Ryan Chau, Philip Kwan

EE114

Computer Assignment 3: Temporal Analysis of Speech

Abstract:

Short-time energy analysis:

For sampling rate = 8000Hz:

* 25ms window ->
* 3ms window ->

Q1 = time\_analysis(data, 'rectwin', 200, 1);

Q2 = time\_analysis(data, 'rectwin', 200, 2);

Q3 = time\_analysis(data, 'rectwin', 200, 3);

Qa = time\_analysis(a, 'rectwin', 200, 3);

Qsh = time\_analysis(sh, 'rectwin', 200, 3);

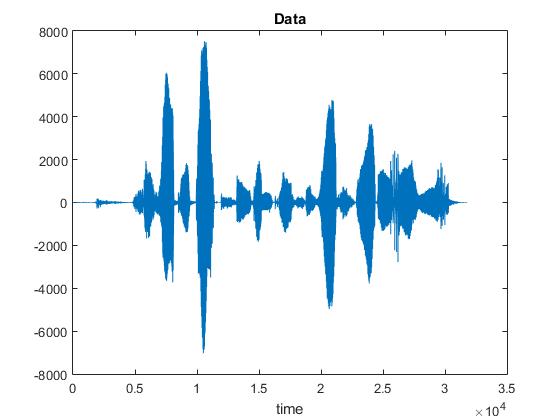


Figure 1: Data Signal – “She had your dark suit in greasy wash water all year”

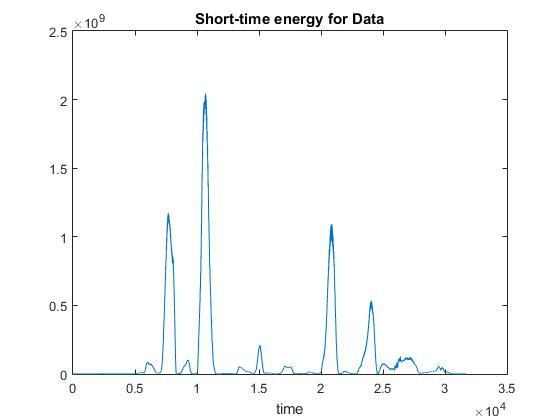


Figure 2: Q1 – Short-Time Energy

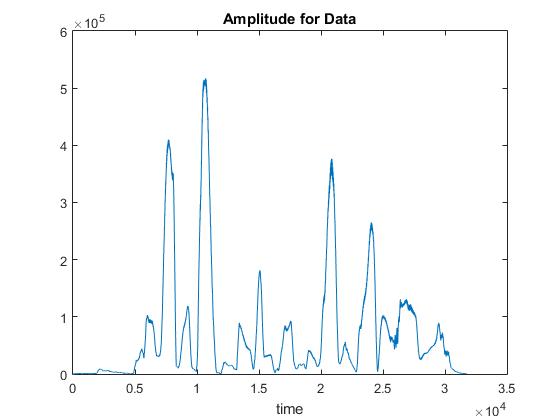


Figure 3: Q2 – Amplitude

1. This analysis with the code above is a narrowband analysis. We look at the plots of Q1 and Q2 to segment the sentence. We know that the peaks represent vowels because vowels have high energy. By looking at Q1, it’s easy to see that there are 12 peaks, which represent the number of syllables present in the sentence (each syllable contains a vowel sound). The amplitude plot is very similar to the energy plot in that there’s a dependence on the speech signal. Q2’s peaks are at the same area as Q1’s. When looking at the plot of Data, we see that the high swings correspond to the vowels as well.

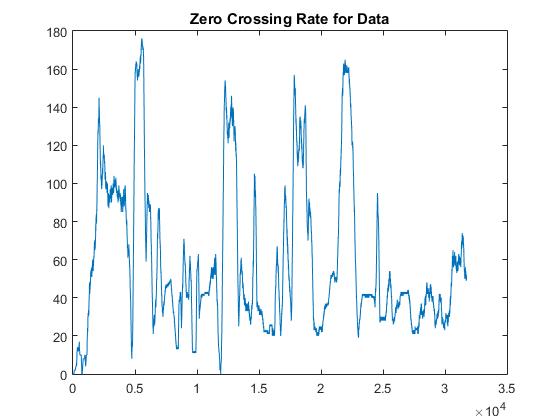


Figure 4: Q3 – Zero Crossing Rate

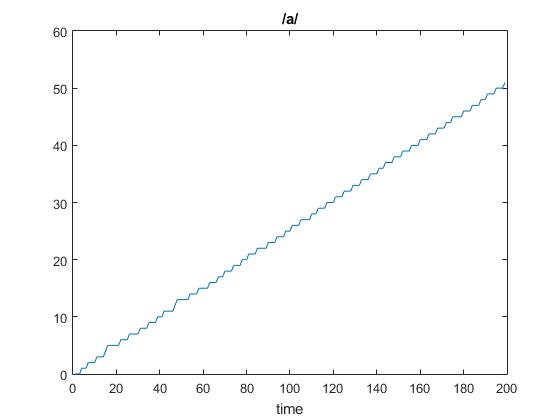


Figure 5: Qa - /a/ sound

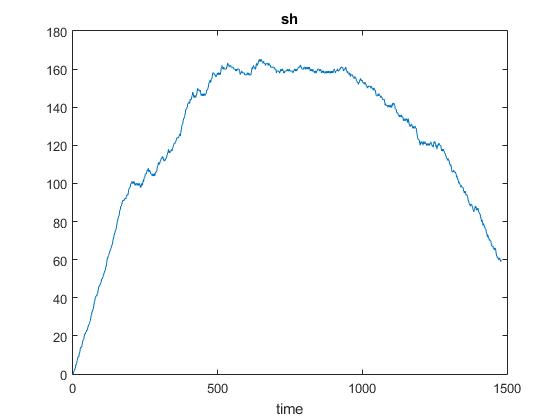


Figure 6: Qsh - sh sound

2. The “sh” sound is unvoiced, whereas the /a/ sound is voiced. Therefore, the ZCR threshold we obtain is: around 50 crossings per second for voiced and 150 crossings per second for unvoiced, maybe give or take 10 crossings/sec. The sentence is “She had your dark suit in greasy wash water all year.” The unvoiced sounds are: sh, s, sh, and t in order. There are 4 peaks that start after 0.5x104, which represent these unvoiced sounds. The rest are voiced, fricatives, vowels, or silence.

3. For this part, let’s use a wideband analysis.

Q1 = time\_analysis(data, 'rectwin', 200, 1);

Q2 = time\_analysis(data, 'rectwin', 200, 2);

Q3 = time\_analysis(data, 'rectwin', 200, 3);

Qa = time\_analysis(a, 'rectwin', 200, 3);

Qsh = time\_analysis(sh, 'rectwin', 200, 3);

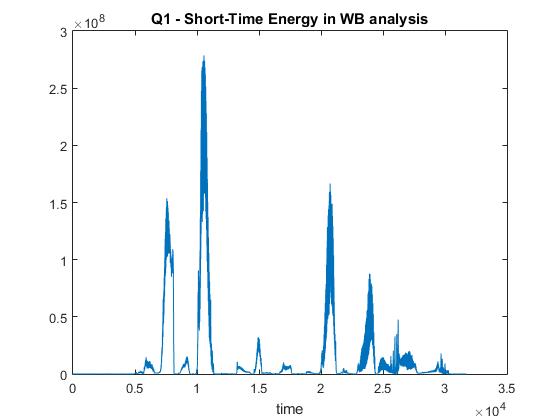


Figure 7: Short-Time Energy in WB analysis

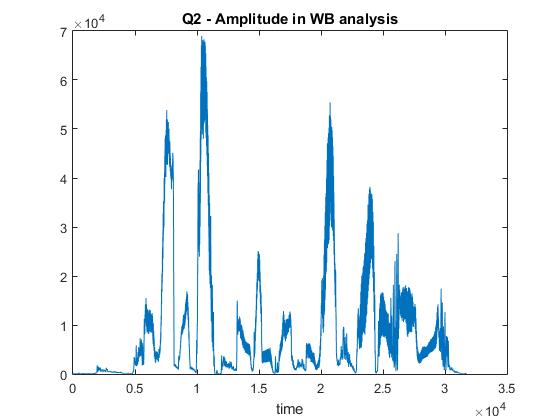


Figure 8: Amplitude in WB analysis

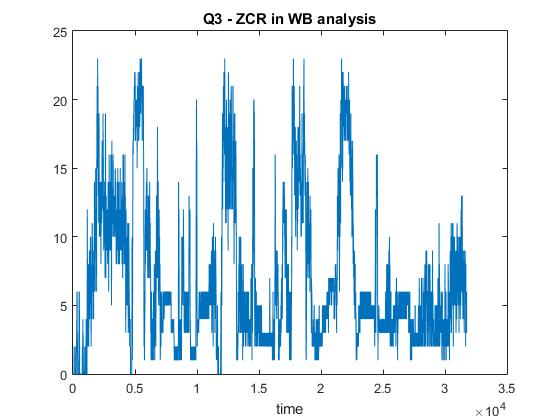


Figure 9: ZCR of Data in WB analysis

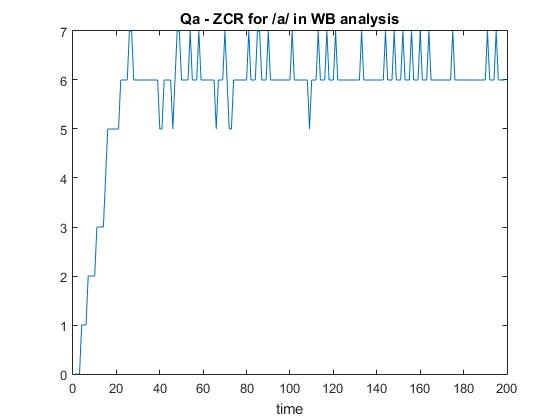


Figure 10: ZCR of /a/ in WB analysis

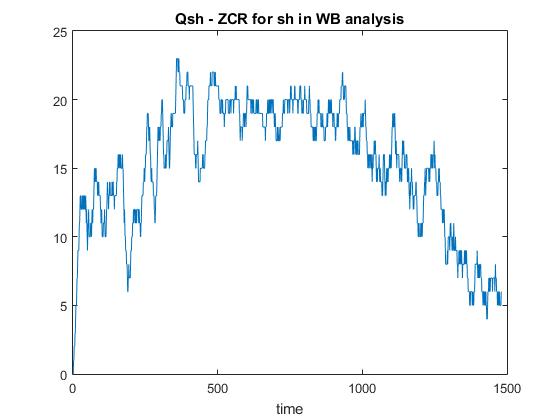


Figure 11: ZCR of sh in WB analysis

It looks like there’s a significant amount of spectral leakage, blurring the harmonics. It looks tougher to perform the analysis, especially for ZCR, because there just seems to be a huge blur between the two different thresholds in the Data plot, even if we define /a/ to be ~7 crossings/sec and sh to be ~20 crossings/sec.

4. In this part, we are looking at short-time autocorrelation and short-time average magnitude difference for the /a/ and sh sounds.

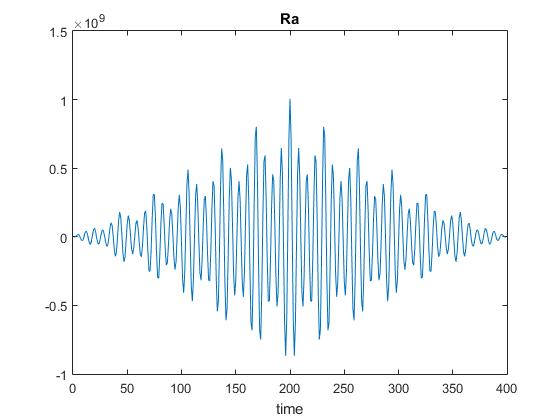


Figure 12: ST Autocorrelation of /a/

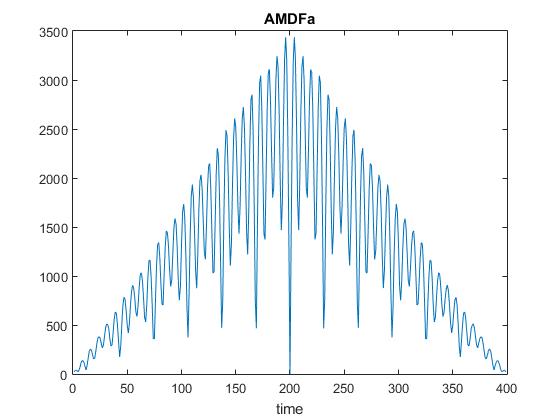


Figure 13: ST AMDF of /a/

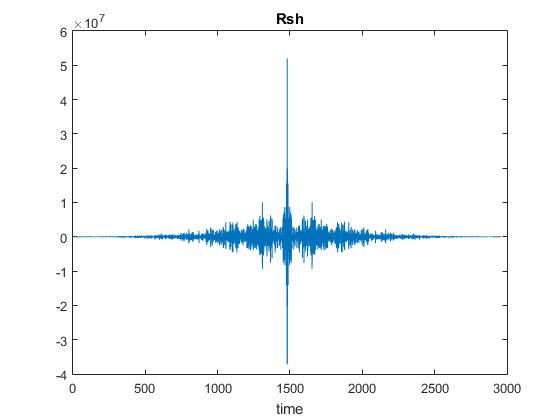


Figure 14: ST Autocorrelation of sh

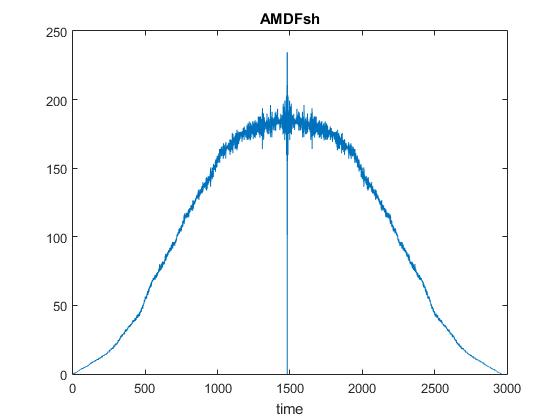


Figure 15: ST Autocorrelation of /a/

Looking at the two plots for sh, we can see that the relative swing with respect to time stays about the same, although the autocorrelation is symmetric with respect to the y-axis, whereas the AMDF is pushed up as we approach the middle from either side. Looks like the same goes for /a/.

**The Effect of Window Size:**

About 70 is when the harmonics become unresolved – where the lobes seem to merge over with the adjacent ones – for the female voice. About 140 is when the harmonics become unresolved for the male voice, although the male voice is much more sensitive to the window sizing and have many more lobes than the female voice.

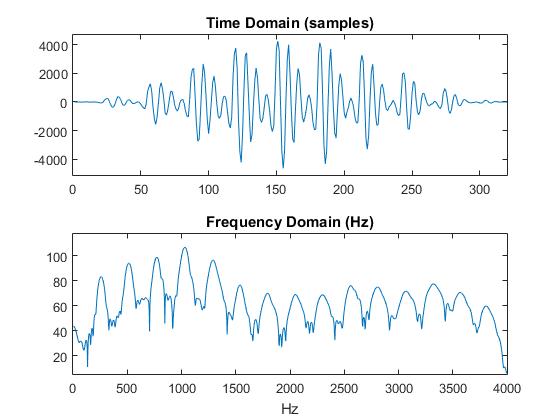


Figure 16: Feale\_a voice – 320 samples

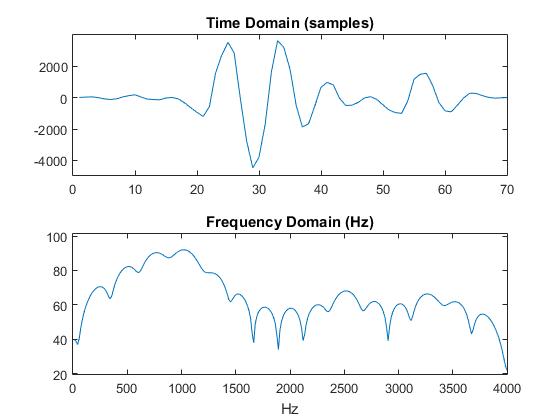


Figure 17: Feale\_a voice – 70 samples

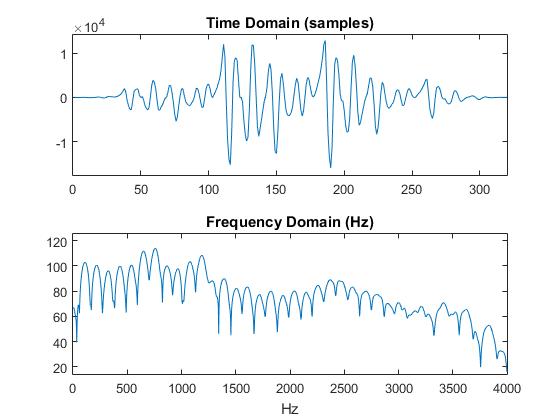


Figure 18: Male\_a voice – 320 samples

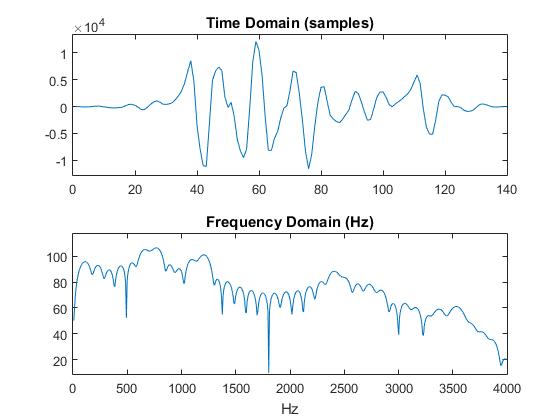


Figure 19: Male\_a voice – 140 samples