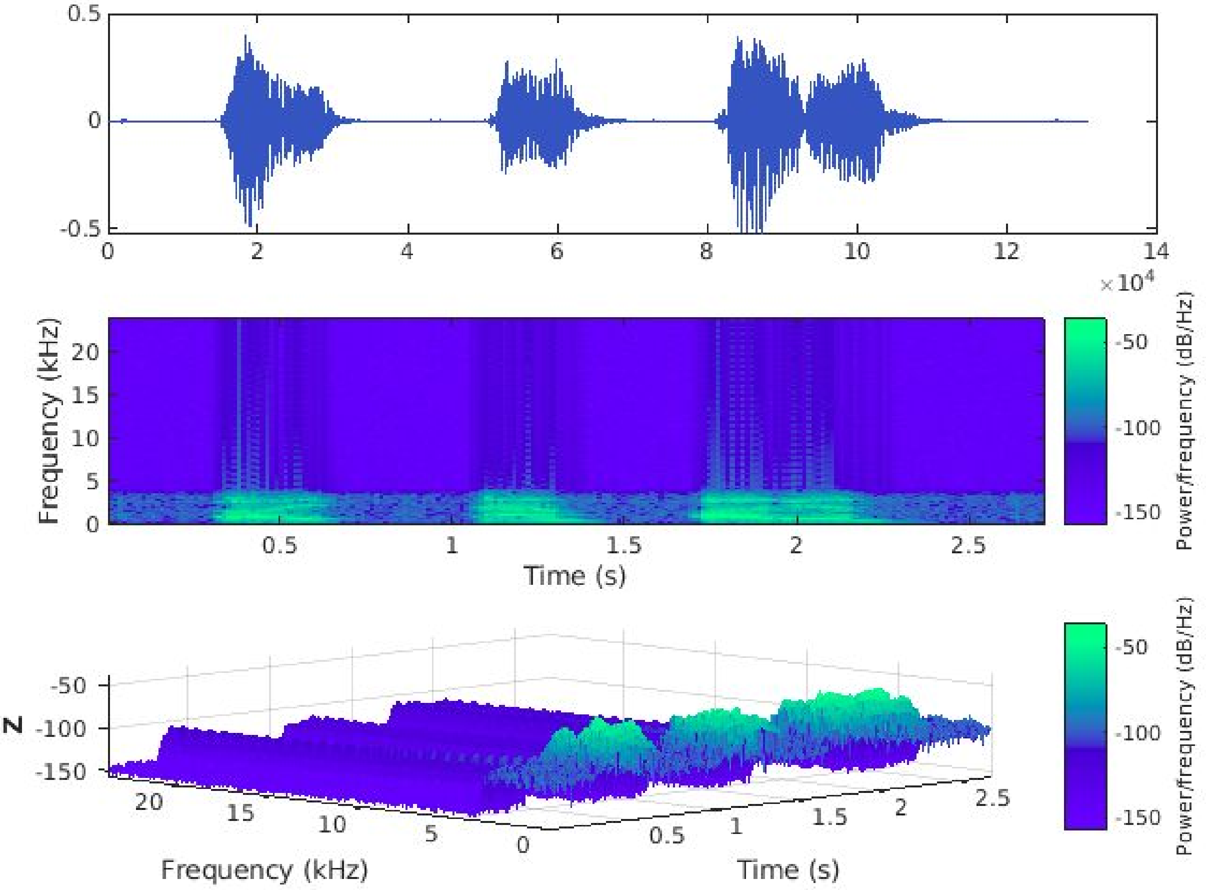
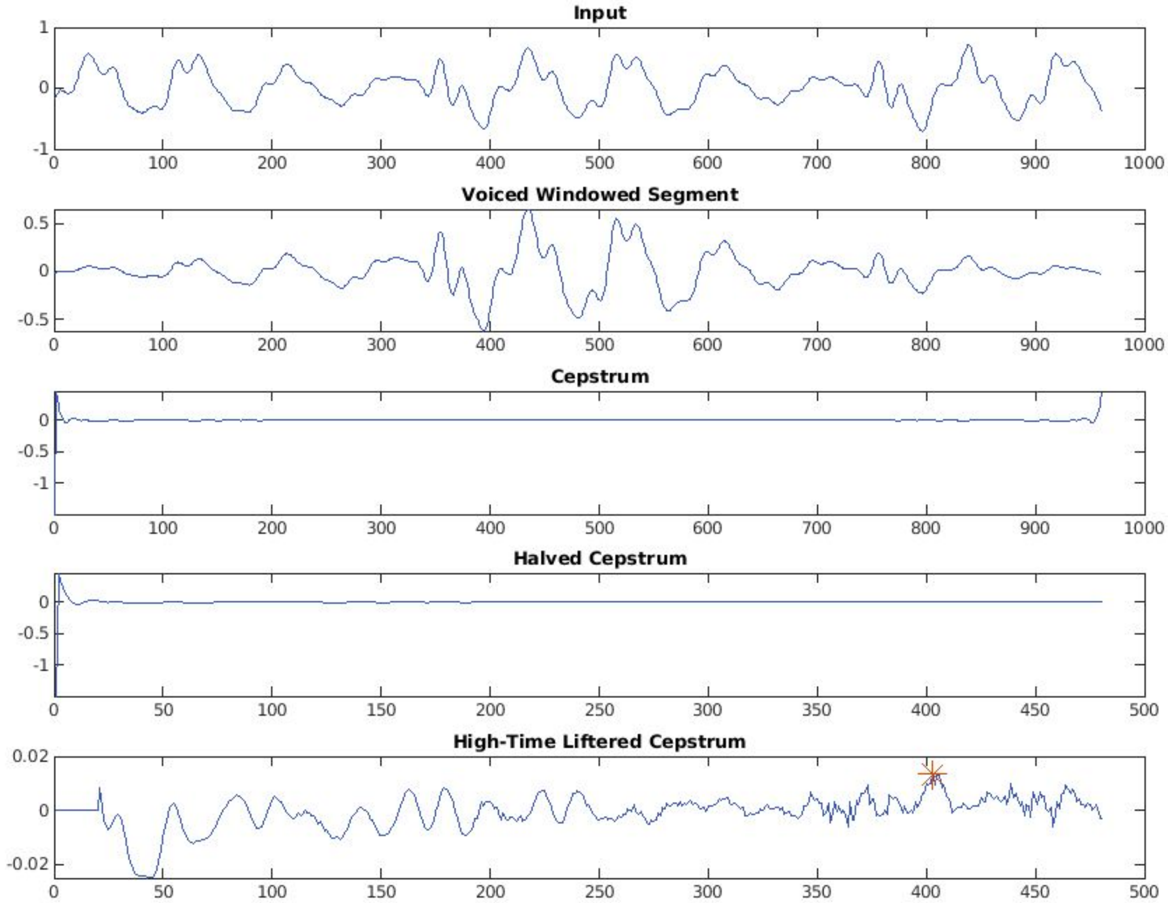
Question 1


0.5
0
-0.5
0
2
4
6
8
10
12
14
×10^4
Frequency (kHz)
20
15
10
5
0
0.5
1
1.5
2
2.5
Time (s)
-50
-100
-150
Power/frequency (dB/Hz)
Z
-50
-100
-150
20
15
10
5
0
Frequency (kHz)
0.5
1
1.5
2
2.5
Time (s)
-50
-100
-150
Power/frequency (dB/Hz)

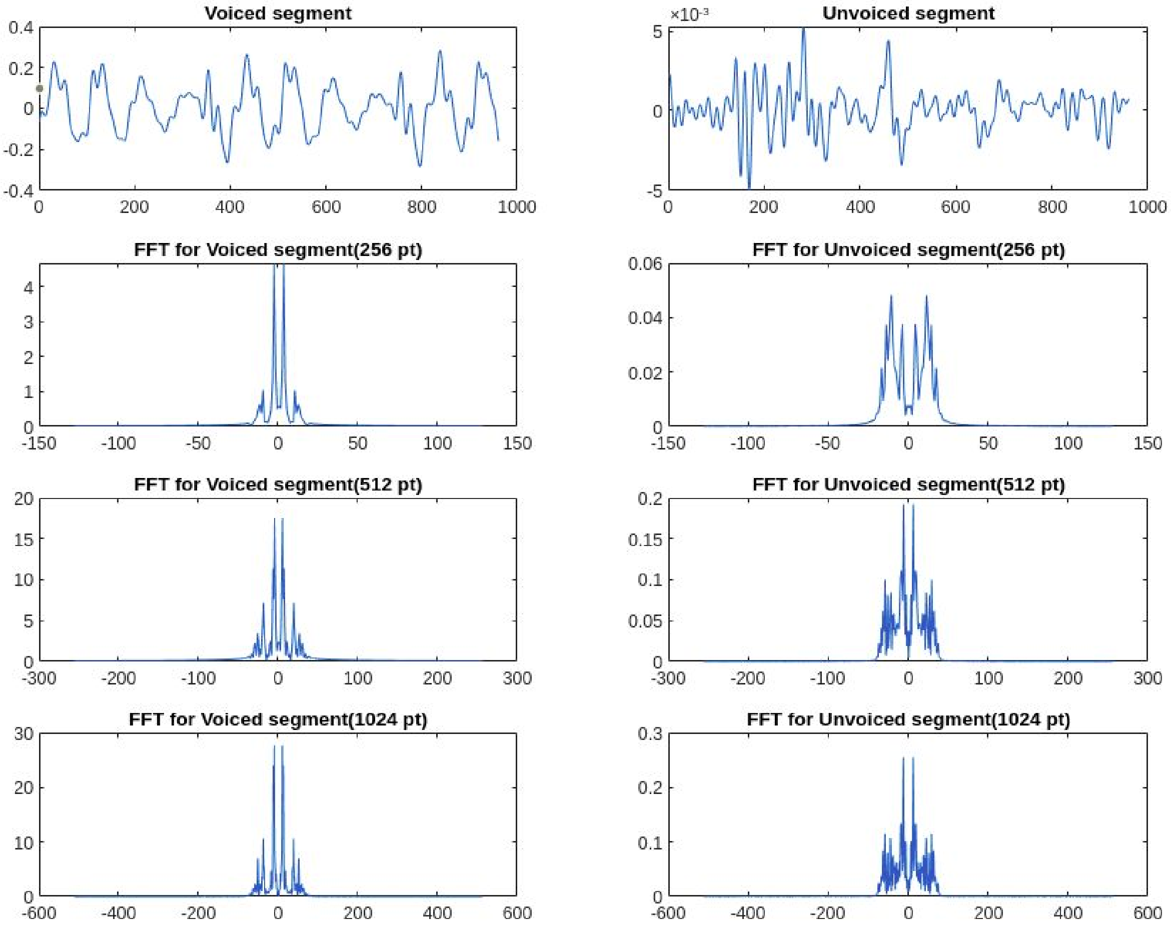
Question 2



The first plot is a 20ms voiced segment. The second plot indicates a windowed voiced segment. The windowing is done using the Hamming window. The third plot shows the inverse of the log of the spectrum, the cepstrum. As Cepstrum has repetition in coefficients, only half of them are considered as shown in the fourth plot. The fifth plot shows the High-Time Liftered plot. The cutoff frequency that I have used for Liftering is 21. We see that the peak in this plot indicates peak in excitation characteristics. The location of peak gives us a pitch period in quefrency[1] samples. The pitch location comes at x = 403. The peak has been marked with a '*' in the plot. Therefore the Pitch frequency comes out to be 119.1 Hz.

Question 3

Part (a)



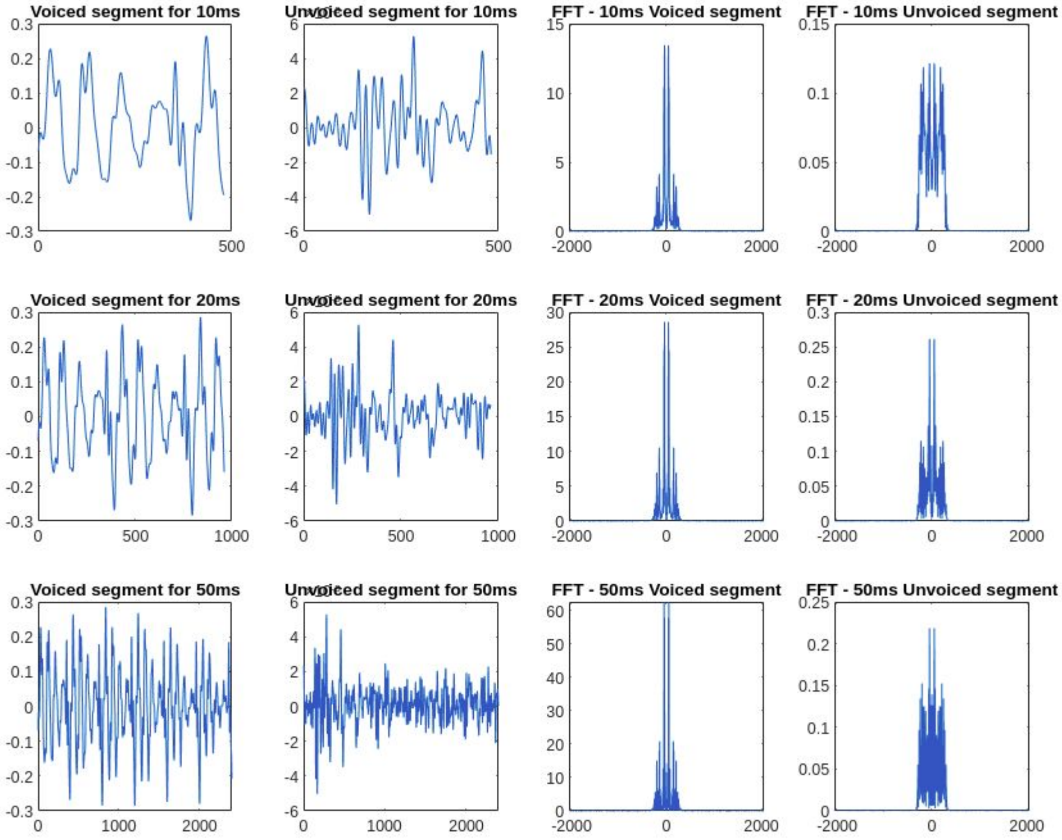
[1] "Cepstral Analysis of Speech (Theory) - Amrita Virtual Lab." http://vlab.amrita.edu/?sub=3&brch=164&sim=615&cnt=1. Accessed 18 Sep. 2020.

I took a 20ms voiced and 20ms unvoiced frame and applied 256-point DFT, 512-point DFT and 1024-point DFT and plotted them. I used Hamming window for windowing all segments.

The first observation is that the FFT of the voiced part has impulses or spans over a short period while the FFT of the unvoiced part spans a longer length on the x axis. The rightmost non-zero value of 1024-point DFT for the voiced segment comes approximately at x = 68 and the rightmost non-zero value of 1024-point DFT for unvoiced segment comes approximately at x = 80. Although the difference is not too much, because I don’t have a perfect/high frequency unvoiced segment in my name, the difference is still visible. This shows that the voiced part contains less frequency as it has impulses at lower frequencies and the unvoiced part has higher frequencies as it is more random and non-periodic as seen in the 1024-point DFT plots.

So as my frame is of 20ms and Sampling Frequency is 48 kHz, I have 960 samples in one frame. If I do 256-point DFT, my segment is truncated from 960 samples to 256 samples and then FFT is calculated and thus we can see both voiced and unvoiced plots for 256-point DFT contain less information than 1024-point DFT. If I do 512-point DFT, my segment is truncated from 960 samples to 512 samples and then FFT is calculated and thus we can see both voiced and unvoiced plots for 512-point DFT contain less information than 1024-point DFT. The 1024-point DFT contains the whole information of the segment and is ideal.

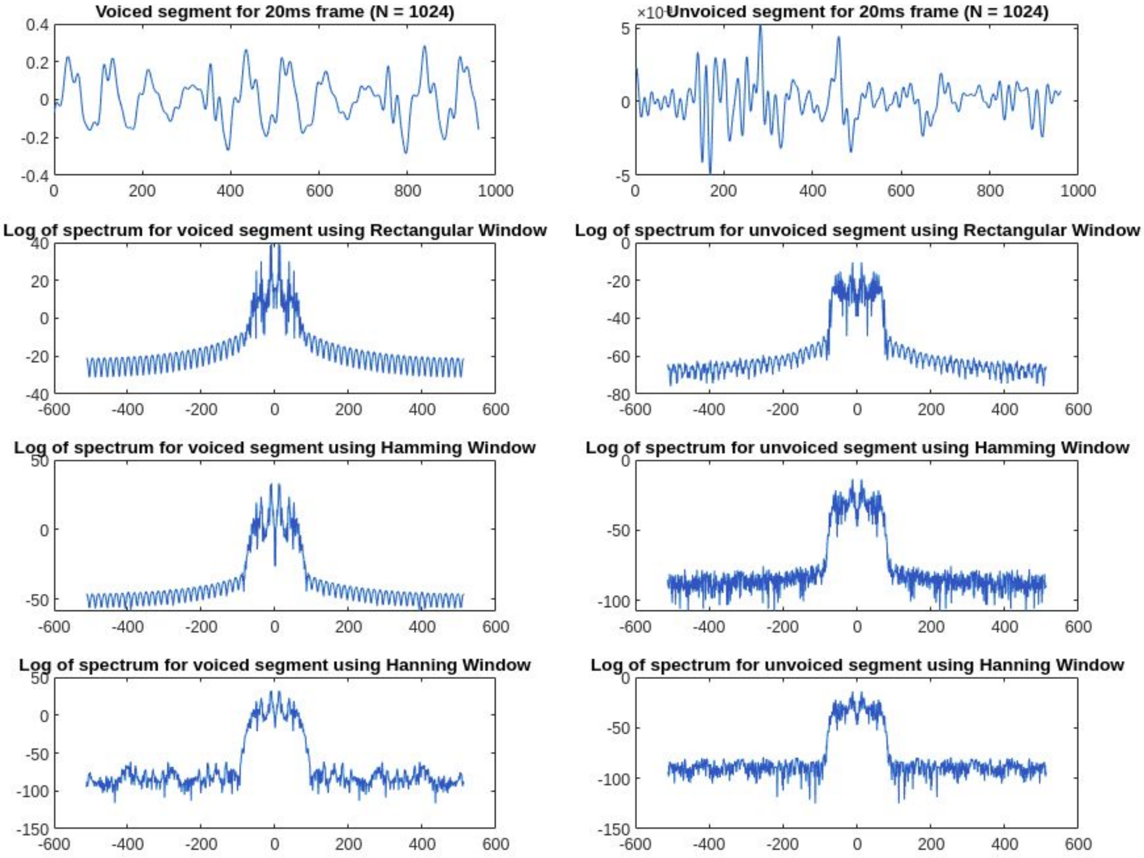
Part (b)



I took 3 sizes of window for this question, 10ms, 20ms and 50ms and applied 4096-point DFT to all of them with a hamming window. The first row is for 10ms, the second row is for 20ms and the third row is for 50ms. We can compare FFT plots for the voiced and the unvoiced regions and see that unvoiced are having impulses at more and higher frequencies than voiced because of random and non-periodic nature. The FFT plot for the unvoiced region of 50ms is very dense and has many frequencies (wider plot) as compared to the plot corresponding to voiced. This is due to the random and non-periodic nature of the unvoiced region. So one important thing to see is that as the size of the window increases we are getting more information about frequency and thus our frequency resolution is increasing. But at the same time our time resolution is decreasing. So the 50ms window has high frequency resolution and low time resolution. The 10ms frame has high time resolution but low frequency resolution[2]. Plots can be observed closely in the code.

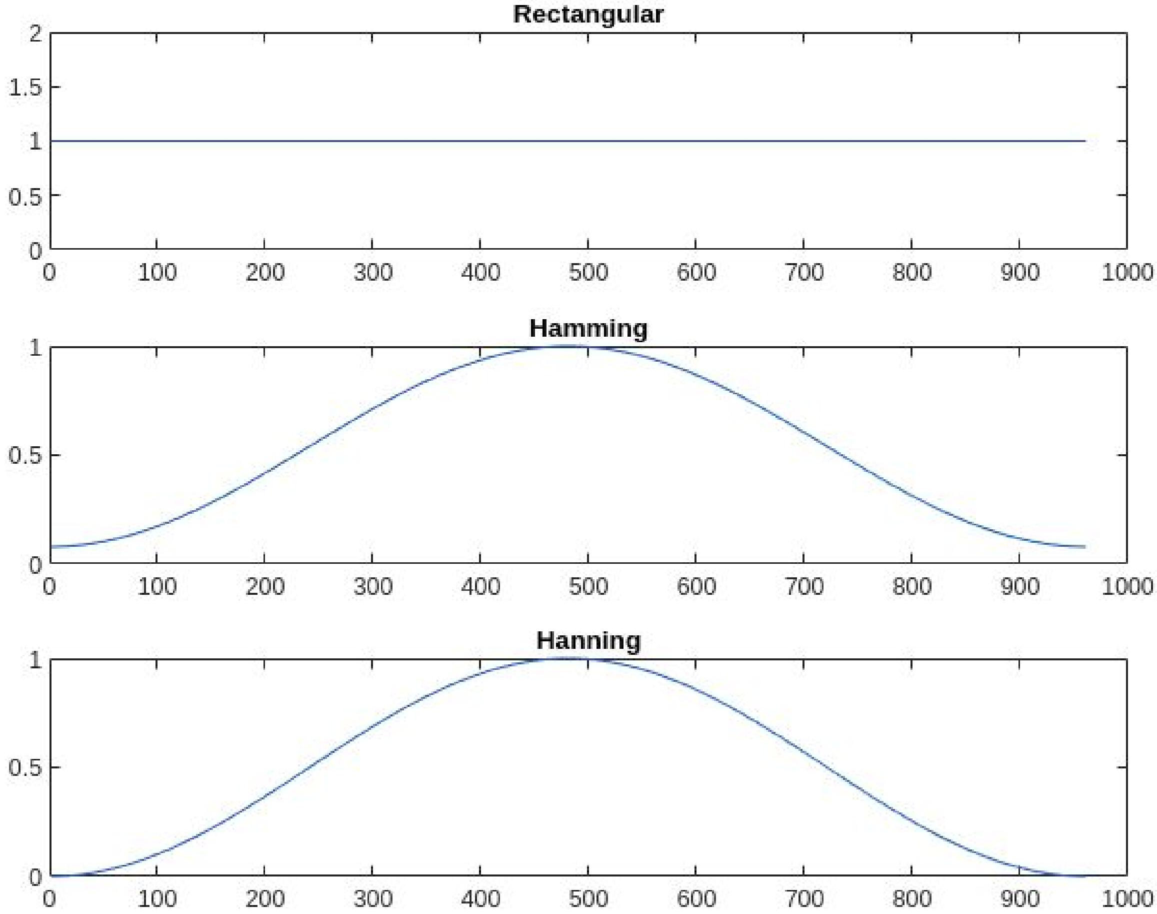
[2] "Short-Term Frequency Domain Processing of Speech ...." http://vlab.amrita.edu/?sub=3&brch=164&sim=908&cnt=2. Accessed 18 Sep. 2020.

Part (c)



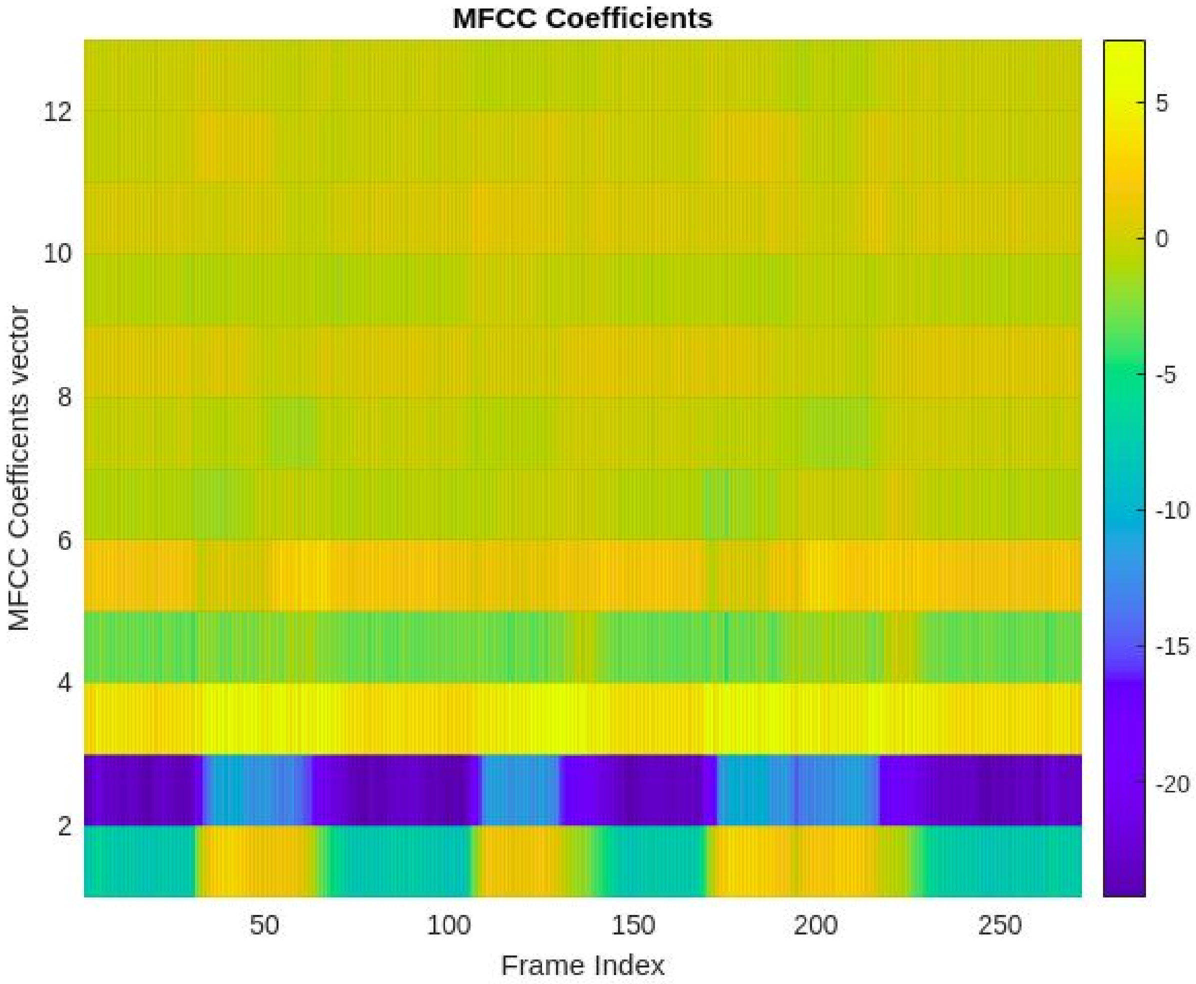
I took 20ms voiced and unvoiced frames and applied 1024-point DFT to all of them using different windows and plotted log of the spectrum. Plotting the log of spectrum makes it easy to compare the windows. We can see that FFT plots of voiced and unvoiced regions for the rectangular window have more noise (When comparing above 0) as compared to other plots using hamming and hanning windows. This is due to high spectral leakage in the rectangular window as compared to other windows. The plots for windows are shown below. We see as rectangular has a sharp drop, this can lead to spectral leakage[3] in the spectrum. Hamming and Hanning windows are smooth and do not lead to aberrations in the spectrum after windowing.

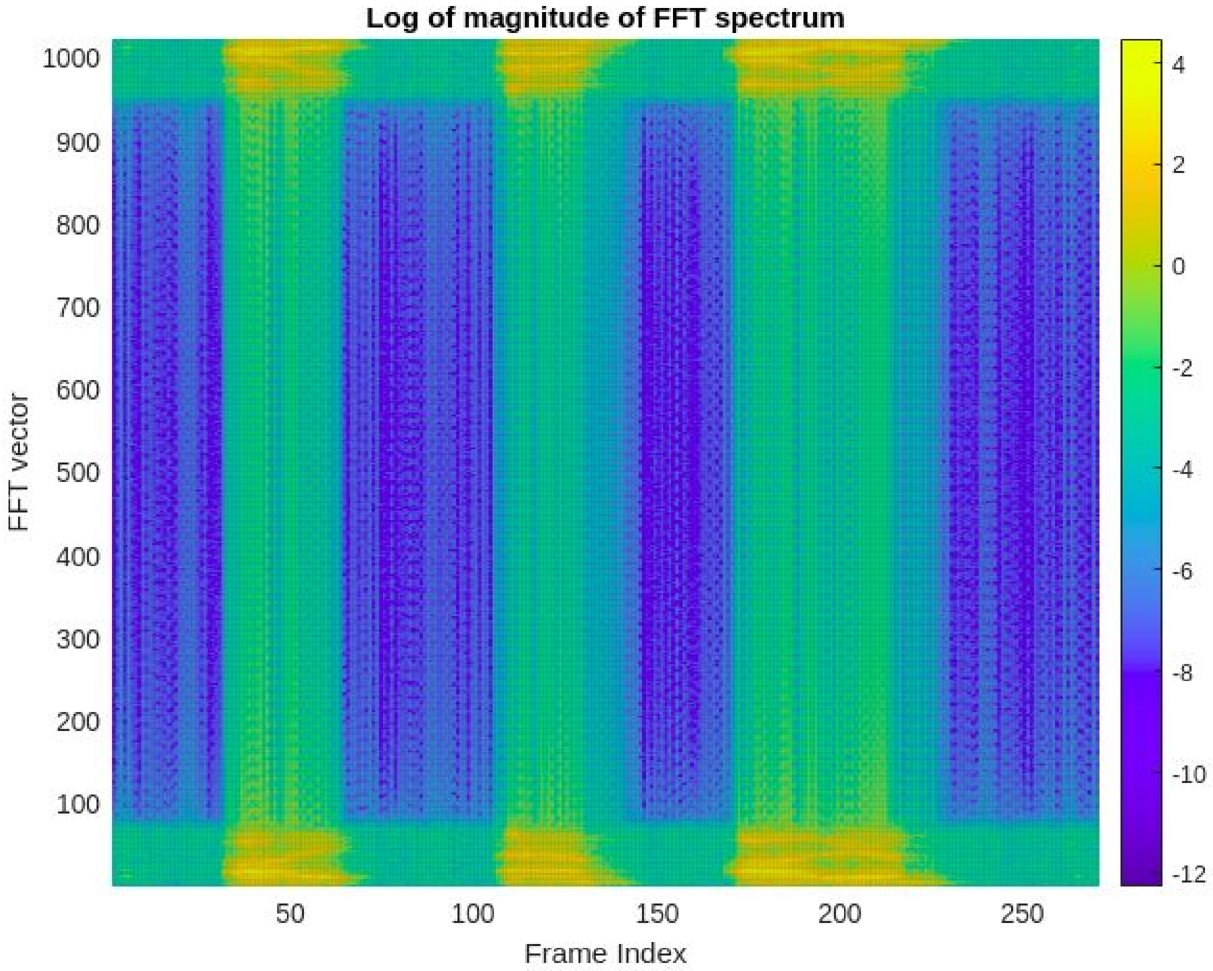
[3] "Short-Term Frequency Domain Processing of Speech ...." http://vlab.amrita.edu/?sub=3&brch=164&sim=908&cnt=2. Accessed 18 Sep. 2020.



The Hanning window falls steeply as compared to Hamming.

Question 4


MFCC Coefficients
MFCC Coefficents vector
12
10
8
6
4
2
50
100
150
200
250
Frame Index
5
0
-5
-10
-15
-20



In the first plot the y-axis indicates MFCC coefficient vector of size 13 and x-axis contains the Frame index. The colour indicates the value of that particular coefficient corresponding to a particular Frame index. We can observe that wherever there are voiced regions in audio, for example between frame number 100 and 150, the coefficients less than 2 correspond to yellow colour as compared to green colour in other regions. This tells us that the yellow region (High MFCC coefficient value), has voiced region and blue region is silent or has some unvoiced part. Also as we rise in the plot we see we can’t really distinguish values as those coefficients contain some excitation characteristics with vocal tract information. This is a great way to visualize and understand the importance of MFCC features.

In the second plot y-axis contains the FFT vector (1024 points) and x-axis contains the Frame index. The colour indicates the log of magnitude of value of the FFT spectrum of a particular point. I have taken log as without log, the range of the values was too big and the plot could not give us important information as this plot does. We observe that wherever there are voiced regions in audio, for example, frame indices between 100 and 150, they are yellowish or greenish in colour as compared to blue colour of other frames. This indicates some higher value of log of magnitude of FFT spectrum due to presence of voiced regions. Whereas blue regions are either silent or unvoiced and thus have low magnitude in the FFT spectrum. Also we see greenish regions at FFT points of the end or the beginning as by default the FFT function returns a symmetric spectrum with values in starting and ending. We can bring the spectrum in centre using the fftshift function. The yellow regions in the beginning of FFT points and ending (the lowest and highest points in the plot) are voiced regions whereas green regions are silent or unvoiced.

If we vary the overlapping length, we can see some minute changes in the plots like, if we increase overlapping length the colours would be more mixed and smooth in transition because changes would be small and the number of frames would increase.