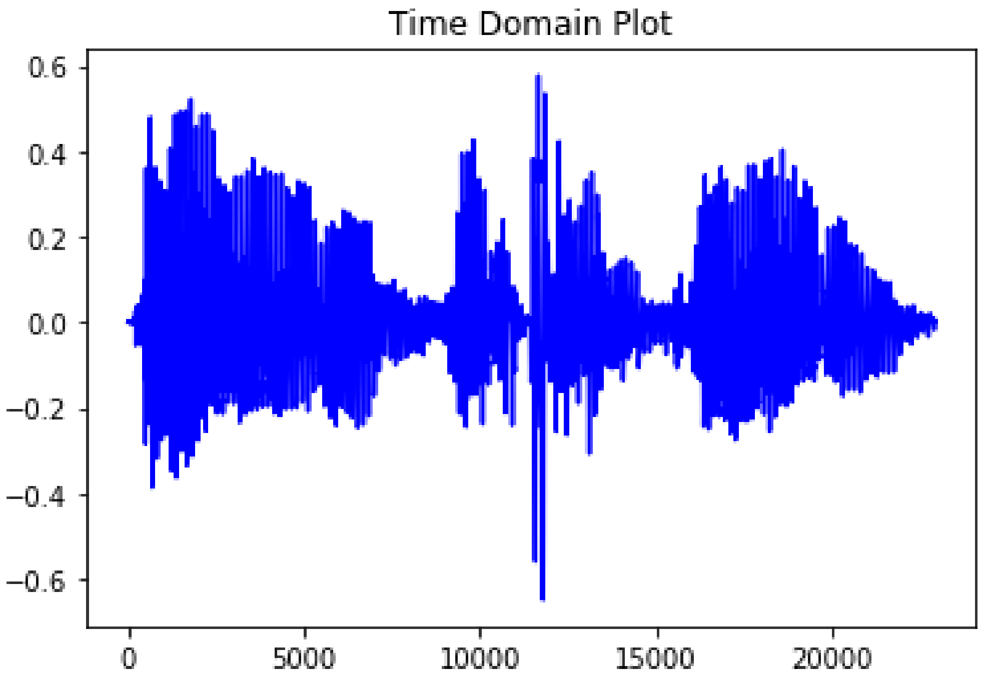
Abhigyan Ghosh - 20171089

Speech Signal Processing [ECE448]

20 August 2020

Assignment 2

1) Question 1



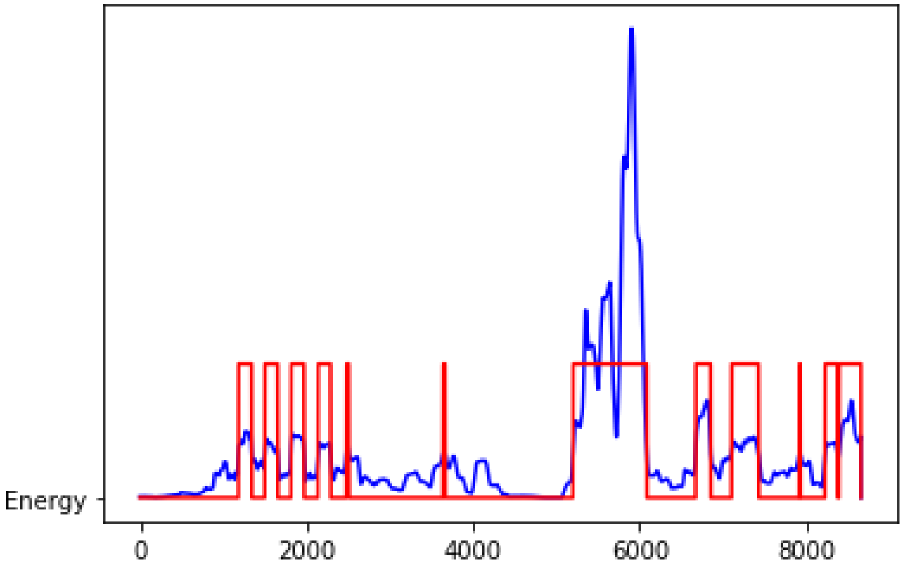
a) The entire thing is voiced because my entire name is voiced.

I am Abhigyan /ai/,/ae//m/, /a/,/bh/,/i/,/g/,/y/,/A/,/n/

b) The entire speech is voiced so I guess the answer is yes to all the below

i) There are 149 `zero crossings.`

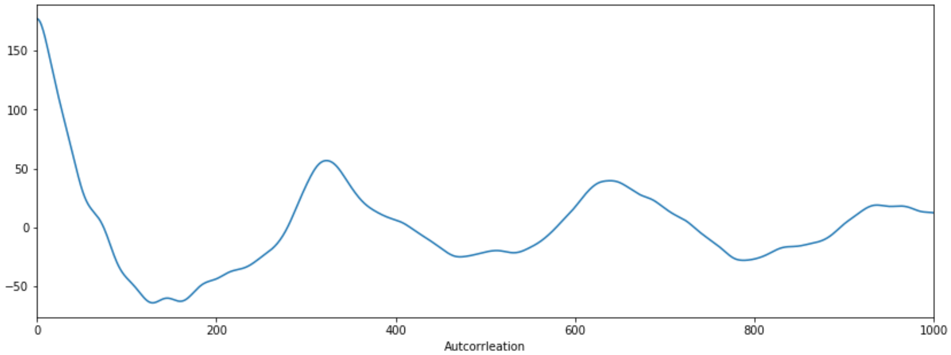
ii)



The areas in red are the spaces which have higher than average in the frame.

But the entire signal has high energy.

iii)



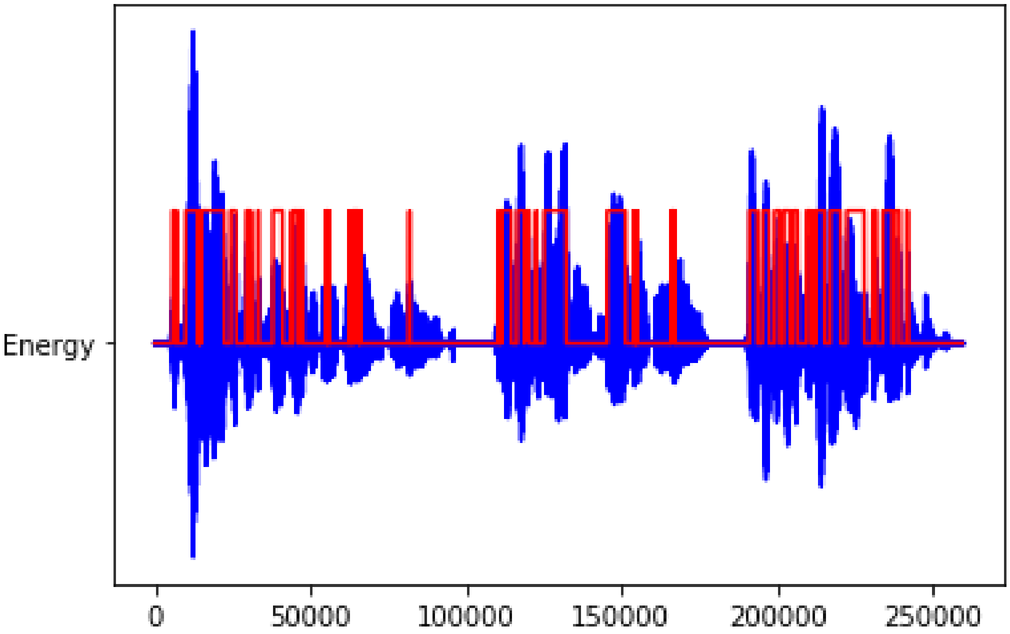
2) Question 2

Epoch is the instant of significant excitation of the vocal-tract system during production of speech.

It enables extraction of important acoustic-phonetic features such as glottal vibrations, formants, instantaneous fundamental frequency, etc. Epoch sequence is useful to manipulate prosody in speech synthesis applications. Accurate estimation of epochs helps in characterizing voice quality features.[1]

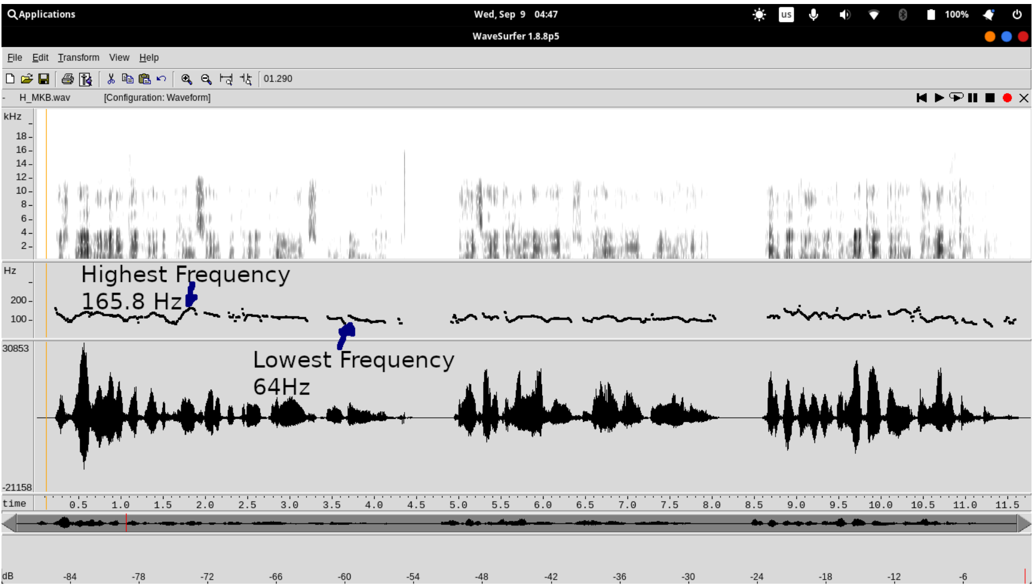
3) Question 3

[1] https://link.springer.com/article/10.1007/s12046-011-0046-0



a)

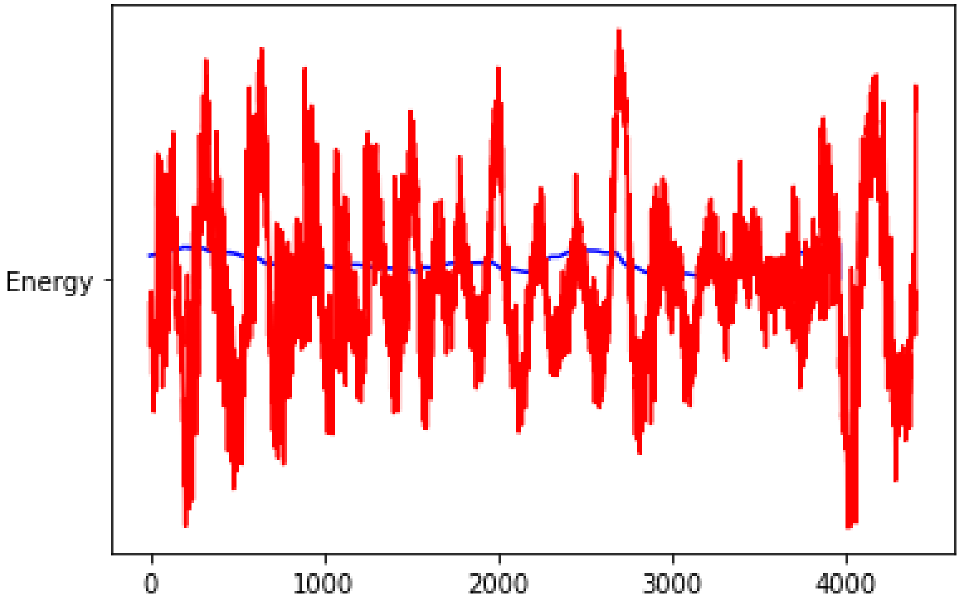
The places where the red line is high is voiced and rest is unvoiced or silent. The signal was too long to manually mark.



b) `

The highest pitch is 165.8 Hz and lowest pitch is at 64Hz

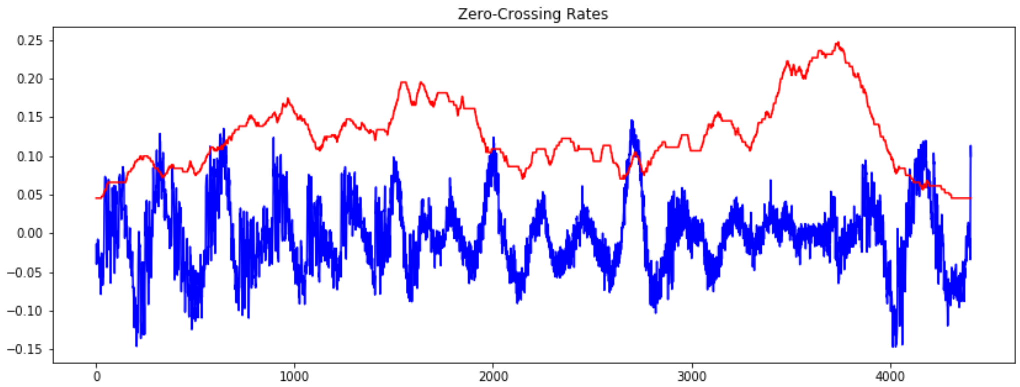
c) 117.647 Hz is the fundamental frequency (pitch).



d) 0.354

Hz Is the total energy of the frame. The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific, energy associated with short term region of speech. By the nature of production, the speech signal consists of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy. Thus, short term energy can be used for voiced, unvoiced and silence classification of speech. The relation for finding the short-term energy can be derived from the total energy relation defined in signal processing. The total energy of an energy signal is given by

$$E_T = \sum_{n=-\infty}^{\infty} s^2(\mathrm{n})$$



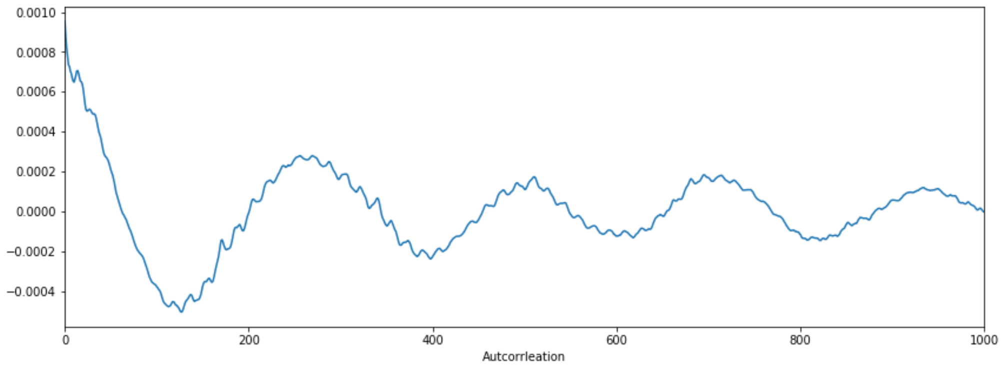
e)

Zero crossing rates are higher in this region as it is a voiced region. Zero Crossing Rate gives information about the number of zero-crossings present in each signal. Intuitively, if the number of zero crossings are more in each signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information. On the similar lines, if the number of zero crossing are less, hence the signal is changing slowly and accordingly the signal may contain low frequency information. Thus, ZCR gives an indirect information about the frequency content of the signal. The ZCR in case of stationary signal is defined as,

$$z = \sum_{n=-\infty}^{\infty} |sgn(s\ (n)) - sgn(s\ (n-1))|$$

$$\text{where } sgn(s\ (n)) = 1 \text{ if } s\ (n) \geq 0$$

$$= -1 \text{ if } s(\ n) < 0$$



f)

Cross correlation tool from signal processing can be used for finding the similarity among the two sequences and refers to the case of having two different sequences for correlation. Autocorrelation refers to the case of having only one sequence for correlation. In autocorrelation, the interest is in observing how similar the signal characteristics with respect to time. This is achieved by providing different time lag for the sequence and computing with the given sequence as reference. The autocorrelation is a very useful tool in case of speech processing. However due to the non-stationary nature of speech, a short-term version of the autocorrelation is needed. The autocorrelation of a stationary sequence rxx(k) is given by,

$$r_{xx}(k) = \sum_{m=-\infty}^{\infty} x(m).x(m+k)$$