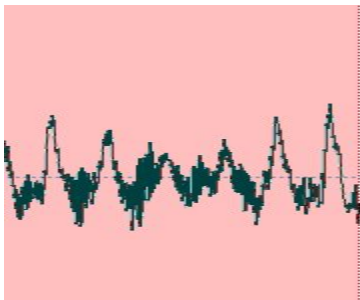


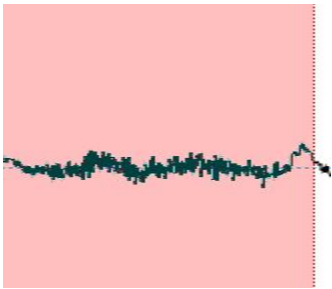
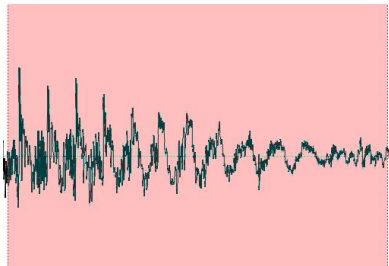
Voiced and Unvoiced Parts : (Major ones)

Time	Wave Segment	Unvoiced/Voiced/Silence/Plosive
t= 0.2412 to t= 0.334		Voiced Region (Quasi Stationary Wave) /I/ {Pronounced as ee}
t= 0.45 sec to t= 0.51 sec		Unvoiced (Bh)

t= 0.53 sec to t= 0.631 sec		Voiced (Vowel) /aa/ in /Bhar/
t= 0.79 to t= 0.83		/m/ in /maine/ Voiced
t=0.69 sec to t= 0.79		Voiced (/i/ in /bhi/)
t=1.029 sec to t= 1.11 sec		/dh/ in /dheka/ Unvoiced (partly silence included too)
t=1.10 sec to t= 1.20 sec		/ee/ Voiced Part

<p>t= 1.40 sec to t= 1.49 sec</p>	 <p>A waveform plot showing a series of irregular, low-amplitude pulses. The background is light red, and the waveform is black. The pulses are distributed across the time interval, with no sustained high-frequency oscillations.</p>	<p>/k/ - Voiceless velar Unvoiced</p>
<p>t=5.75 to t= 6.0</p>	 <p>A waveform plot showing a series of regular, high-amplitude pulses. The background is light red, and the waveform is black. The pulses are closely spaced, indicating a high frequency of vibration.</p>	<p>/n-i/ Voiced (ee part of it is more focussed in this image)</p>
<p>t= 6.02 to t=6.09 secs</p>	 <p>A waveform plot showing a series of irregular, low-amplitude pulses. The background is light red, and the waveform is black. The pulses are distributed across the time interval, with no sustained high-frequency oscillations.</p>	<p>/H/ in /hai / Unvoiced</p>
<p>t= 6.09 to t= 6.331 sec</p>	 <p>A waveform plot showing a series of regular, high-amplitude pulses. The background is light red, and the waveform is black. The pulses are closely spaced, indicating a high frequency of vibration. A small text label '0.237802 (4.205 / s)' is visible at the top of the plot area.</p>	<p>/ai/ (Mixed - elongated due to pronunciation) Voiced</p>

t= 6.32 to t= 6.55 sec		/s/ from /sujav/ Unvoiced part .
t=8.61 sec to t=8.71 sec		/h/ from /har/ Unvoiced.
t=8.73 sec to t = 8.79 sec		/r/ in /har/ Unvoiced {Looks like voiced , but it isn't really}
t=8.84 sec to t = 8.93 sec		/oi/ in /koi/ Voiced .
t= 9.11 to t = 9.17 sec		/ch/ in /kuch/ Unvoiced


t=9.8 to t= 9.84		/ch/ in /chahtha/ Unvoiced
t=11.00 to t=11.15 sec		/t/ in /tha Unvoiced.

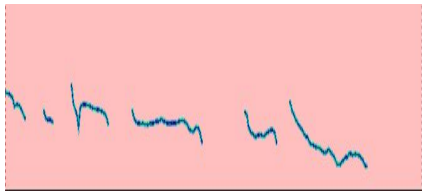

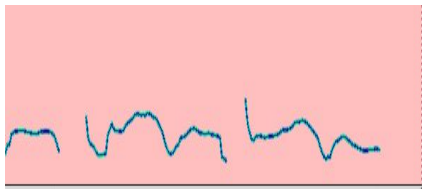

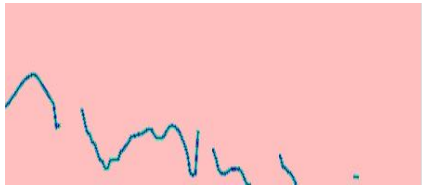
Note : The table does not consist of all the voiced and unvoiced parts of the sample , I've chosen a few interesting areas and mentioned the details.

- b) In the time-domain plot, mark the regions where the pitch is the highest and the lowest. What are the pitch frequencies in those regions?

So, to understand where the pitch is highest/lowest , I've used the Pitch option in Praat Tool, it provides the pitch statistics.

Note : I've splitted the entire signal into 2 sec length duration to analyse the pitch values . This ensures that there is no variance in pitch across samples.

Time Frame	Highest Pitch	Lowest Pitch	Mean Pitch	Pitch Figure
t= 0 to t = 2 secs	165.561165754 33063 Hz	81.247272455 76675 Hz	123.797559681 73984 Hz	

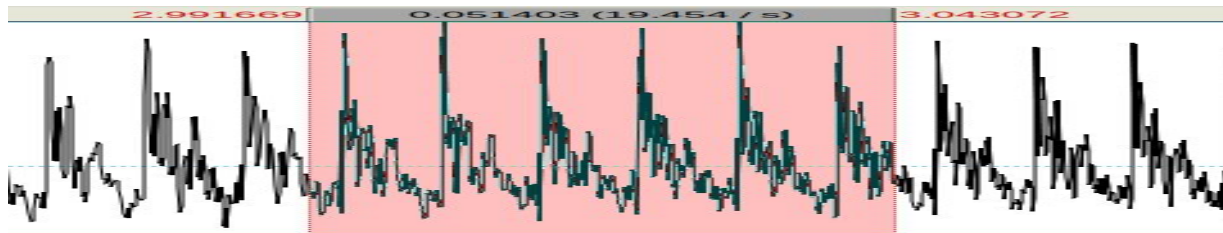
t= 2 to t = 4 secs	137.73572873 046666 Hz	89.579516066 32139 Hz	112.249500491 5609 Hz	
t=4 to t=6 secs	143.11944649 234124 Hz	84.876283849 4735 Hz	114.449279138 20985 Hz	
t=6 to 8 secs	129.43241070 783324 Hz	91.1573258767 1098 Hz	106.45854707 04419 Hz	
t= 8 to t=10sec	161.199659579 48273 Hz	106.94689377 735797 Hz	126.184196372 41051 Hz	
t= 10 to t=11 secs	156.906162907 06367 Hz	78.933741933 3375 Hz	113.175224868 58398 Hz	

Note : In the rightmost column , one can easily observe/mark the maximum & minimum pitch .

c) Fundamental Frequency in my case :

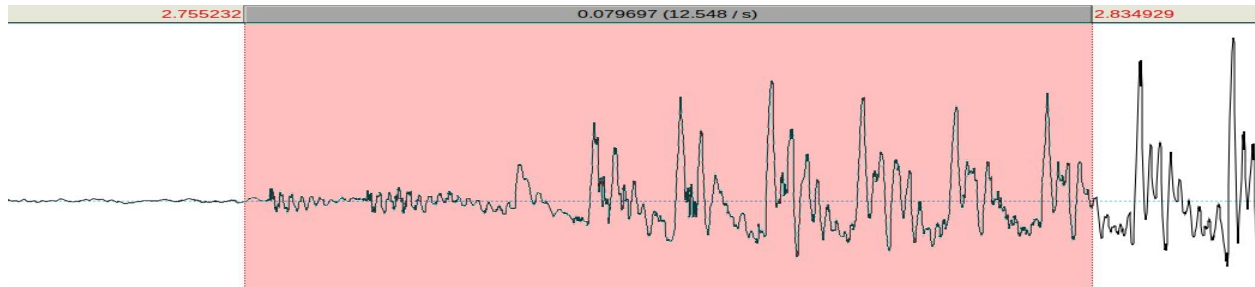
So, in my case I will be choosing a 50 ms frame from t=2.99 secs to t =3.04 secs , it's a part of /comments/ word , specifically /m/ .

Here, its voiced structure ,and we can observe the quasi-periodic nature of the speech segment.

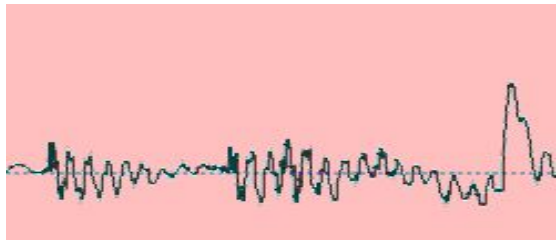
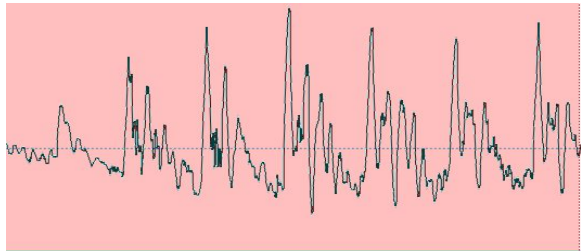


According to Praat tool's pitch option ,the fundamental frequency turns out to be : **115.15 Hz**

d)Let's take a 80ms frame ,that consists of both unvoiced & voiced components . So, this part was extracted from /comments/ initial part , and : 'c' part was unvoiced and 'm' was voiced region .



Selected frame of the speech signal
t= 2.755 sec to t = 2.83 sec

Parts of Frame	Image
Unvoiced Region t=2.75 sec to t=2.78 sec	
Voiced Region t=2.78 to t = 2.83 sec	

Q. 1. (d , e ,f)

Identify a voiced regions and unvoiced regions in a signal and for one particular

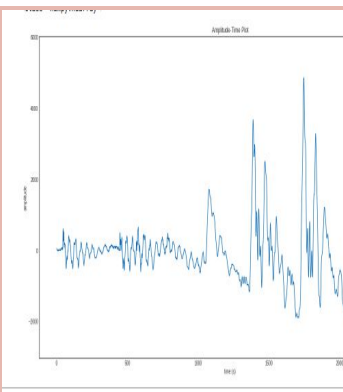
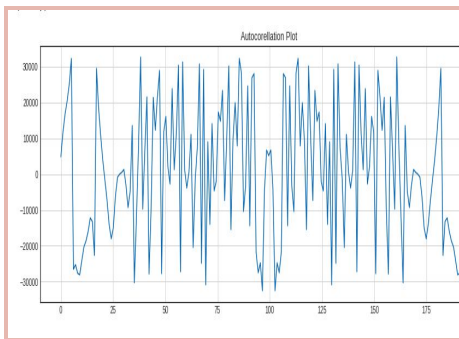
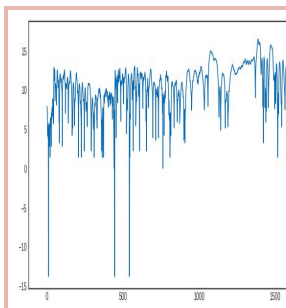
frame, compute of frame energy and comment of it. (2 pts)

(e) Implement zero-crossings and comment of it.(2 pts)

(f) Autocorrelation and comment of it.(2 pts)

Note : We need to understand that the frame consists of both voiced and unvoiced regions , but there is dominance of the voiced region (refer the picture) . Hence , we are bound to observe more voiced region features.

Observation Table :

Frame	Autocorrelation	Log Short Term Energy	ZCR
-Contains voiced & unvoiced components. 	Graph 	Total Energy : 139.40 dB 	Zero Crossing Rate : 0.041207670338637 294

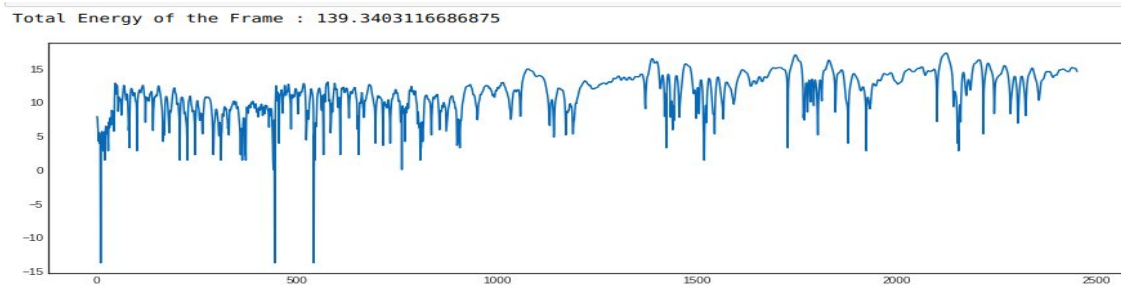
- Comments for each of the above category :

- ZCR** : The zero crossing rate here is less than 0.5 , it shows dominating voiced signals characteristics .We also know that Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count , whereas the unvoiced

speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count.

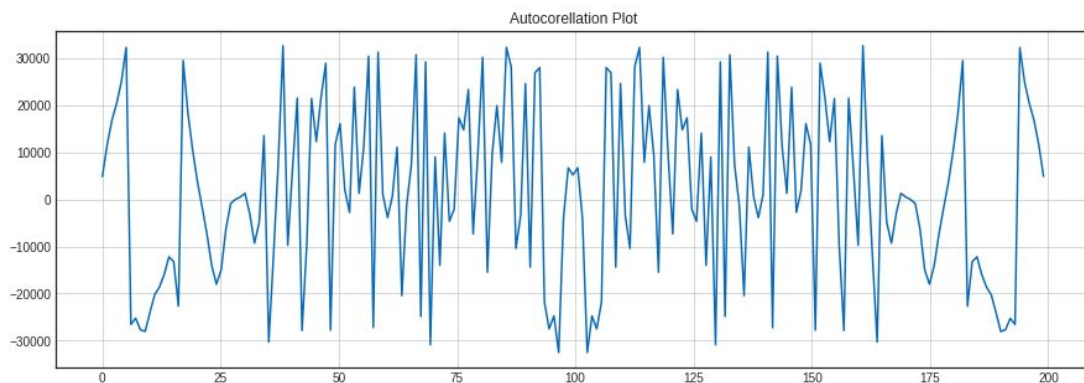
2. Short Term Energy :

So, the frame's energy is around 139.34 db .



3. Autocorrelation :

Since , the frame is dominated by the voiced part , the autocorrelation shows highly correlated signal parts. And at the center of the graph , we can observe symmetrical structure . An important note is that



Due to the concentration of low frequency energy of voiced sounds, adjacent samples of voiced speech waveform are highly correlated and thus the coefficient parameter is close to 1. On the other hand, the correlation is close to zero for unvoiced speech. So, in this case the coefficients will be closer to 1. Also, pitch calculation can be done via consecutive formants .

Q.1 Record your name and the utterance should be "I am <your name>".

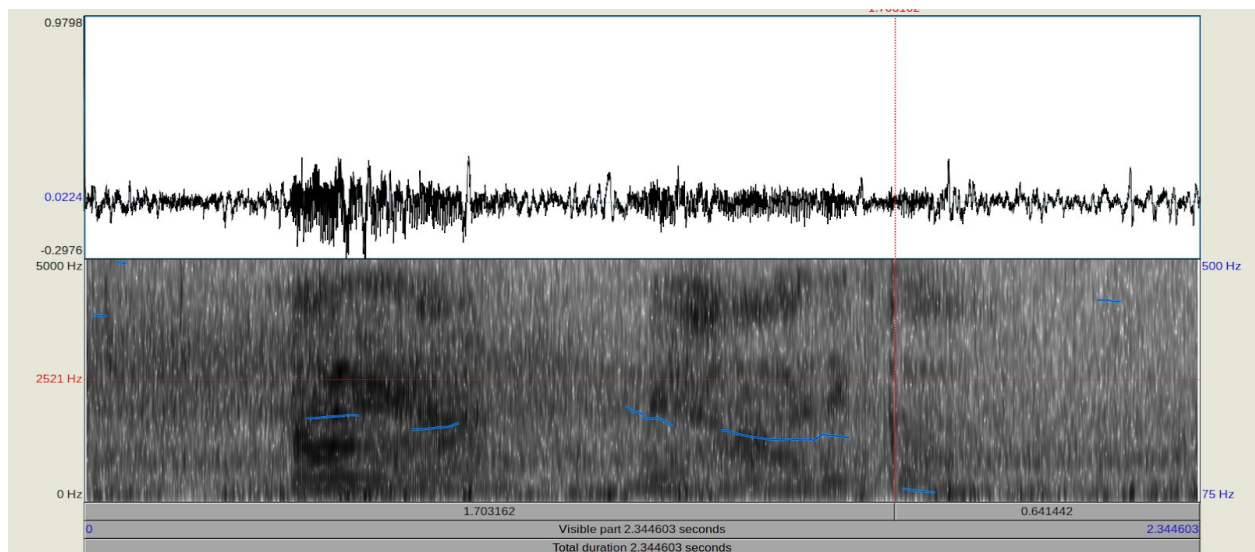
A: The corresponding speech voice sample is added to the corresponding question's directory .

The sample contains the following line : {Added first name}

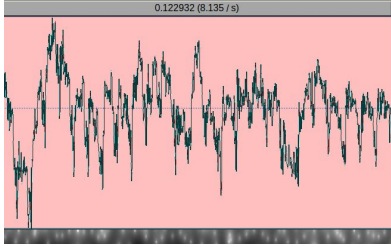
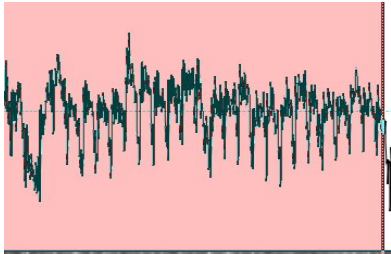
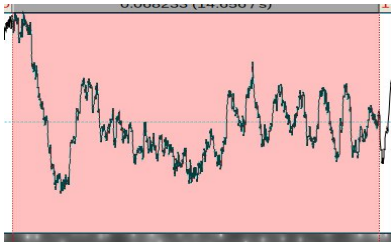
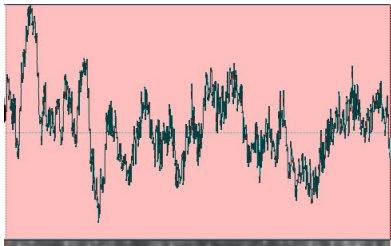
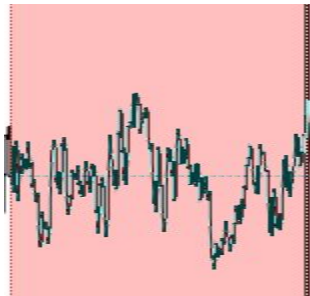
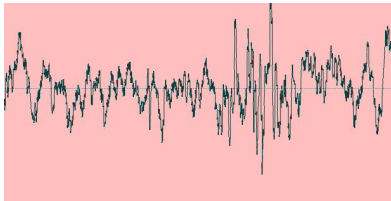
< I AM Niharika >

Note : The plot seems to have a lot of background noise which crept in during recording .

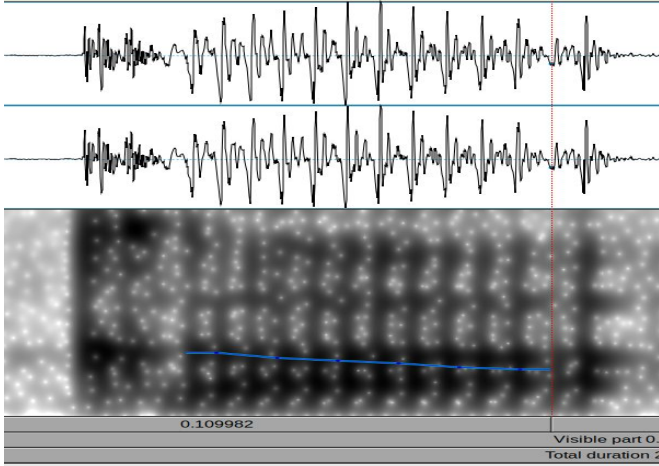
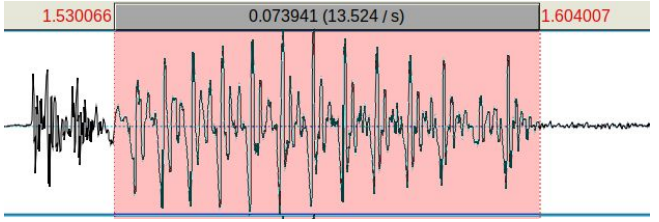
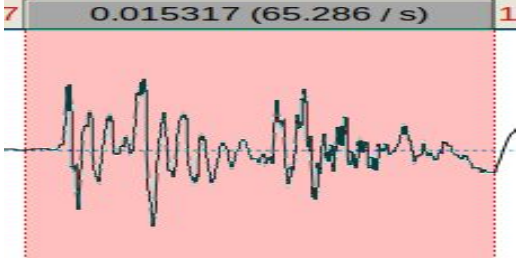
a) Time Domain plot :



Time Period	SpeechSegment	Category
t=0 to t=0.48 sec		Voiced //

t= 0.58 to t= 0.75 sec		Voiced /a/
t=0.663 sec to t=0.79 sec		/m/ Voiced
t=1.102 sec to t=1.17 secs		/n/ Unvoiced Nasal Constriction
t=1.24 sec to t =1.334 sec		/h/ Unvoiced
t=1.41 sec to t=1.44 sec		/r/ in /Niharika/ Unvoiced
t=1.65 sec to t=1.71 sec		/k / in /Niharika/ Unvoiced {No formant seen for spectrogram}

b) Now, we can obtain a 15ms frame , to observe the following properties . To understand the variation in values of voiced & unvoiced regions , I have split the wave into its respective types and performed the analysis. Also, analysis is done to a cleaner sample stored as niharika_frame.wav in the Zip Folder. The voiced segment is stored as - a_frame.wav and the unvoiced segment as k_frame.wav

Time Bounds	Speech Segment	Category
t=1.50 sec to t =1.60 secs		Voiced + Unvoiced Part (/ka/) /k/ - Unvoiced /a/ - Voiced
t=1.53 sec to t= 1.60 sec		/a/ - Voiced
t=1.51 to t=1.53 secs		/k/ - Unvoiced Segment In /ka/ of /niharika/ word,

i) Zero Crossing Rate :

A zero crossing is said to occur if there is a sign difference between successive samples. The rate at which zero crossings happen is a simple measure of the frequency content of a signal.

$$Z_m = \sum_n | \text{sign}[x(n)] - \text{sign}[x(n-1)] | w(m-n)$$

$$\text{where the sign function is } \text{sign}[x(m)] = \begin{cases} 1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases}$$

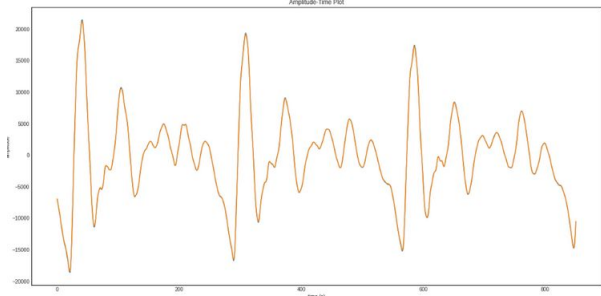
So , the Python Code written for this rate is :

```
# Zero Crossing Rate.

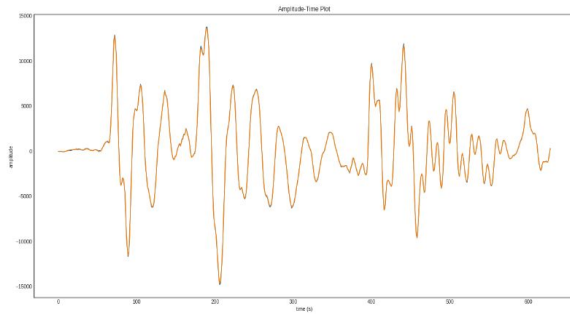
def sign(x):
    if(x>=0):
        return(1)
    return(-1)

def zeroCrossingRate(fileName):
    Fs,data=readFile(fileName,False)
    if(data.shape[1]==2):
        data=data[:,0]
    signs = np.sign(data)
    signs[signs == 0] = -1
    return len(np.where(np.diff(signs))[0])/len(data)
```

Observations :

Speech Segment	Zero Crossing Rate
<p>1.Voiced Part</p> 	<p>0.0350</p>

2.Unvoiced Part



0.07154

Comments :

The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count, whereas the unvoiced speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count. Hence, we can confirm from above experiment that ZCR of the unvoiced is 0.075 compared to Voiced part's ZCR.

ii) Log Short Term Energy :

The energy E of a discrete time signal $x(n)$, esp for many audio signals such a measurement is of less importance, since it gives little information about time dependent characteristics of such signals.

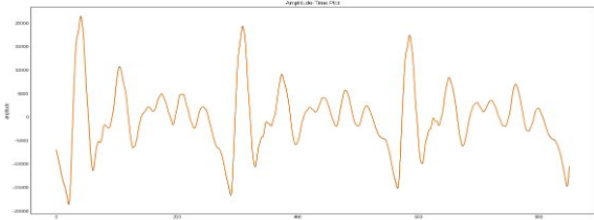
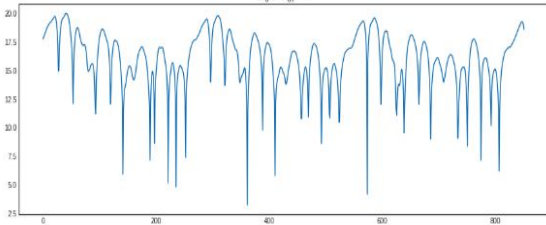
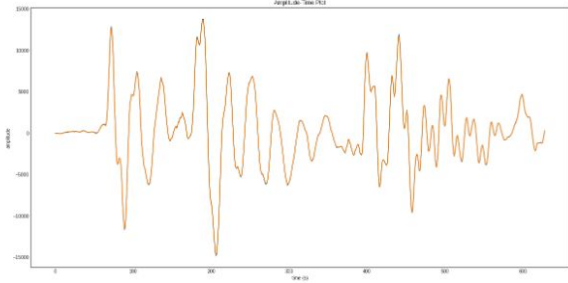
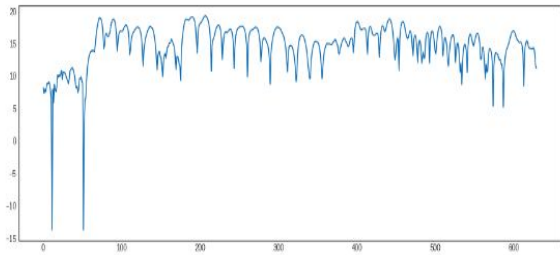
$$E_s = 10 \log \left(\varepsilon + \frac{1}{N} \sum_{n=1}^N S^2(n) \right)$$

Python Code Snippet :

```
def logEnergy(fileName,flag=False):
    Fs,data=readFile(fileName,flag=False)
    ep=0.0001
    if(data.shape[1]==2):
        data=dt=data[:,0]
    #Compute FFT .
    y=abs(data)
    N=len(y)
    y_conj= np.conj(y)
    tot_energy = abs(np.dot(y,y_conj))
    en= [y[i]**2 for i in range(len(y))]
    logen=[log(y[i]**2) for i in range(len(y))]
    totalEnergy = 10*log(ep+(1/N)*sum(en))
    if(flag):
        t=np.linspace(0,N,N)
        fig=plt.figure(figsize=[15,5])
        plt.plot(t,logen)
        plt.title('Log Energy Plot')
    print('Total Energy of the Frame : ',totalEnergy)
```

So, the above equation we normalise it and compute $10 \cdot \log(E)$, where E is the above energy term. Here, the log energy graph is on the right side .

Observation Table :

Signal Segment	Log Short Term Energy
1.Voiced Region 	Total Energy : 176.48 dB 
2.Unvoiced Region 	Total Energy : 167.50 dB 

Comments :

Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis and usually shows a low zero crossing count. The energy of unvoiced data is usually lower than for voiced sounds but higher than for silence. The above results confirm the same, the energy of the voiced is more than the unvoiced part.

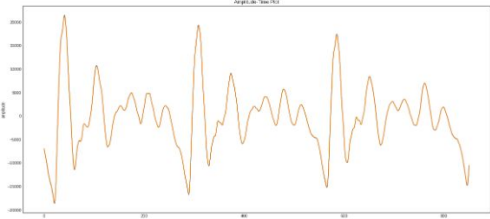
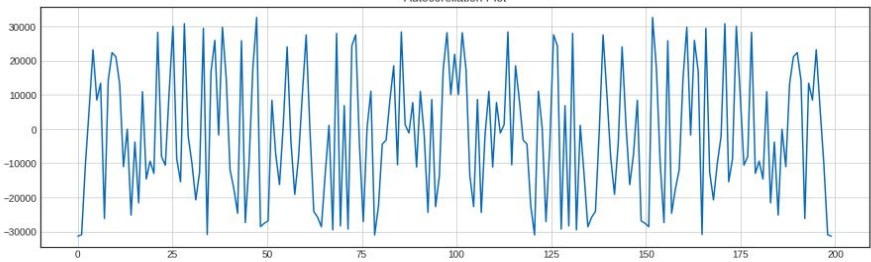
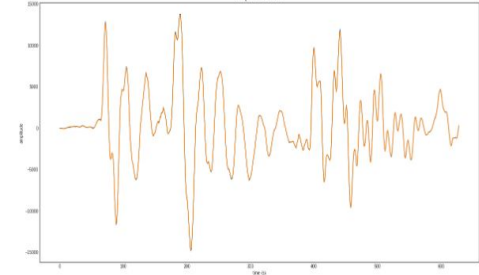
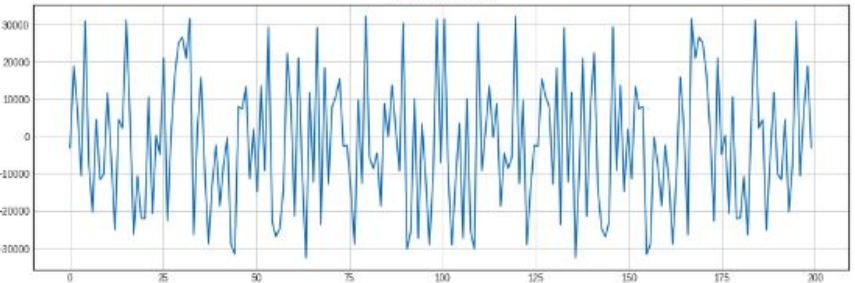
iii) Autocorrelation :

Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. So, for autocorrelation to be clearly visible, I've taken a smaller set of 100 data points for the voiced sample.

Python Code Snippet :

```
# Autocorrelation
import matplotlib.pyplot as plt
def autoCorretion(fileName, flag=False):
    Fs, data = readFile(fileName, False)
    # Apply autocorr to the numpy data.
    data = np.asarray(data)
    print(data.shape)
    dt = data[:, 0]
    # Note : Here only 100 samples are taken out for graphical purpose
    dt = dt[0:100]
    corr = scipy.signal.correlate(dt, dt, mode='full', method='auto')
    if(flag):
        N = len(corr)
        t = np.linspace(0, N, N)
        fig = plt.figure(figsize=[15, 5])
        plt.plot(t, corr)
        plt.grid()
        plt.title('Autocorrelation Plot')
```

Observation Table :

Speech Segment	Autocorrelation Graph
1.Voiced Segment 	
2.Unvoiced Segment 	

Comments :

If you observe the autocorrelation graph of the voiced region , the center of the graph contains 2 peaks and the entire graph is symmetric in nature . If we compute the highest peak and the next-to-next peak's value , the pitch of the segment can be calculated . In case of unvoiced signals , it's simply an erratic signal with no symmetry / pattern . Also,due to the concentration of low frequency energy of voiced sounds, adjacent samples of voiced speech waveform are highly correlated and thus the coefficient parameter is close to 1. On the other hand, the correlation is close to zero for unvoiced speech.

(CONTD)

Conclusion :

Hence , out of all 3 traditional methods for classification , according to me the best method that worked is observing the graph of Autocorrelation for the frame . If there are periodic structures , symmetry is clearly visible and formant calculation is easier too. If you look at ZCR or Log Based Energy methods , they have dynamic parameters (e.g. threshold per window ,etc) , and there can be misclassification if there is excess noise in the speech frame , hence it will be difficult then . In case of autocorrelation procedure , the noise in the sample will clearly be separated .