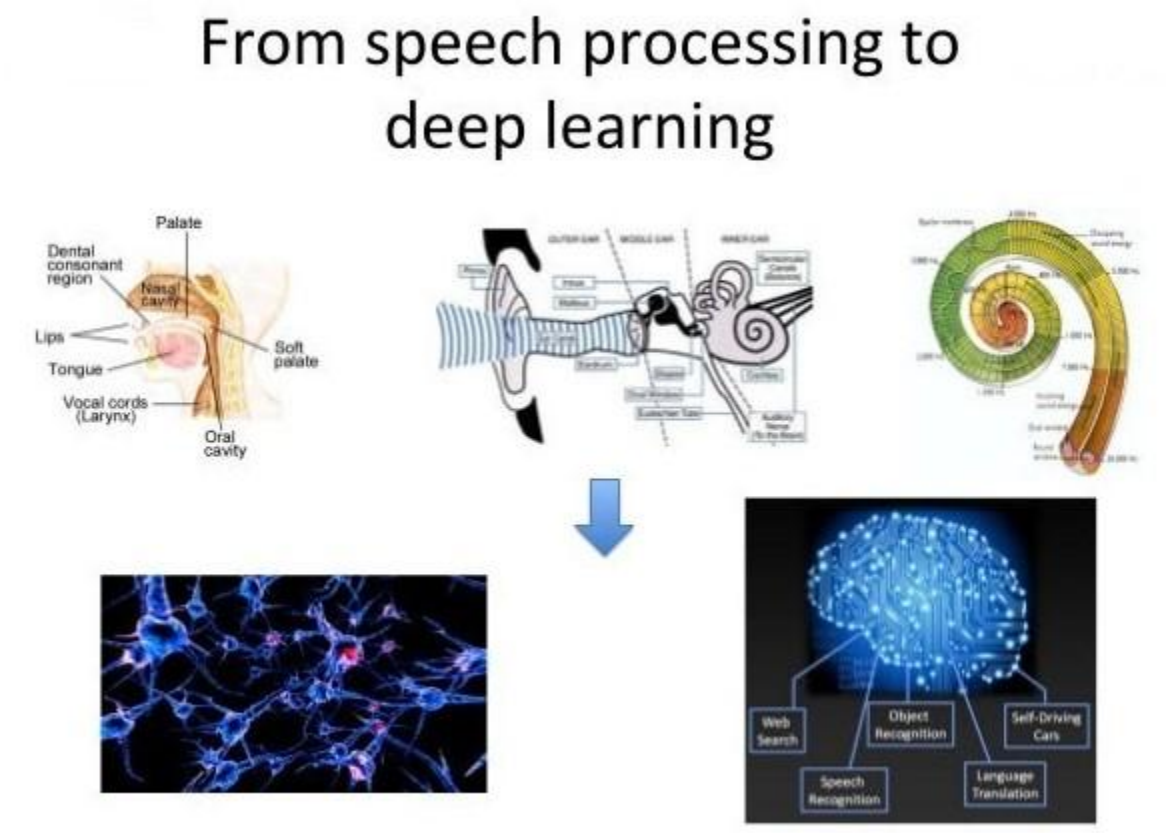
MACHINE LEARNING FOR SPEECH AND COMPUTER VISION

Assignment - 01

Course Code: EC353



Case study 1: Speech Processing Basics

Submitted By:
Abhishek Singh Kushwaha, 19BEC001
3rd Year, 6th Semester, Undergrad ECE Student@IIIT Dharwad

Submitted To:
Dr. Nataraj K, EC353 Instructor, Assistant Professor
Department of ECE@IIIT Dharwad

All codes for this case study and future assignments will be easily available at my GitHub, https://github.com/shakya-abhishek/mlscv/casestudy1

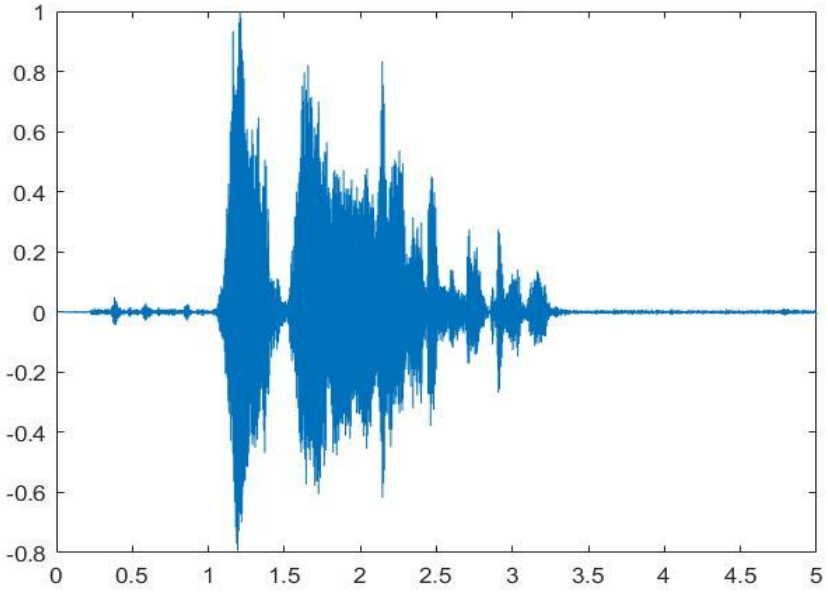
Question 1: Record the utterance "My name is "your-name" in your own voice with a sampling frequency of 16 kHz. Normalize the speech data to lie between -1 and +1.

a) Save this utterance to a wave file named "your-name_record.wav"

Successfully saved the file names abhishek_record.wav

b) Read back the file "yourname_record.wav" to a variable (speech signal) and plot the speech signal with time as the x-axis. (Note: sampling period = (1/sampling frequency)). What is the relation between the time duration of the utterance and the total number of samples in the utterance?

(Total number of samples in the utterance) = (sampling frequency) * (time duration of the utterance)



c) Play and listen to the utterance using the sampling frequency.

Successfully executed

d) Now multiply the variable with 0.5 and 2 (do not normalize after multiplying), and listen to the sounds. Describe the difference between these sounds and the sound played previously in part c.

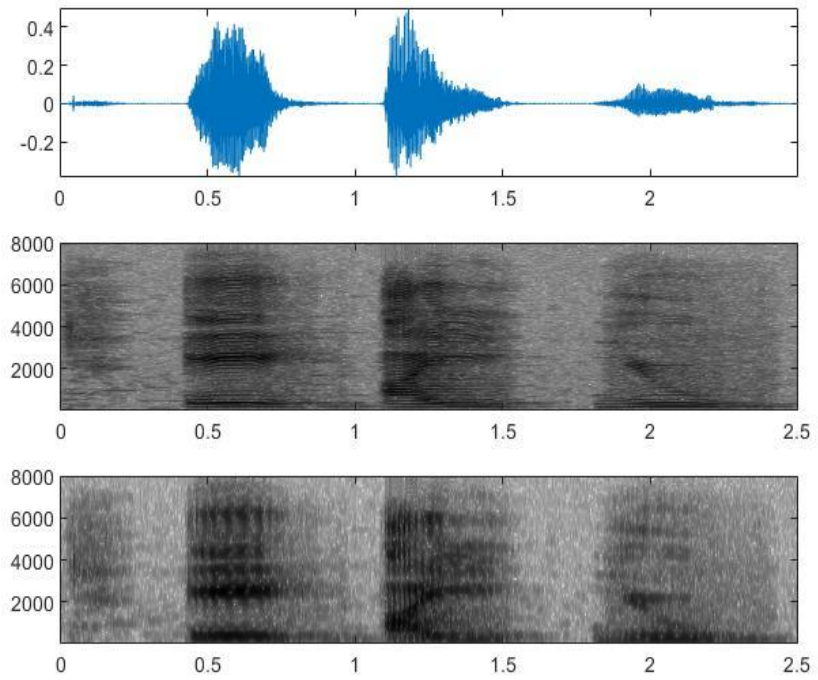
Amplification of normalized speech signal reduces after multiplying it by 0.5 factor
Amplification of normalized speech signal increases after multiplying it by 02 factor

When we multiply by 0.5 the sound heard is low level when compared to the original and for 2 the sound is louder than the original.

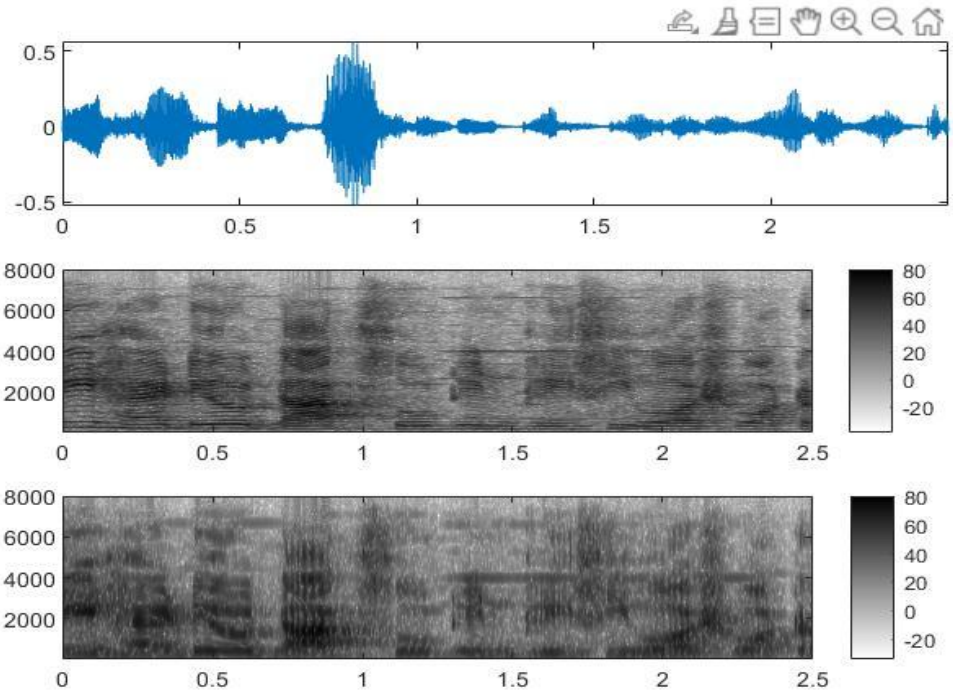
Question 2: Record two utterances in your own voice: the first utterance is three cardinal vowels a, i, u uttered in sequence, the second utterance is "She had your dark suit in greasy wash water all year" Obtain a spectrogram (use the FFT function to obtain after splitting the speech signal into frames) for both the utterances using Hamming window of length 30 ms (narrowband spectrogram) and 5 ms (wideband spectrogram) for window shift of 5 ms and FFT length of 512.

(a) Plot the speech waveform, wideband, and narrowband spectrogram (as grayscale image) as subplots (use "subplot" function in MATLAB) of a figure one below the other with time alignment for both the utterances.

First utterance /a, i, u/



The second utterance “She had your dark suit in greasy wash water all year”

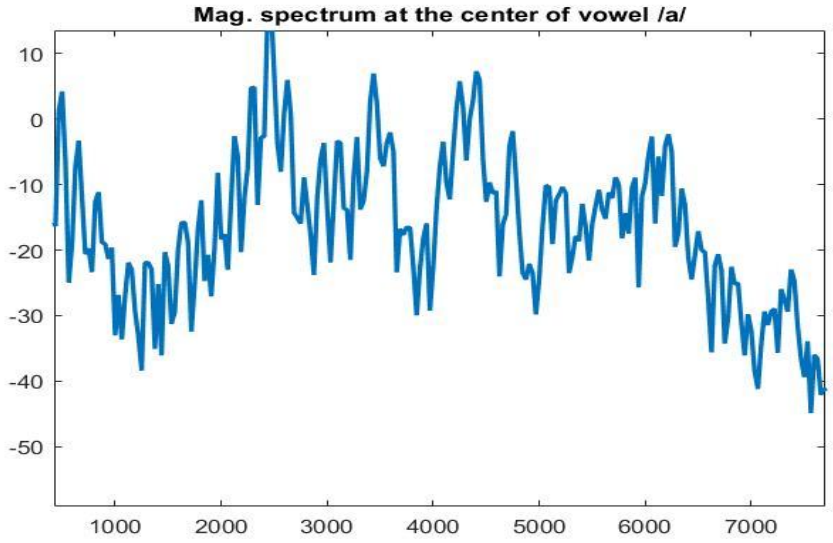


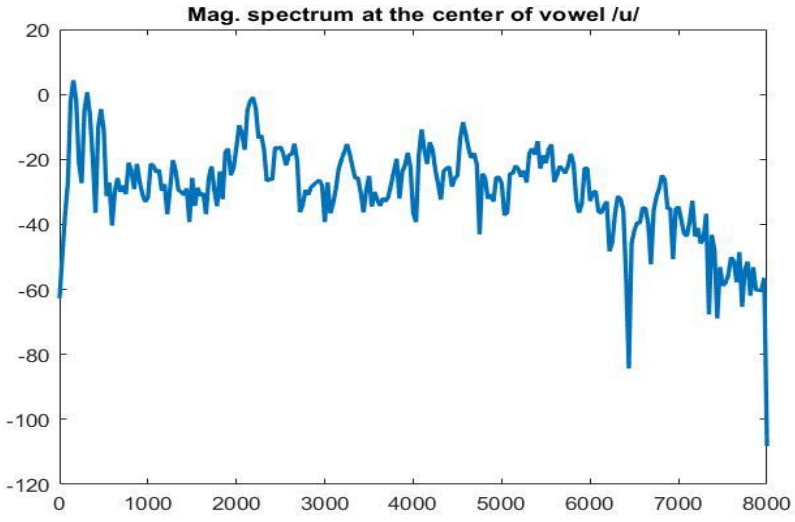
The first spectrogram is a wideband spectrogram and the second one is a narrowband spectrogram.

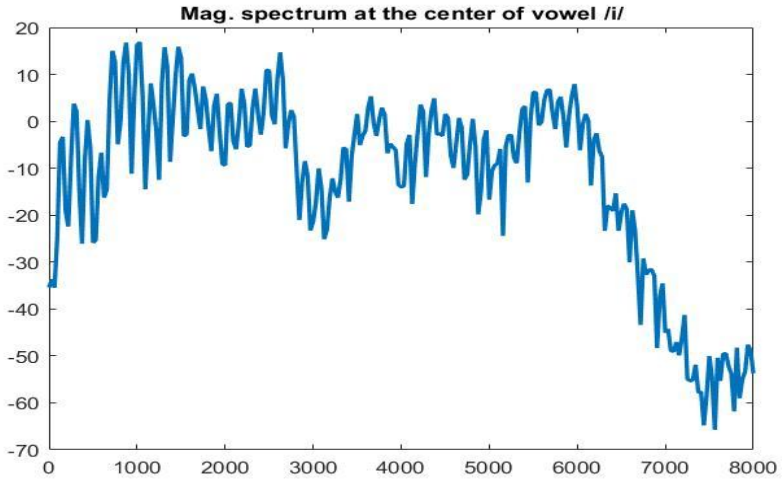
(b) Note down the first three formant frequencies for three cardinal vowels of the first utterance using the spectrogram. Also, obtain the magnitude spectrum at the center of three vowels. Obtain the first three formant resonant frequencies and pitch frequency from the spectrum for all three vowels.

From spectrogram: It is difficult to visually obtain the formant frequencies exactly. /a/: 812 Hz, 1219 Hz, 2500 Hz, /i/:312 Hz, 2281 Hz, 2875 Hz, /u/: 312 Hz, 875 Hz, 2469 Hz

From spectrum at the center frame:







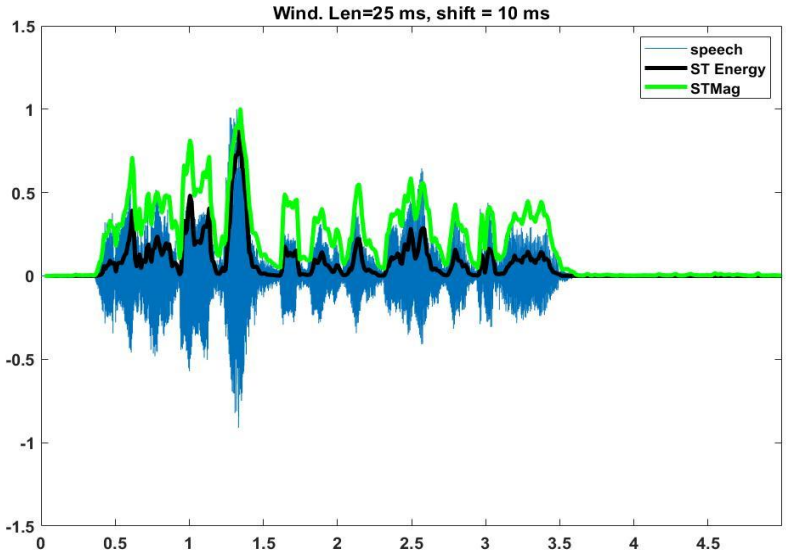
a/: 812.5 Hz, 1188 Hz, 2531 Hz, /i/:250 Hz, 2344 Hz, 2844 Hz, /u/: 281 Hz, 812.5 Hz, 2406 Hz
pitch frequency= 125 Hz (measured as the distance between two pitch peaks in the spectrum)

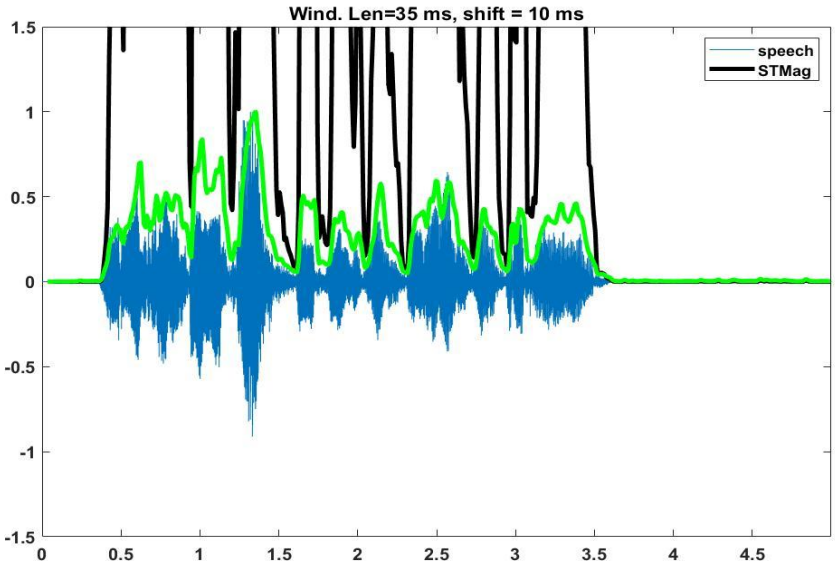
(c) List the set of sounds heard in the IPA notation for the second utterance recorded. Briefly describe spectral characteristics for stop consonants (/k, t/), fricatives (/s, ʃ/), and semivowels (/w, r/). The description should match the spectrogram shown.

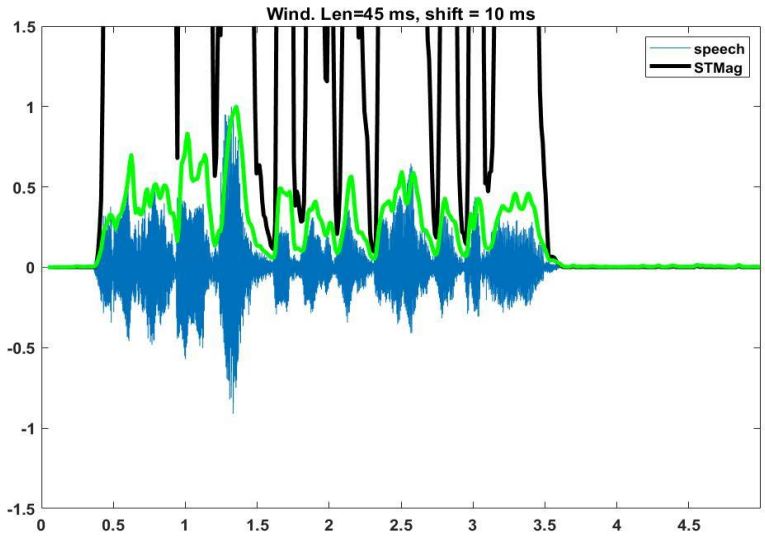
Set of sounds heard in IPA notation: /ʃ I, h ae d, y U r, d a r k, s u t, I n, g r i s i, w a ʃ, w a t /\r, a l , y I r /
/k,t/: Low energy regions with no spectral information, for /t/ shows a voice bar.
/s, ʃ/: Spectral energy starts at 3797 Hz for /s/ and at 2797 Hz for /ʃ/.
/w, r/: Semivowel /w/ has lower first and second formant frequencies than vowel /a/ and /r/ has low third formant than adjacent vowel

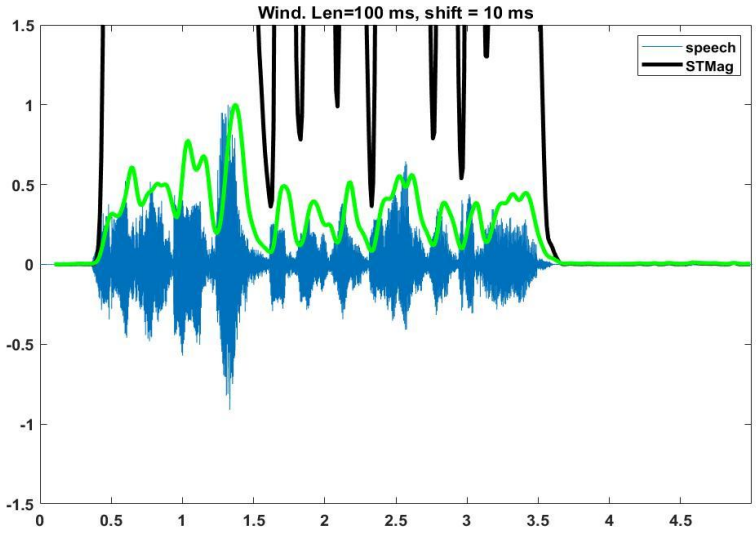
Question 3: Obtain the time-domain features for the second utterance recorded in question 2.

a) Obtain the short-time energy and short-time magnitude using a window length of 25 ms with a window shift of 10 ms. Plot both features overlapping on the speech waveform. Obtain the short-time magnitude for window lengths of 35 ms and 45 ms with the window shift of 10 ms. Explain the effect of the change in window length on the short-time magnitude.



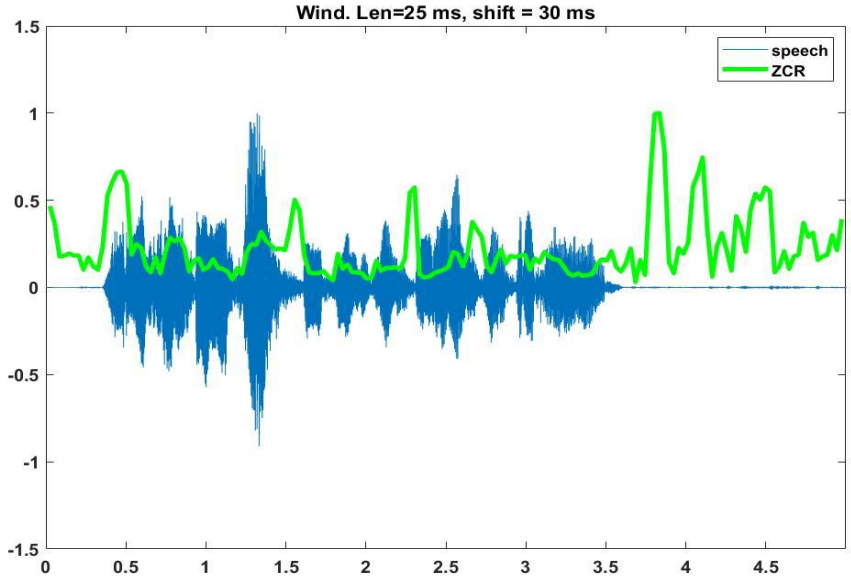




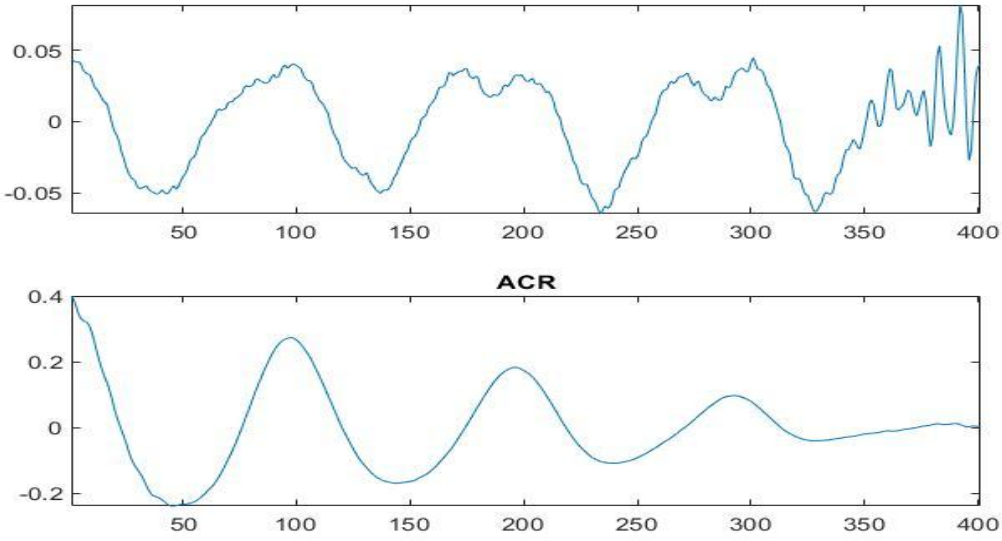


There were no significant changes for different window lengths. However, if we increase the window length to 100 ms we can see that the sudden changes are suppressed and we can get a better speech outer envelope.

b) Obtain the short-time zero-crossing rate (ZCR) for a window length of 25 ms and window shift of 30 ms. Plot the ZCR overlapping on the speech waveform.



c) Obtain and plot the short-time autocorrelation for the center frame of vowel /a/ in word 'dark' and obtain the pitch period based on the autocorrelation function.



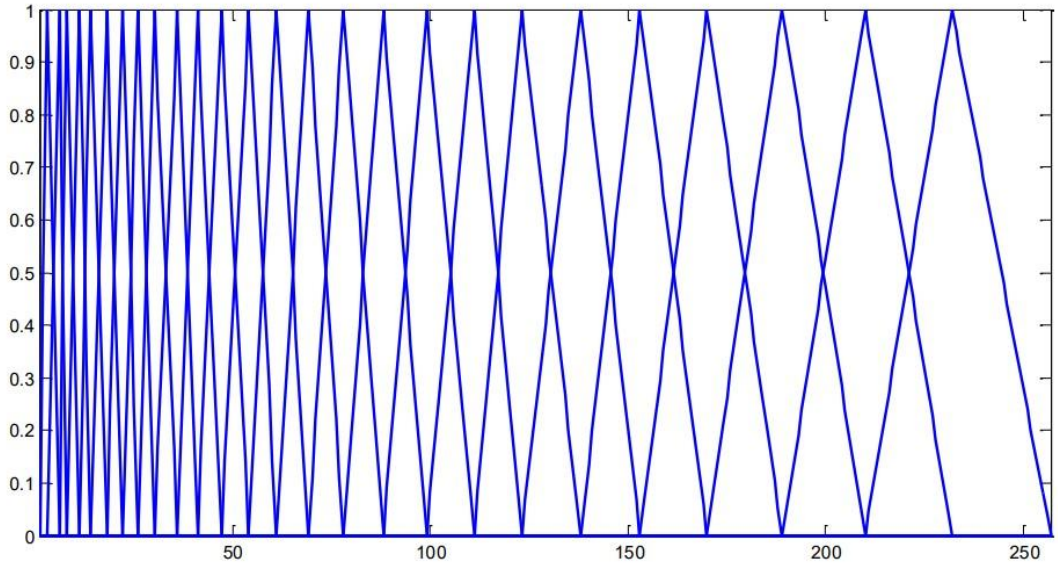
Pitch period = 140 samples => (140/fs) S = 0.0088 S
Pitch frequency = (1/pitch period) = 113.63 Hz

d) Briefly describe how the short-time energy/magnitude and short-time ZCR can be used to obtain the boundaries for voiced and unvoiced sounds using example values in the utterance.

Voiced regions (vowels, semivowels) have high short-time magnitude (>0.3) and low ZCR (<0.36)
Unvoiced regions (fricatives, unvoiced stops) have low short-time magnitude (<0.2) and high ZCR (>0.5)
However, voiced stops had low short-time magnitude (<0.2) and low ZCR (<0.2) due to voice bars.

Question 4: Obtain the first 13 MFCC, delta MFCC, and delta-delta MFCC features for the first utterance recorded in question 2 using the method described in the class with 26 mel-filters, window length of 25 ms, and window shift of 10 ms. Use T=2 for calculating the delta and delta-delta coefficients, and α=0.9 for pre-emphasis. You should explain the code for MFCC computation if asked during evaluation.

Mel-filterbank generated using the equations in ppt for FFT Length of 512



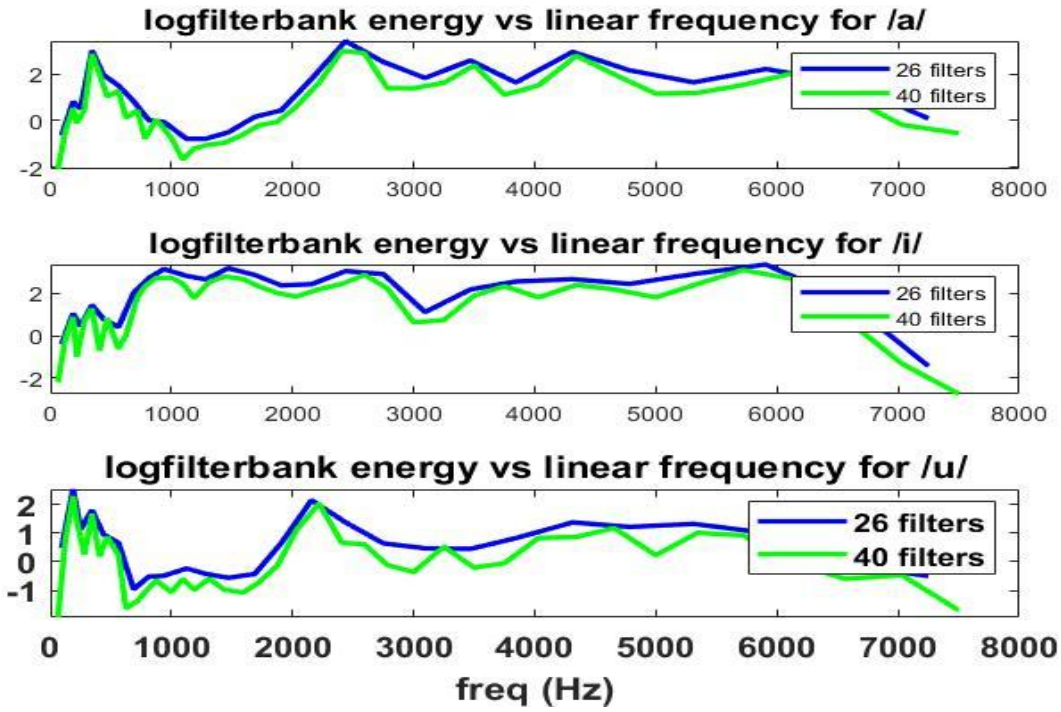
a) Calculate the distance between the MFCC values at the center frame of the three cardinal vowel sounds. i) distance between /a, i/, ii) distance between /a, u/, and iii) distance between /i, u/. Explain your observation on the distance values. For distance calculation, ignore the first coefficient and use the remaining 12 coefficients and calculate the distance as √ ∑ ( ), where mfcc1 and mfcc2 are the MFCC coefficients for two vowels.

(i) /a, i/ distance = 195.2, (ii) /a, u/ distance = 122.12, and (iii) /i, u/ distance =165.01

The vowels /a/ and /u/ are similar as they are both back vowels. But front vowel /i/ is different from /a/ and /u/ and thus results in large distance. Observation is similar to that observed in vowel triangle.

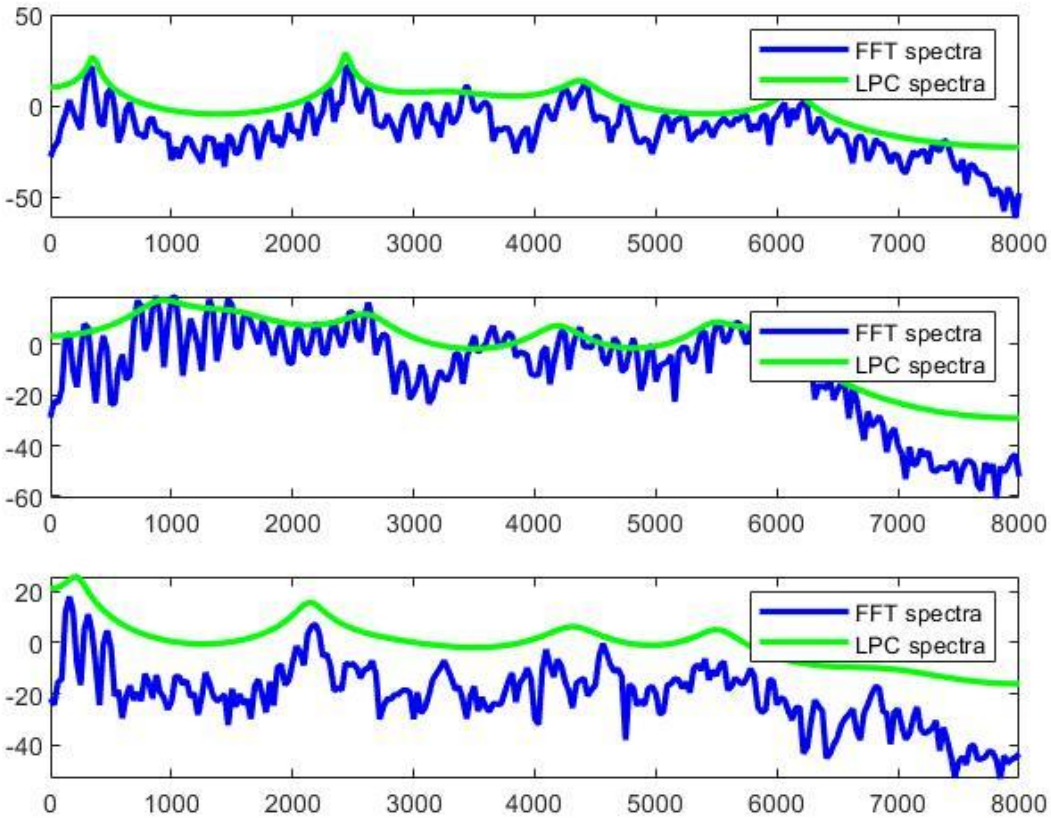
b) Obtain log-filterbank features at the center frame for three cardinal vowel sounds. Plot the log filterbank features versus the frequency (linear frequency value) of the center frequencies of the filter bank.

c) Compute log-filterbank features using 40 mel-filters for the center frame for three cardinal vowels. How do the log-filterbank features change when compared to 26 filters as in part b?



We can see a rise and fall of energy for vowel /i/ with clear formant in 2000 to 4000 Hz. For vowel /a/, energy is large between 1000 Hz due to first and second formant frequencies in this region. For vowel /u/, we can see broad low-frequency energy due to low first and second formant frequencies. Increasing the number of filters reduces the smoothness of the filterbank outputs.

Question 5: Obtain the linear prediction coefficients at the center frame for three cardinal vowel sounds using the first utterance of question 2. Use the "lpc" function of MATLAB to compute the coefficients. Obtain the LPC spectrum for three vowels at the center frame using the vocal tract transfer function assuming G=1.



Observe that the LPC spectrum is a smooth spectrum of the original spectrum as we approximate the speech spectrum using poles. (Prediction order was not mentioned in the question, assumed commonly used order of 12)

THE END