CS6300: Speech Technology
Lab Assignment 3:Report

Abhishek Sekar EE18B067
Keshav Rao MM18B021

March 2021

1 Question 1

This question requires us to compute the envelope spectrum using LPC analysis. We employed two methods to do this, one by finding out the LPC coefficients using the covariance matrix of the respective speech frames and then employing the Levinson Durbin recursion algorithm. The second method employs the inbuilt function LPC of the librosa library in python which is computed via Burg's method.
We used an order of 12 for the LPC filter which roughly corresponds to 6 formants per speech frame. Given below are the log magnitude plots of the DFT spectrum and the corresponding LPC envelope. A hamming window was used on the speech frames before computing the DFT and the Linear Prediction Coefficients as it reduces unwanted spectral leakage about the peaks.A gain equal to the RMS of the speech frame is multiplied to the LPC spectrum in order to get an accurate result.

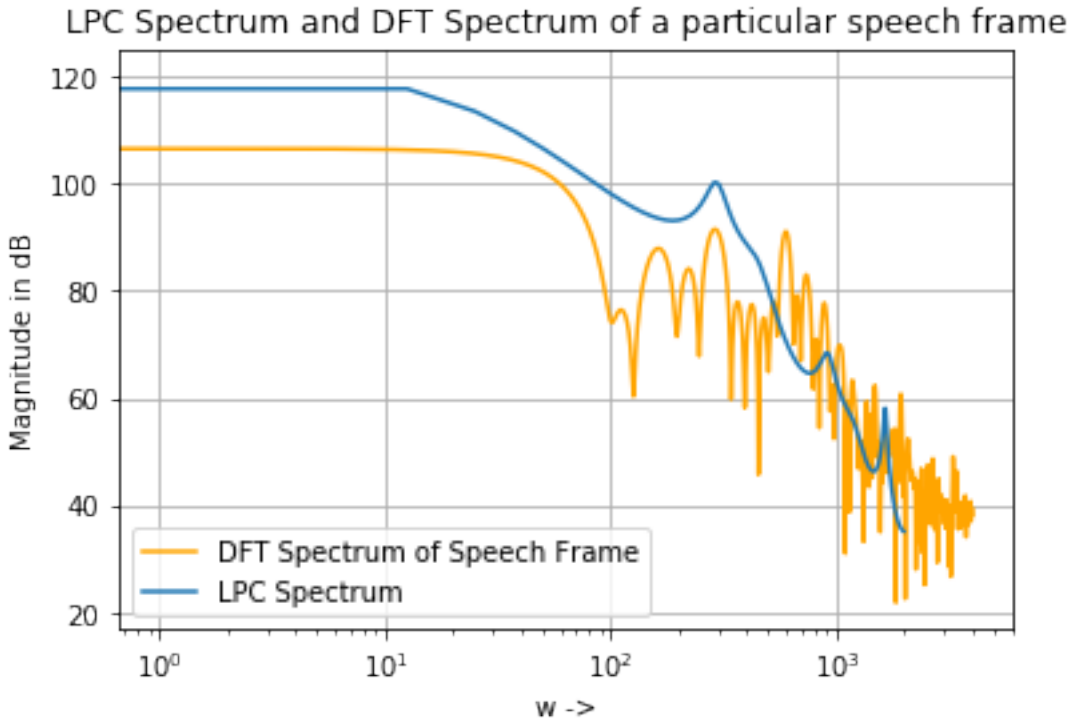

Figure 1: The LPC Spectrum for the word Mask

As seen in the plot, the LPC spectrum somewhat envelopes the DFT spectrum of speech. We can also observe 6 formant peaks (2 are very close to other formant peaks) in this particular speech frame.

2 Question 2

After getting the linear prediction coefficients for a frame like in Question 1, we apply the Linear Prediction Analysis to the entire signal and get LPCs for it.

Using the Librosa.lpc() function in Python, we push the coefficients into an equation, solve it and get the roots. The formant frequencies are nothing but the hyperbolic tan inverse of the roots.
We get 6 formant frequencies for each of the 36 frames in this particular signal. Since, the word "Mask" has both vowels and consonants, the formant contour plots does not show a linear plot for each frequency but instead show some distortion.

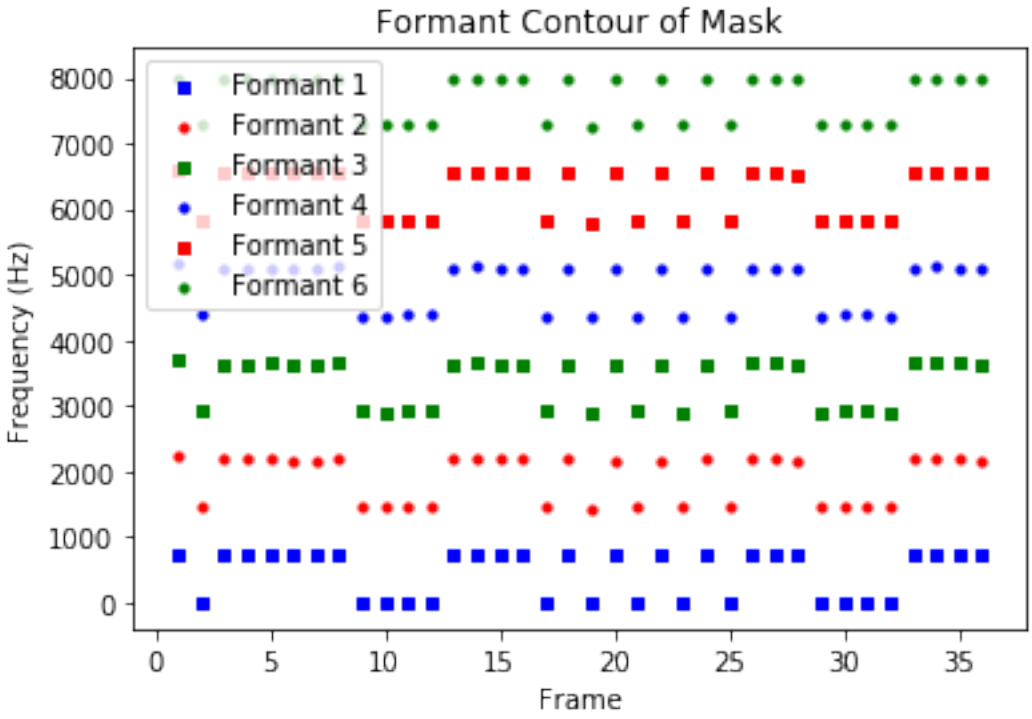

Figure 2: The Formant Contours of the word mask

3 Question 3

The short-time energy of speech signals reflects the amplitude variation. Whereas, Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero.
The simplest method to distinguish between voiced and unvoiced speech is to analyze the zero-crossing rate. A large number of zero crossings implies that there is no dominant low-frequency oscillation.
The voiced speech should have higher zero crossing rate and that is what the plots show. The voiced speech also has a higher average energy as shown by the short term energy plot.
These two are combined to determine the voiced portion of speech in VAD and this is determined in the last plot which roughly corresponds to the sound /a/ in mask.

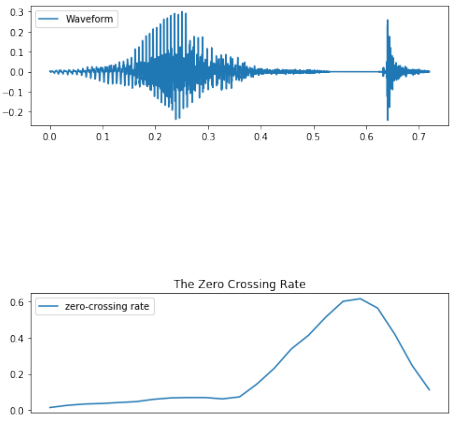

Figure 3: Plot of zero crossing rate

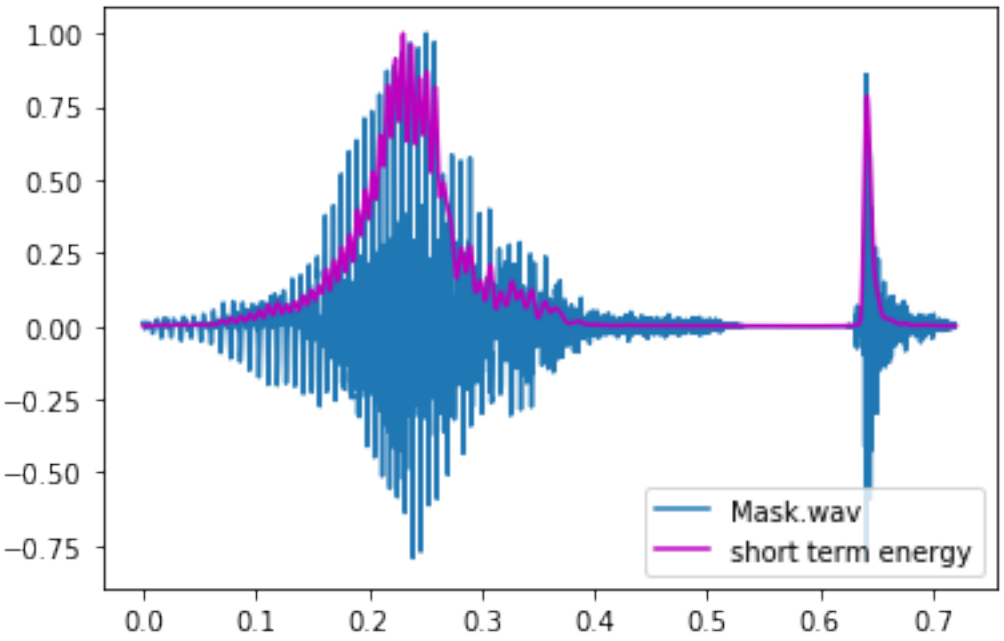

Figure 4: The speech spectrum and corresponding short term energy plot

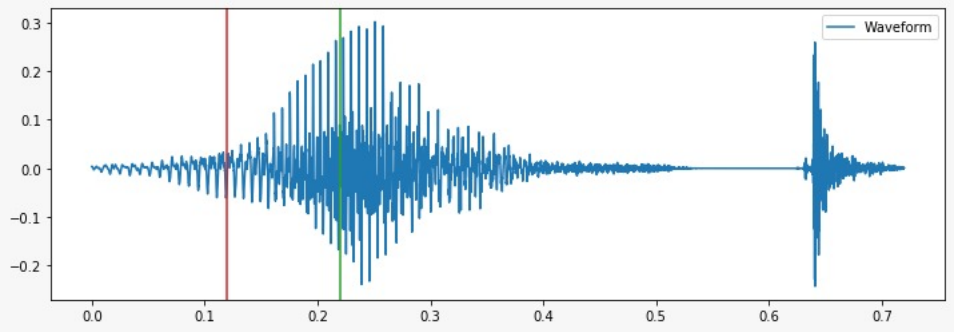

Figure 5: The voiced region of speech as determined by VAD

4 Question 4

The pitch was estimated using the auto-correlation method.Just like in the previous questions, speech frames of duration 20 ms are taken and their autocorrelation is computed. The autocorrelation value will be maximum at the start since the signals completely correlate with each other with the next maxima occurring at the pitch period. This is used to determine the pitch period and subsequently its frequency in Hertz.

From the plots we observe that except for some portion in between, the pitch frequency is too high and erratic. This is because, the portion in between corresponds to the sound /a/ in the word mask and that is voiced. We can also observe that the uniform pitch period discussed above corresponds to the portion of the signal determined as voiced in Question 2.

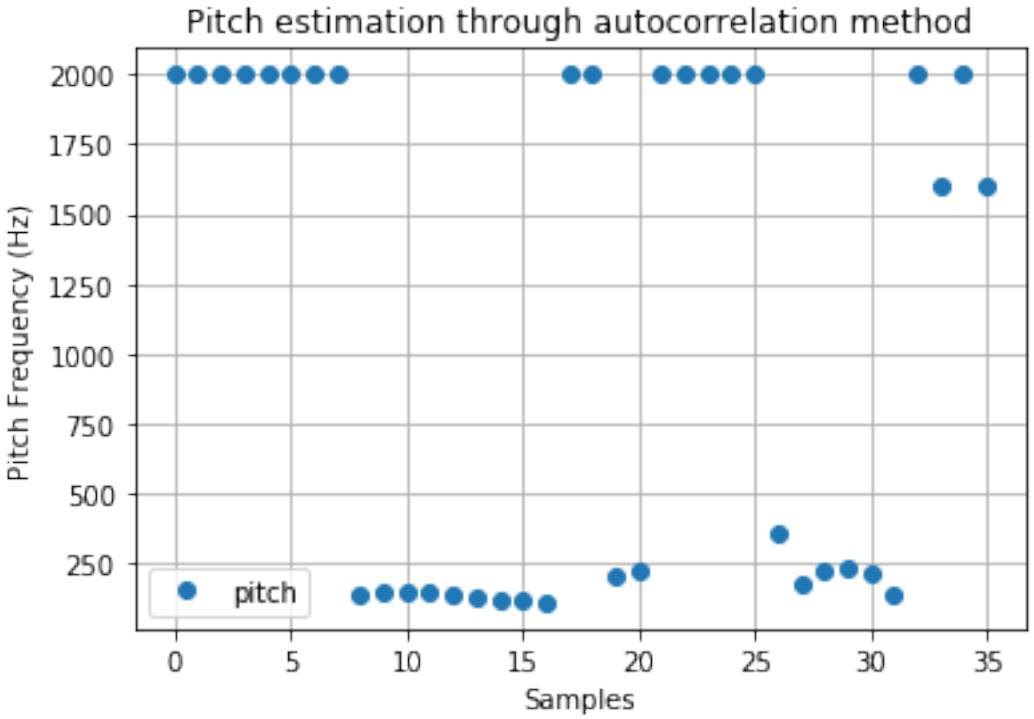

Figure 6: Pitch Frequency plot

5 Question 5

In this question, the speech frames are convolved with the hamming window before their short term Fourier transform is taken. This STFT gives frequency-time characteristics of the speech frame for short durations. This STFT is then plotted in a dB scale and that is presented as a spectogram where the brightness represents the energy intensity. This is done twice for the sentence, 'We were away a year ago', spoken once each by a male and a female speaker. The formant frequencies and the pitch periods are also computed. From the spectogram plots we observe, that the female plot has higher intensity at higher frequency components when compared to the male voice. Observation on the pitch period tells us that the female speech's pitch frequencies are higher than the male speech.

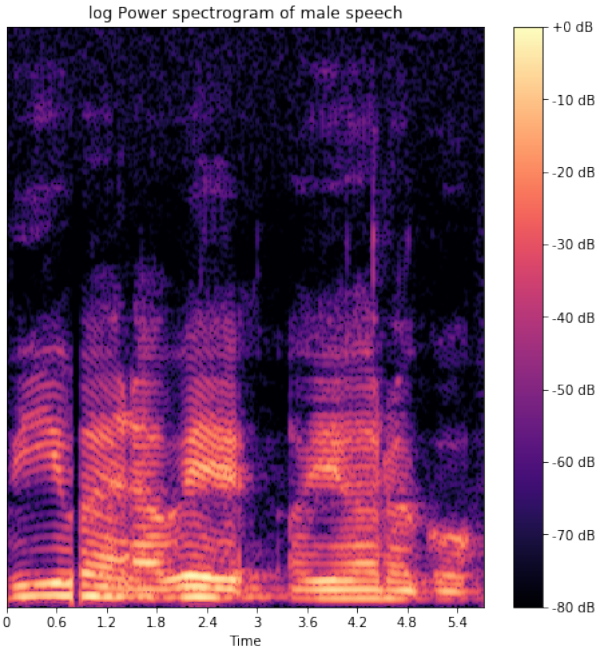

Figure 7: Spectogram of male speech

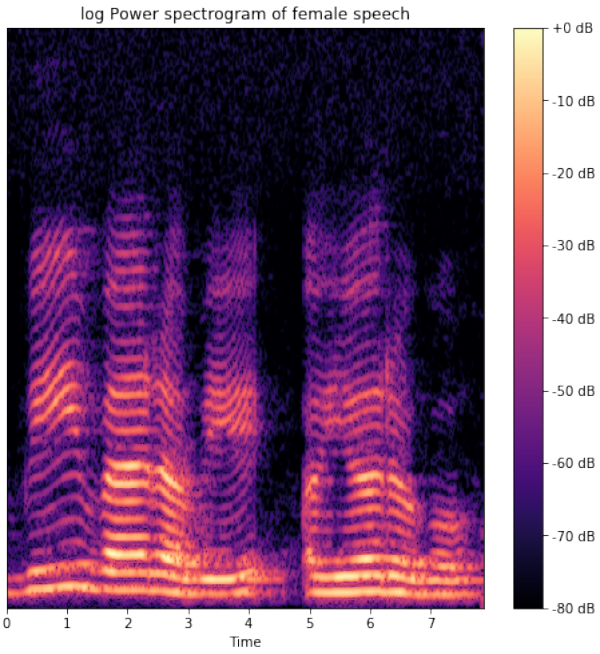

Figure 8: Spectogram of female speech