Assignment 4,5 - Speech Technology

EE17B047 - EE17B035

March 2020

1 Smoothed Spectrum - Cepstral Processing

This section give a description of the process to compute the smoothed spectrum of a speech signal employing the cepstral processing method

1.1 Implementation

Below is a description of the various steps which need to followed to obtain the smoothed spectrum :

- Firstly, window the speech signal for a short time say about 25ms.
- The, take the windowed signal and pad it with zeros to the next closest power of 2 for ease of fft calculation.
- The, compute the fft of the padded signal and take the magnitude spectrum. We can observe from the plot of the magnitude spectrum that there are many lobes in the spectrum which makes it difficult to identify the formants of the signal.
- Below is the formula for the cepstral calculation :

  $$cepstrum = F^{-1}(log(\|Ff(t)\|^2))$$

  We can see that to compute the cepstrum we need to take log of the magnitude specturm followed by an inverse fourier transform. The log is to replicate the functioning of the ear i.e. the sense of perception of sound vs frequency is in form of a logarithm.
- Now, from the cepstrum we observe that there are two peaks in the magnitude spectrum wherein the later corresponds to the high frequency component in the magnitude spectrum which we want to get rid of in order to smooth out the spectrum.
- We need to choose a suitable window i.e. rectangular or hamming of suitable length and mask the cepstrum.
- Once we have the masked cepstrum, perform the fft of the cepstrum in order to obtain the smoothed signal.

1.2 Plots

Below is the plot of the smoothed spectrum using the rectangular window. Figure 1 has been plotted taking a large window size which includes the pitch component of the cepstrum. Hence, there is no observable smoothing. On the other hand, Figure 1 has been plotted taking a small window size which doesn't include the pitch component of the cepstrum. Hence, there is smoothing of the spectrum.

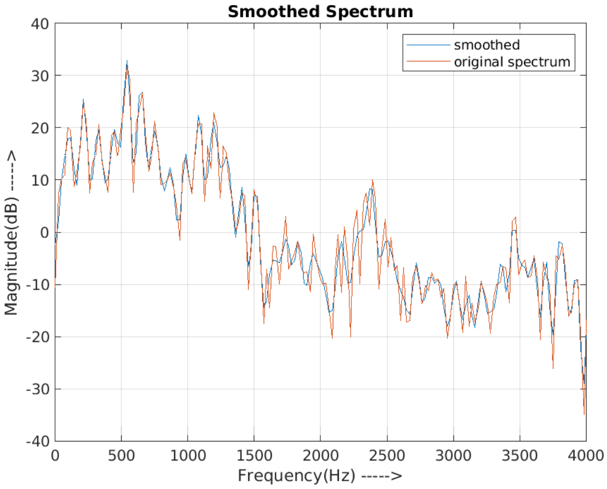

Figure 1: Smoothed Spectrum of the speech signal using large rectangular window

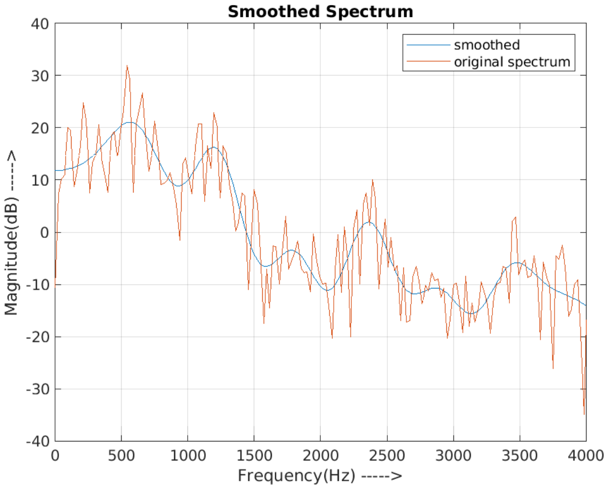

Figure 2: Smoothed Spectrum of the speech signal using small rectangular window

Below is the plot of the smoothed spectrum using the hamming window. Figure 1 has been plotted taking a large window size which includes the pitch component of the cepstrum. Hence, there is no observable smoothing. On the other hand, Figure 1 has been plotted taking a small window size which doesn't include the pitch component of the cepstrum. Hence, there is smoothing of the spectrum.

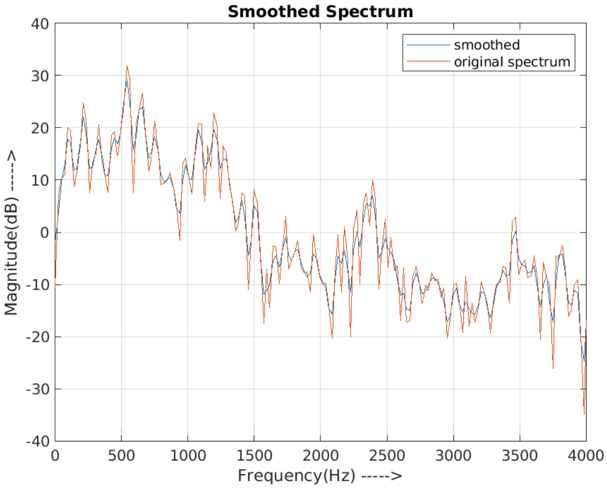

Figure 3: Smoothed Spectrum of the speech signal using large hamming window

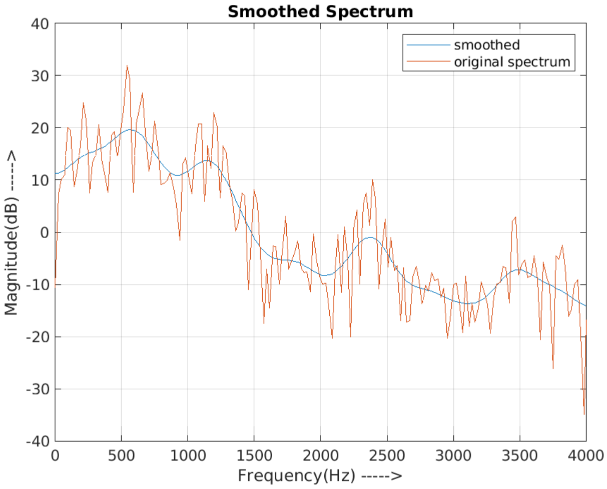

Figure 4: Smoothed Spectrum of the speech signal using small hamming window

1.3 Observations

- We can clearly observe from the smoothed spectrum that the formants are simpler to identify and we can extract them employing signal processing techniques.
- The nature of the window has no noticeable effect on the form of the smoothing. The ability of smoothing remains the same irrespective of the window.

2 Formant Extraction & Formant Contours

From the previous section, it was evident that in the smoothed spectrum, the formants embedded themselves as the peaks of the spectrum. Hence, the formant extraction becomes simpler.

2.1 Implementation

- We can extract the formants using the *findpeaks()* function. This returns all the local maxima in the spectrum and we can pick the maxima which we need.
- It is generally observed that the first formant is between the range 200 - 1500 Hz and the second formant is in the range 500 - 3000 Hz. Hence, we can apply this constraint to the list of peaks obtained.
- Further, we can also see that the peaks of the first formant and second formant are generally the highest among the peaks. Hence, using the above constraints we can get a reasonable estimate of the formants.
- In order to obtain a smooth curve for the formants, we perform a moving average on the curve.

2.2 Observations

- We can observe from the plot in the next page that the formant values are in agreement to the accepted value ranges for most cases. Identification and estimation of first formant is a simple task.
- However, since there are multiple peaks around the second formant, it becomes a harder task owing to which some discrepancies might occur.

2.3 Plots

Below is the plot of the formants as a function of time for all the 5 vowels.

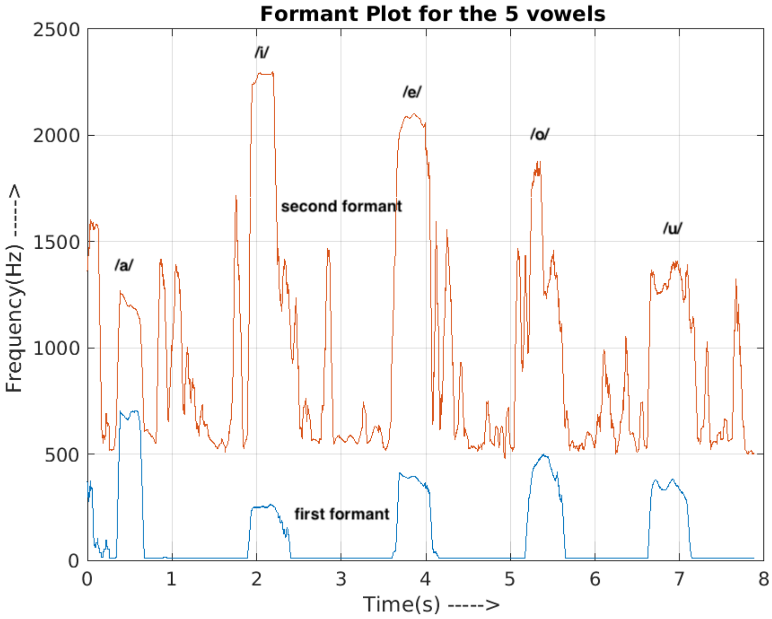

Figure 5: Formant Plot vs time for all the vowels

The corresponding time domain signal of the recording is as given below.

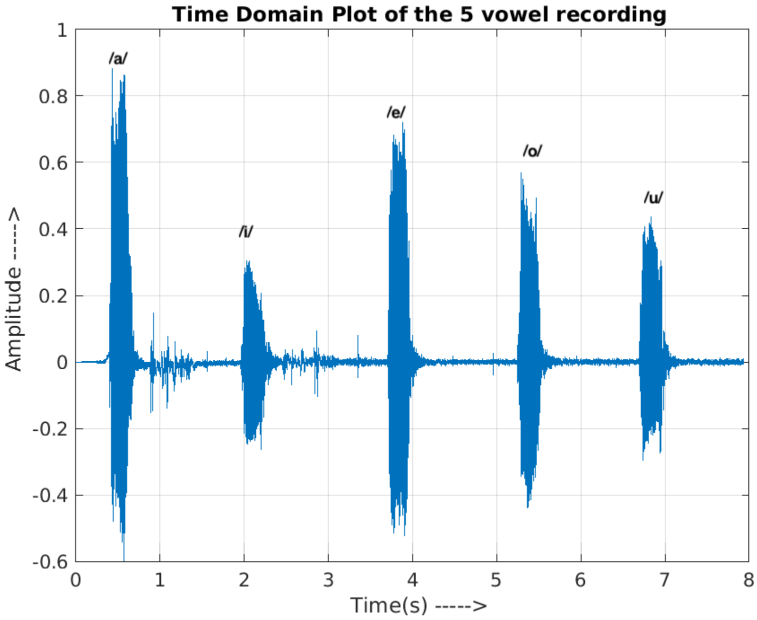

Figure 6: Time domain plot of the 5 vowels

3 Voice Activity Detection

It has been determined that about 3/5th of speech signal is silence or noise. Hence, it becomes extremely vital to identify the parts of the speech signal where we have meaningful information about the signal. First, we will calculate the short term energy and the short term zero crossing rates of the signal. Since voiced speech can be characterised with:

- High-energy
- low zero-crossing rate

we can compare the calculated short term values to some threshold and determine the parts of signal which are voiced speech and which are not.

3.1 Calculating Short Term Energy

The short-term energy corresponds to an estimation of the energy of a signal over short windows. It is calculated as follows:

$$E_n = \sum_{m=n-N+1}^{n} s^2[n]$$

where s is the signal, N is the window size and E_n is the the short term energy for a particular part of the signal.

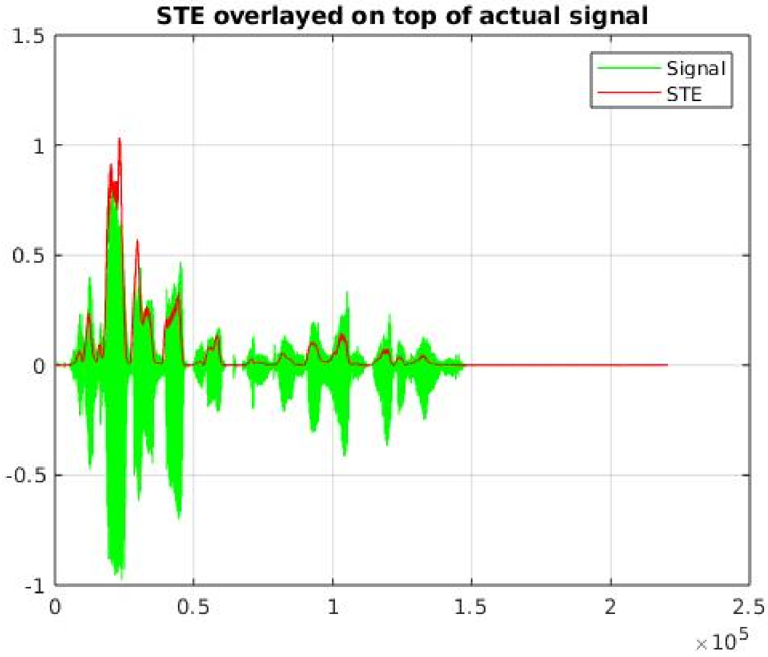

Figure 7: Voice Activity Detection using the short term energy method

3.2 Calculating Zero Crossing Rates

The short term zero crossing rate for a signal is calculated as follows:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x[m]] - sgn[x[m-1]]|w[n-m]$$

gin where x is the signal, sgn is the signum function with a small change, i.e sgn(0)=1, w[n] is a rectangular window function, non zero from n=0 to n=N-1, and with area under it normalized to 1 (N is once again the window length), and Z_n is the short term zero crossing rate for a particular part of the signal.

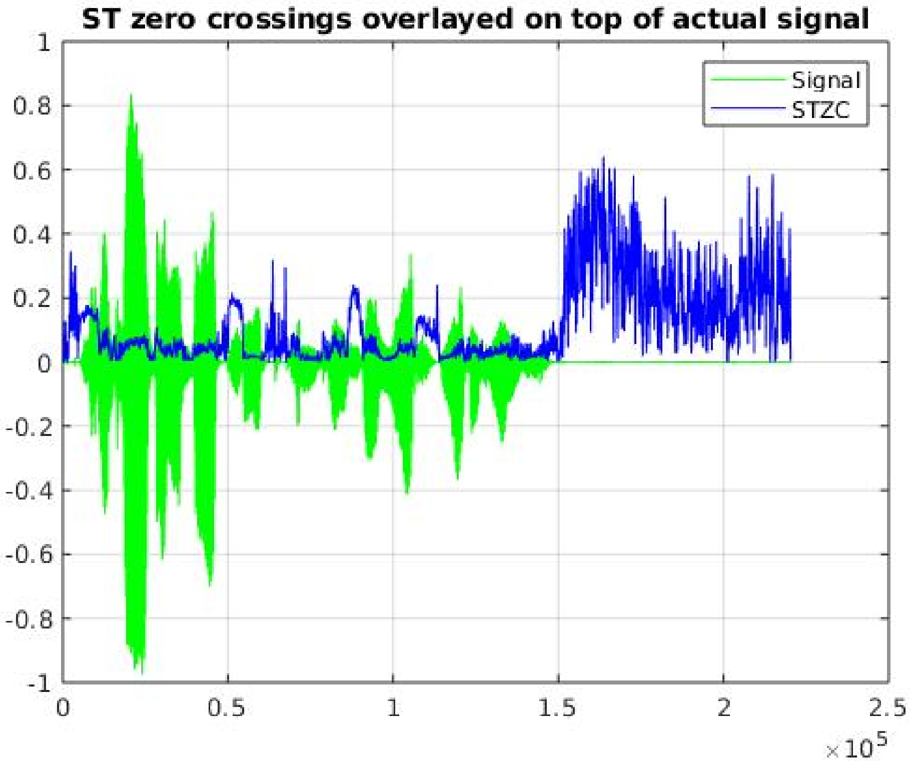

Figure 8: Voice Activity Detection using the zero crossing method

3.3 Performing VAD using the above two metrics

We know that the Voiced part of the speech have high energy and low zero crossing rate. After several trials, we have decided to set the threshold for STE equal to 10% of the average STE of the signal, and the threshold for Zero Crossings as the average zero crossing rate of the signal. The plot below compares the actual signal and signal after VAD processing.

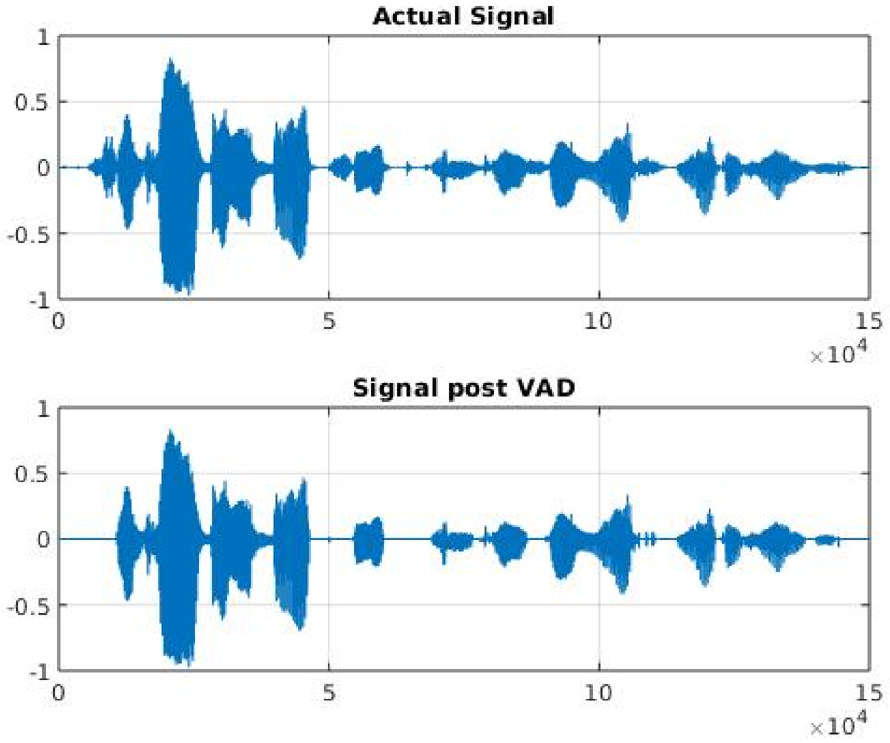

Figure 9: Voice Activity Detection using the two methods

3.4 Observations

- For most parts of the speech signal, the Voice activity detection works well. However, when there are sounds such as fricatives which have very less power and higher crossing rate, these parts might be detected as noise and get neglected in the process.
- A combination of both the short term energy and zero crossing rates work well compared to using them independently.
- While using the short term energy, if there is substantial power to the noise, then the noise might be detected as speech. However, those parts have high zero crossing rates which are detected in the zero crossing rate method and can be neglected.

4 Extracting Pitch using Cepstrum Method

The basis of this method is the fact that in the cepstra, the formant information is found in the beginning following which at the higher ranges, there is pitch information.

4.1 Implementation

- Decide upon the window size for which you want to calculate the pitch period. Here, we can assume that the pitch period remains constant over 5ms and calculate for that value i.e. every 5ms one pitch period.

- Firstly, we calculate the cepstra akin to question 1. However, we need to make sure that the windowed signal is not zero padded.

- Once we have the cepstra, we take the positive half of the cepstra and set a threshold beyond which we expect to find the pitch period.

- Before calculating the peak, we perform moving average on the signal to account for the noise in the cepstra. Then, we find the peaks of the cepstra beyond the threshold point and note its index.

- The pitch period can then be estimated using the following formula :

$$Pitch\ Freq = \frac{Sampling\ Frequency}{index\ w.r.t\ cepstra}$$

4.2 Plots

Below is the plot of pitch estimation of a sentence using the above stated method.

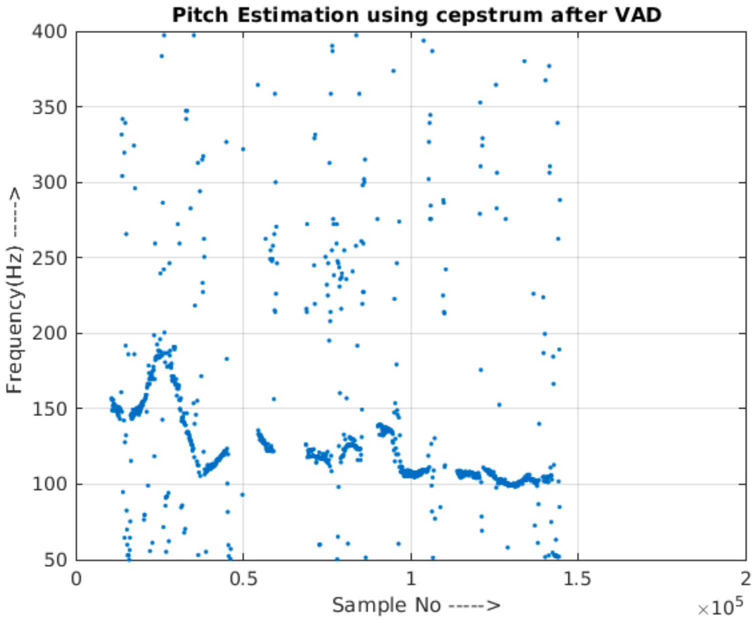

Figure 10: Plot of the pitch estimation using cepstrum

Below is the plot of pitch estimation of a sentence using the pitch in built function.

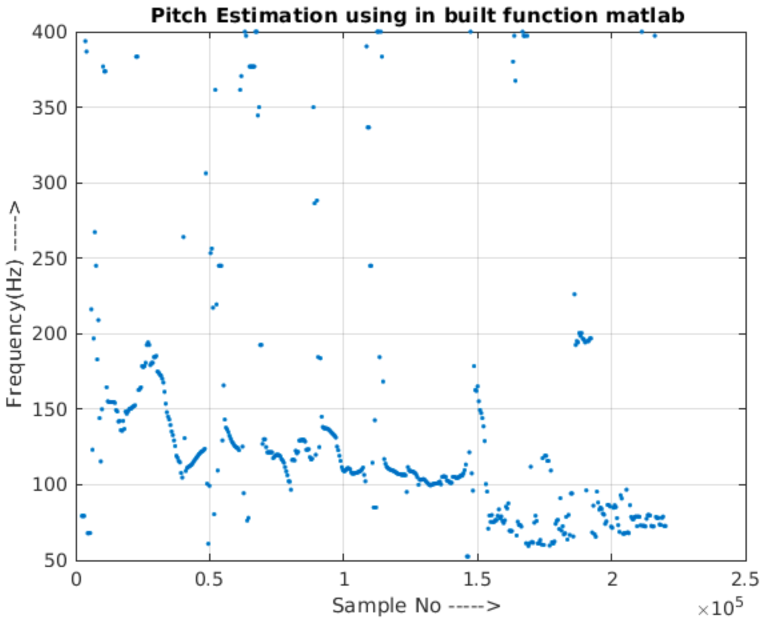

Figure 11: Plot of the pitch estimation using matlab function pitch

4.3 Observations

- From the plot of the pitch, we can see that both of them agree for major parts.
- Since the pitch is estimated after VAD, the parts where there was noise or silence, those parts have been set to pitch 0.

5 "We were away a year ago" - Analysis

5.1 Implementation

- The spectrogram of the signal was calculated using the spectrogram function present in Matlab. The overlayed formants were calculated using the code similar to question 2.
- The VAD and pitch estimation were done in manners similar to the Questions 3 and Question 4 respectively.

5.2 Plots

Below is the plot for Spectrogram overlayed with the formant plot for a male voice.

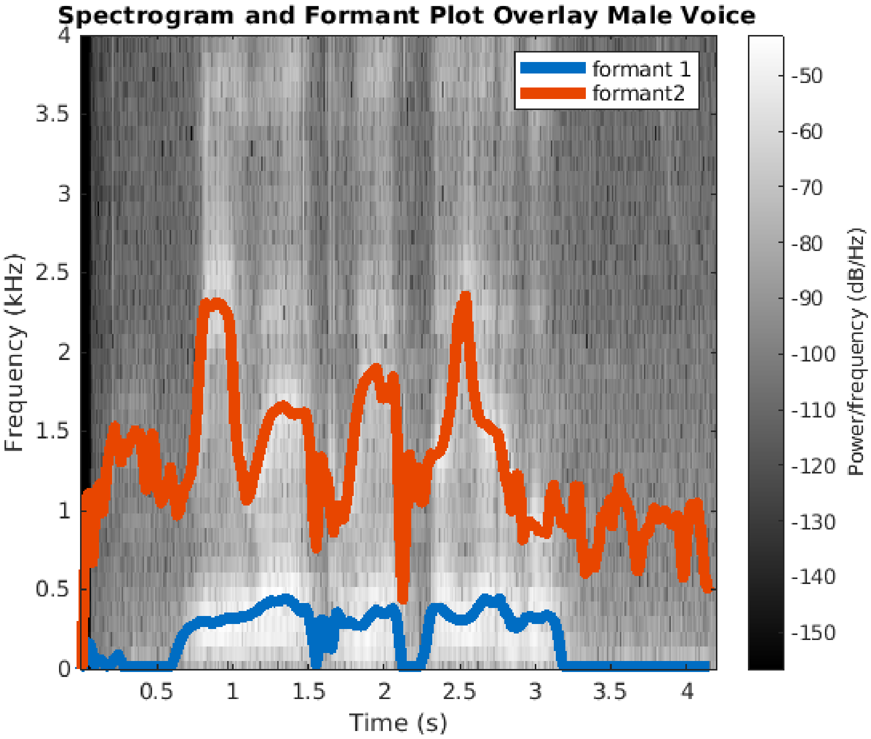

Figure 12: Spectrogram overlayed with the formant plot for male voice

Below is the plot for Spectrogram overlay with the pitch estimation for a male voice.

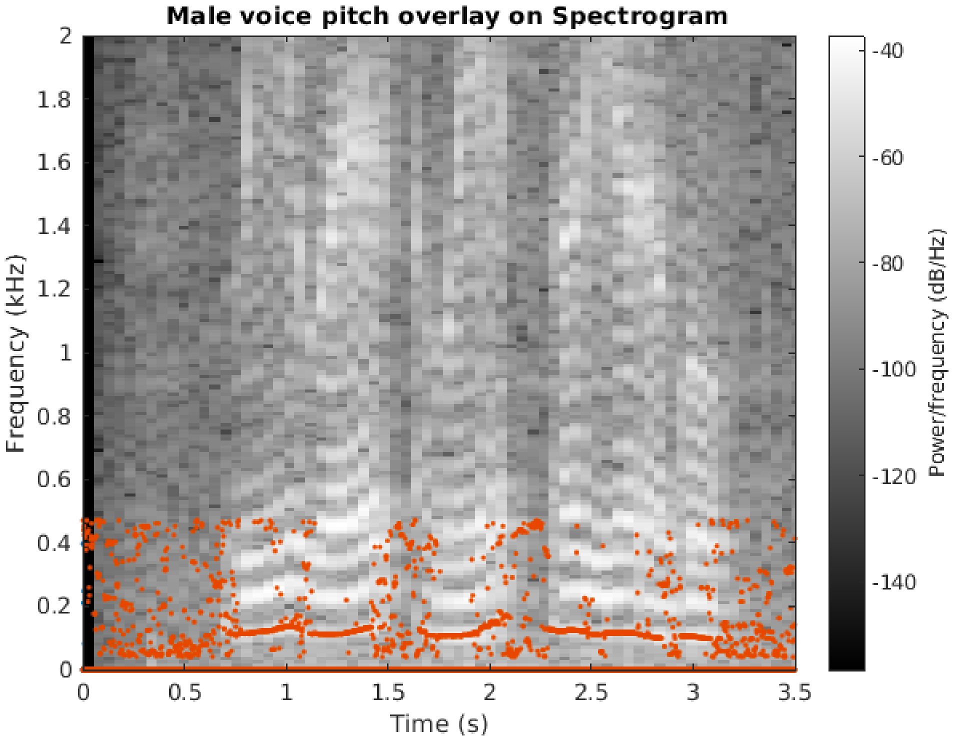

Figure 13: Spectrogram overlayed with the pitch plot male voice

Below is the plot with the Short Term Energy for a male voice.

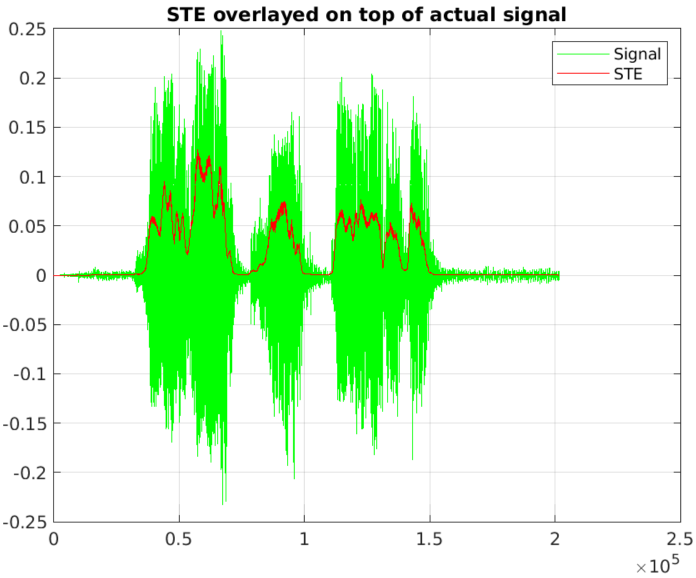

Figure 14: Short Term Energy for male voice

Below is the plot with the zero crossing rate for a male voice.

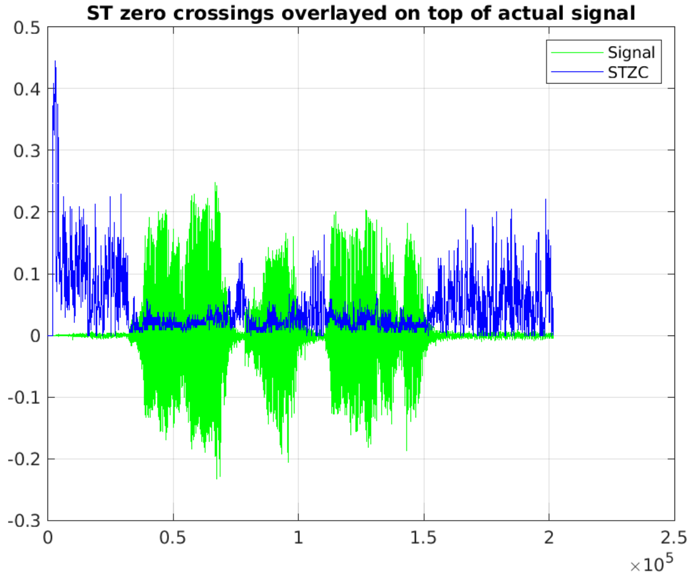

Figure 15: Zero Crossing Rate for male voice

Below is the plot with the voice activity detection for a male voice.

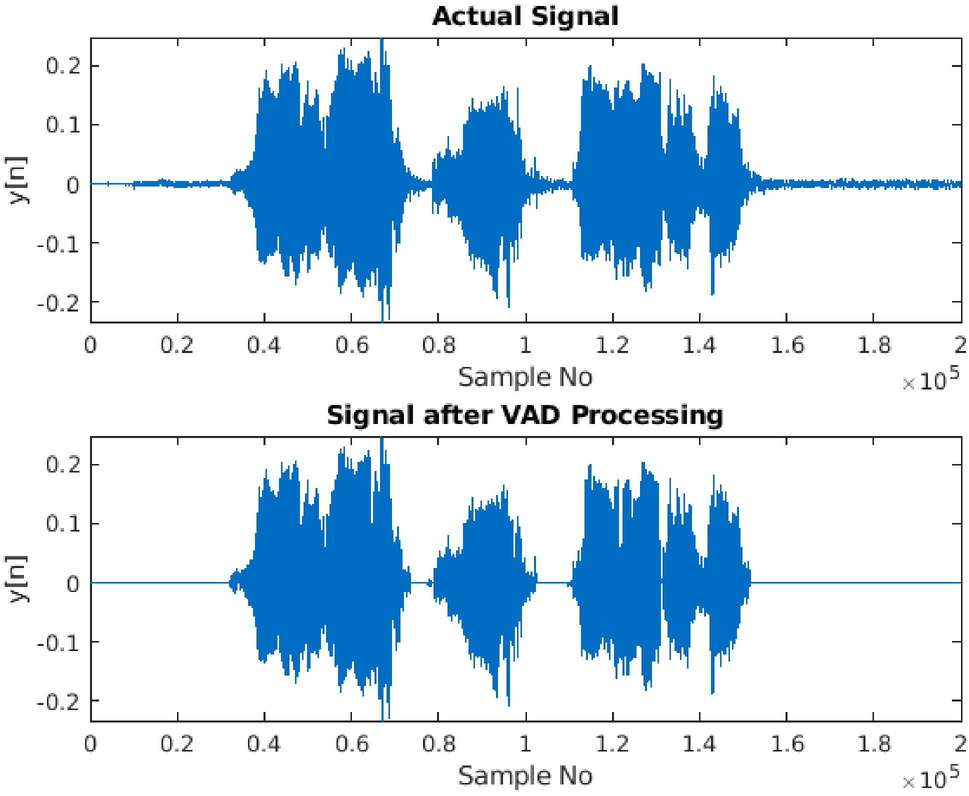

Figure 16: Voice Activity Detection for male voice

Below is the plot for Spectrogram overlayed with the formant plot for a female voice.

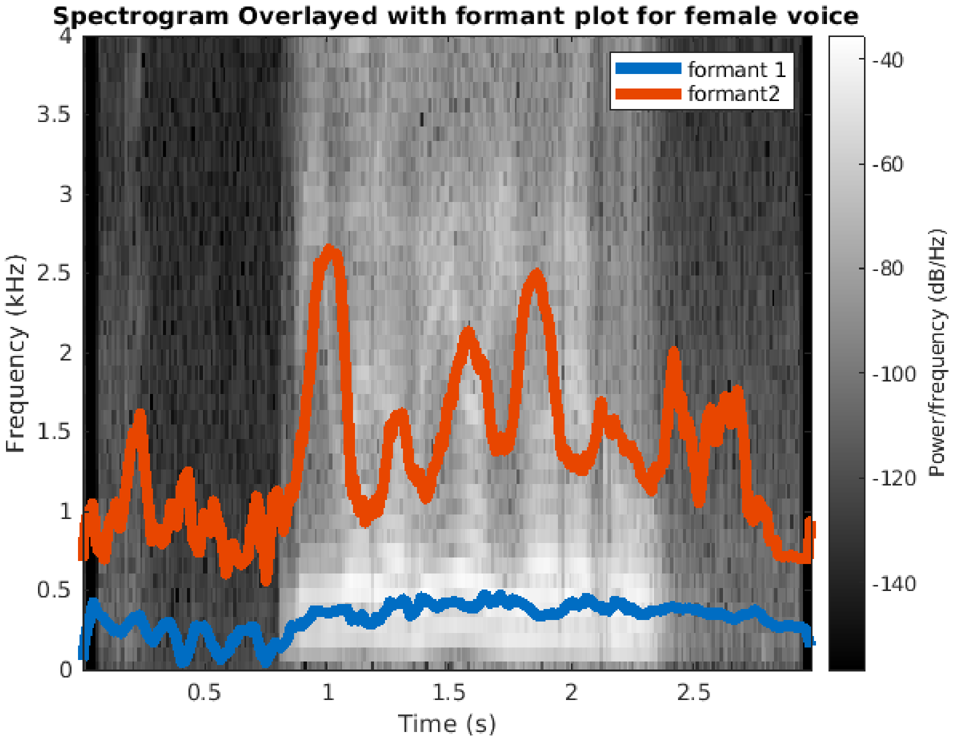

Figure 17: Spectrogram overlayed with the formant plot female voice

Below is the plot for Spectrogram overlay with the pitch estimation for a female voice.

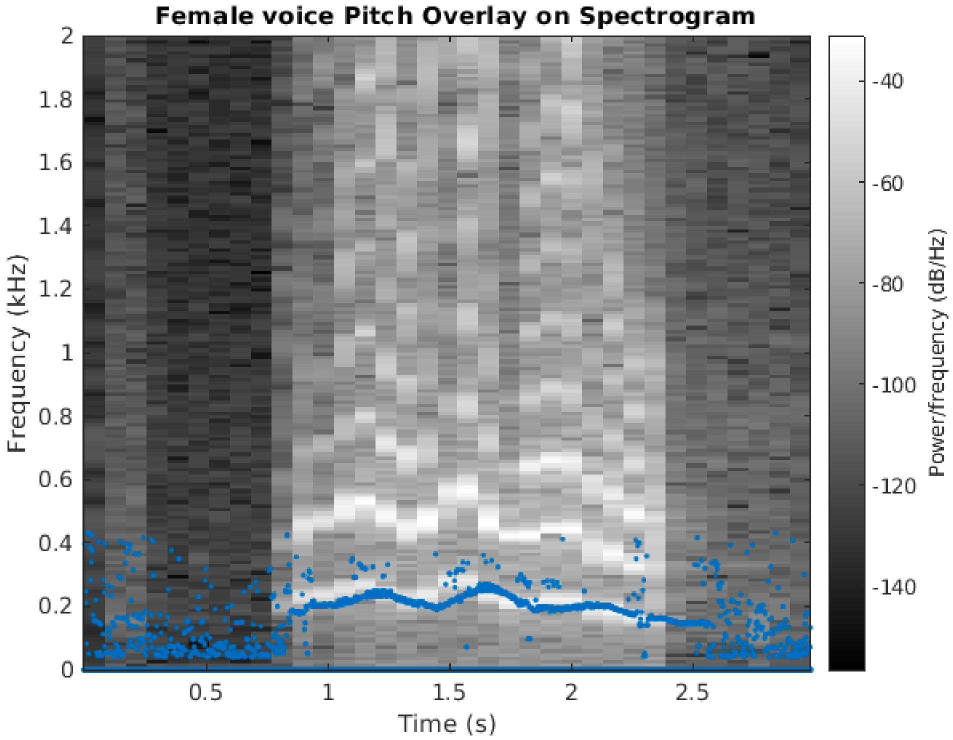

Figure 18: Spectrogram overlayed with the pitch plot female voice

Below is the plot with the Short Term Energy for a female voice.

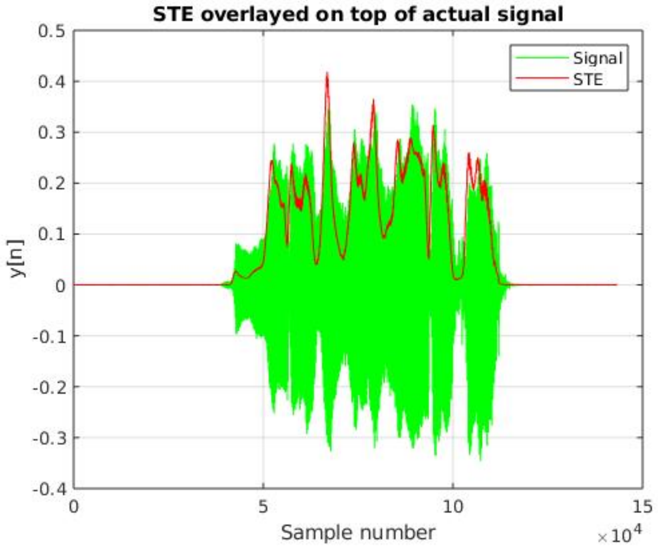

Figure 19: Short Term Energy for female voice

Below is the plot with the zero crossing rate for a female voice.

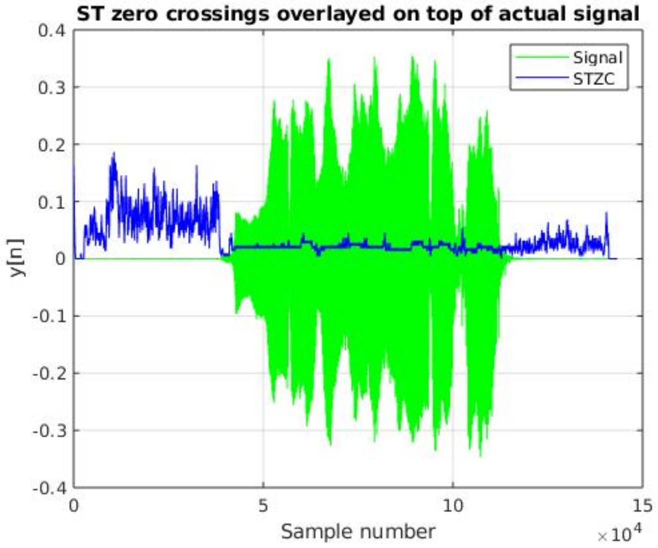

Figure 20: Zero Crossing Rate for female voice

Below is the plot with the voice activity detection for a female voice.

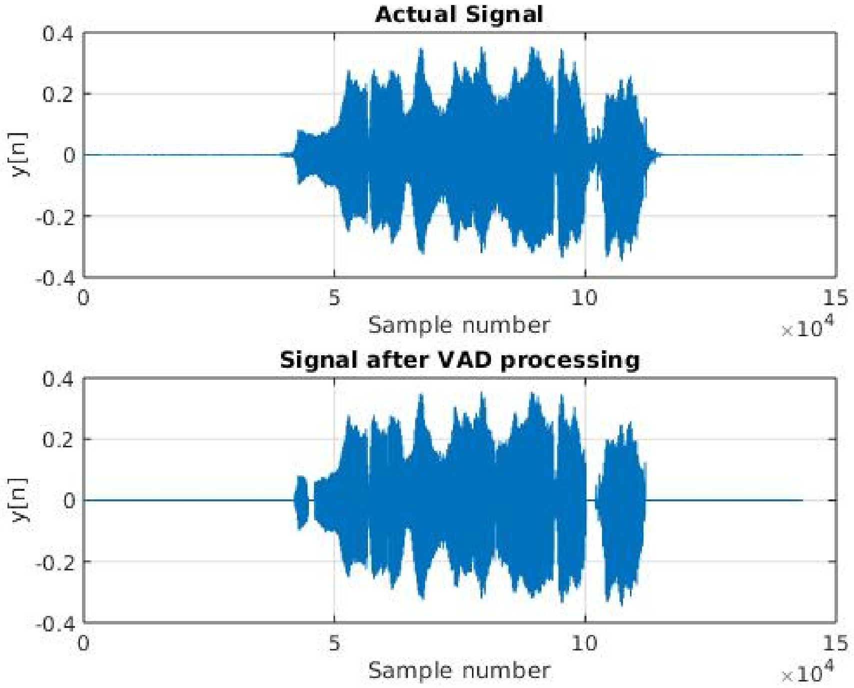

Figure 21: Voice Activity Detection for female voice

5.3 Observations

- From the overlayed plot of the formant and the spectrogram, we can observe that the formants predicted align mostly with the one visible from the spectrogram.

- The pitch plot coincides with one of the bright bands in the spectrogram.

- The VAD identifies the parts of the speech utterance which can be considered as noise and gives the meaningful parts alone as non zero.