Aristotle University of Thessaloniki
Department of Electrical and Computer Engineering

Advanced Signal Processing

"Cepstral Analysis & Synthesis of Vowels / Speech Processing"

4th Assignment – Summer Semester 2020/2021

Kavelidis Frantzis Dimitrios – 9351

29/5/2021

Contact: kavelids@ece.auth.gr

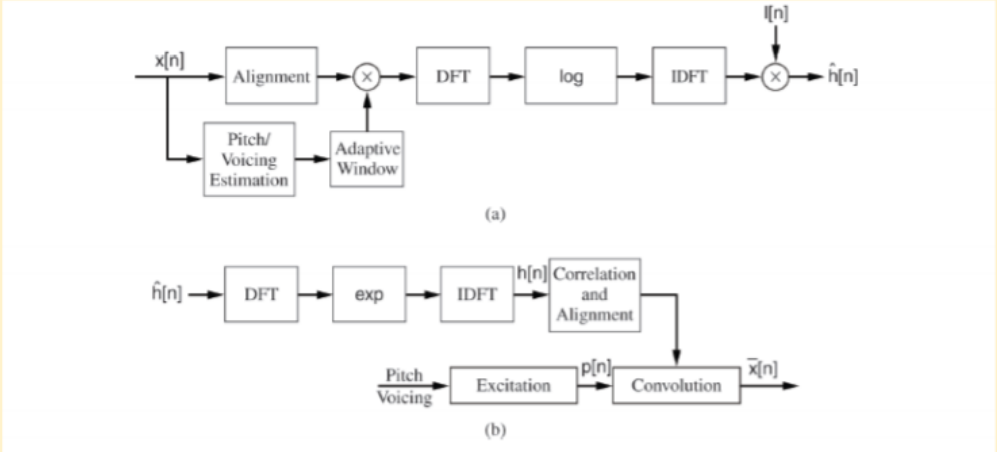
Description - Assignment 4:

1. Acquire voice samples. In this part, please either record or find at least 10 voice samples of a male and female individual making the five vowel sounds – "a", "e", "i", "o", "u". If you are going to record them yourself or using a friend, then exaggerate the sounds a little and keep your voice extended for a while. Please make note of the conditions used to obtain the voice samples (e.g., what smartphone, what type of speaker – built-in or microphone, using which software program, or where or from whom the files were obtained). You should have 10 files in the end. If you want to be keen and impressive, you can get more than one male and female voice to obtain a better understanding of differences between both signals in the cepstral domain. Have fun with this.

2. Compute the cepstrum of each voice signal and discuss any difference qualitatively and quantitatively amongst male and female voices in general and amongst the different vowel sounds. This is an important component of the project, so please be creative and as comprehensive as possible. Your report should provide figures with original time-domain signals as well as cepstrum signals. Female voices should generally have more peaks than male voices in the cepstrum domain. You should discuss why you think this would be the case.

3. Lifter the cepstrum domain signals. Design a window (length is an important design parameter and you should discuss how and what you select – it can be the same or different for each speech sample depending on what you would like to experiment with) to remove the transfer function dependency. Then, compute the time domain signal of the corresponding windowed result to obtain the deconvolved signal. Plot the deconvolved result. Is there anything you can say about the signal and its difference from the original time domain recorded sample? Again, your discussion is an important part of the report.

4. Try to synthesize back the voiced signals as follows and comment on your results.



Analysis:

1. The acquired samples were recorded in a home studio. A total of 10 voice samples were recorded from a male and a female.
The equipment used was:

a. Microphone : **Behringer B1**
b. Loop-station used as audio interface : **Boss RC-505**

Also, to get the samples in the computer, the software used was:

a. Digital Audio Workstation : **Ableton Live 11**
b. Setup Loop-station editor : **RC-505 Editor**

Samples were saved in a folder called 'Samples'. Female samples were saved in the folder 'Samples\Female1' and male samples were saved in the folder 'Samples\Male1'. Each of the sample inside each folder has the name of the vowel that it holds, i.e. 'A.wav'.

These settings were chosen this way to create an easily comprehendible code and to have the compatibility to run it for future recordings of different people that would be saved in similar structured folders, i.e. 'Female2'.

2. The plots of a segment of the samples for each vowel is shown below for female and male samples respectively:

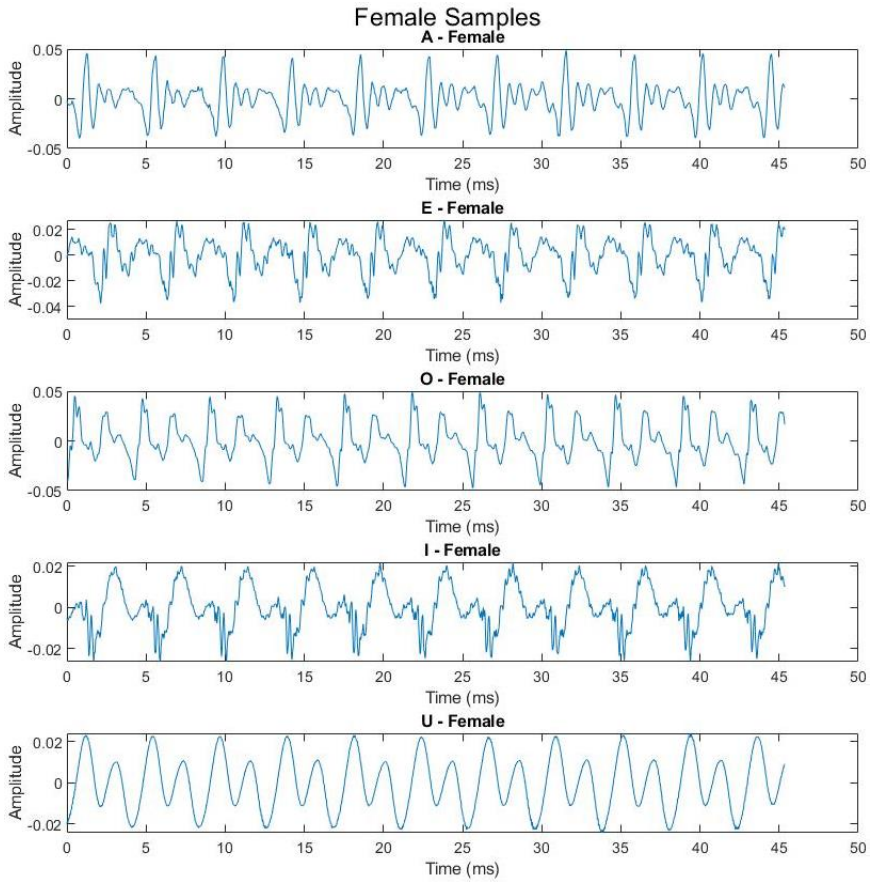


Fig. 1 – Vowel Segments of Female

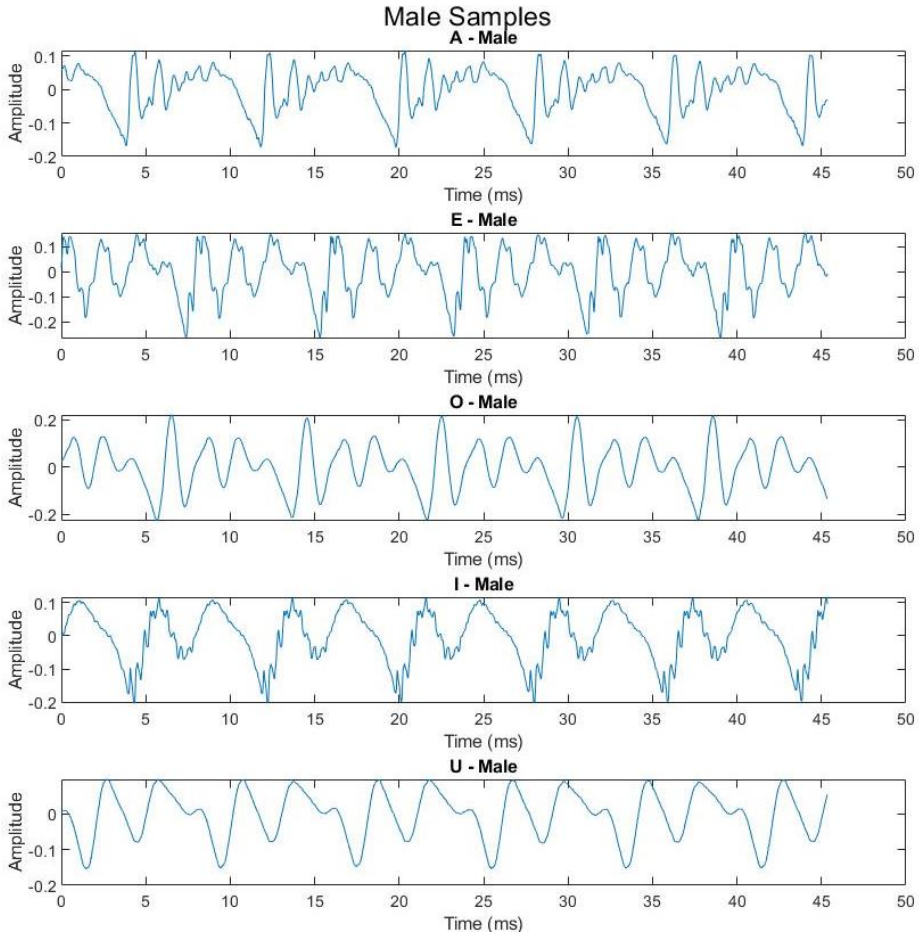


Fig. 2 – Vowel Segments of Male

Even though the speech signal is considered non-stationary, it is easy to see that the signal of a vowel in a short period like the one chosen (~45 ms), seems to have stationary properties **[1]**. The human speech production system according to the source-filter model consists of a source creating a signal that excites the vocal tract, generating the speech signal. Thus, for the short period of time mentioned above, the speech signal $\boldsymbol{s(n)}$ can be considered as the linear convolution of the source excitation signal $\boldsymbol{e(n)}$ with the function describing the vocal tract $\boldsymbol{v(n)}$ which is the impulse response of the vowel production system.

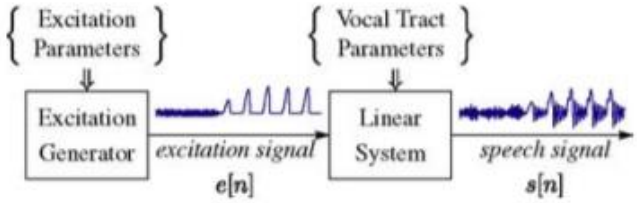


Fig. 3 – Source / System model for a speech signal

<u>Vowel Production (shortly)</u>:

In vowel production, air is forced from the lungs by contraction of the muscles around the lung cavity. The sound source that excites the vocal tract can be either a ***Voiced (periodic source)*** or ***Unvoiced (aperiodic source)***. The vowels are signals coming from a voiced source, explaining the periodicity exists in the plots above. Air flows through the vocal cords, which are

two masses of flesh, causing periodic vibration of the cords whose rate gives the ***Pitch*** of the sound. The ***Pitch*** is the fundamental frequency of the vocal cords vibration, followed by 4-5 ***Formants*** at higher frequencies. The resulting periodic puffs of air act as an excitation input, or source, to the vocal tract.
The typical pitch of the speech is ~85-155 Hz for male and ~165-255 Hz for female. Thus the fundamental frequency for female is higher resulting to shorter fundamental period as it can be seen in Fig. 1 and 2.
In the frequency domain, the convolution $\boldsymbol{s(n) = v(n) * e(n)}$ is transformed to multiplication:

$$\boldsymbol{S(\omega) = V(\omega) \cdot E(\omega),} \qquad \mathbf{(1)}$$

In real systems though, usually a window is applied:

$$\boldsymbol{s(n) = (v(n) * e(n))\, w(n),} \qquad \mathbf{(2)}$$

If we assume that w(n) is slowly varying over the effective length of **v(n),** then $\boldsymbol{s(n)}$ can be written as:

$$\boldsymbol{s(n) \approx v(n) * (e(n)\, w(n)),} \qquad \mathbf{(3)}$$

and therefore:

$$\boldsymbol{S(\omega) = V(\omega) \cdot E_W(\omega),} \qquad \mathbf{(4)}$$

where $\boldsymbol{V(\omega)}$ and $\boldsymbol{E_W(\omega)}$ are the DTFT of $\boldsymbol{v(n)}$ and $\boldsymbol{e(n)\, w(n)}$ respectively.
To extract an estimation of the impulse function of the linear system, a more sophisticated analysis was proposed by Oppenheim (1965), defining the complex cepstrum in his development of homomorphic system theory. The use of cepstrum yields to

$$\boldsymbol{\hat{s}(n) \approx \hat{v}(n) + \hat{e}(n),} \qquad \mathbf{(5)}$$

where $\boldsymbol{\hat{v}(n) = \mathcal{F}^{-1}[log\, V(\omega)]}$ and $\boldsymbol{\hat{e}(n) = \mathcal{F}^{-1}[log\, E_W(\omega)]}$ **.**

In this domain, it is expected for the cepstrum of Excitation signal to have higher values and the cepstrum of vocal tract/impulse response to have lower values. Thus, if there is no overlap, an appropriate filter can be used to separate them, taking the inverse of their cepstrum and retrieve the impulse response (*Homomorphic Filtering – Liftering*).

There are many plots generated from our code because of the number of vowels, so the following are just some of them, chosen for showing both good and bad results.

Differences between male and female vowels:

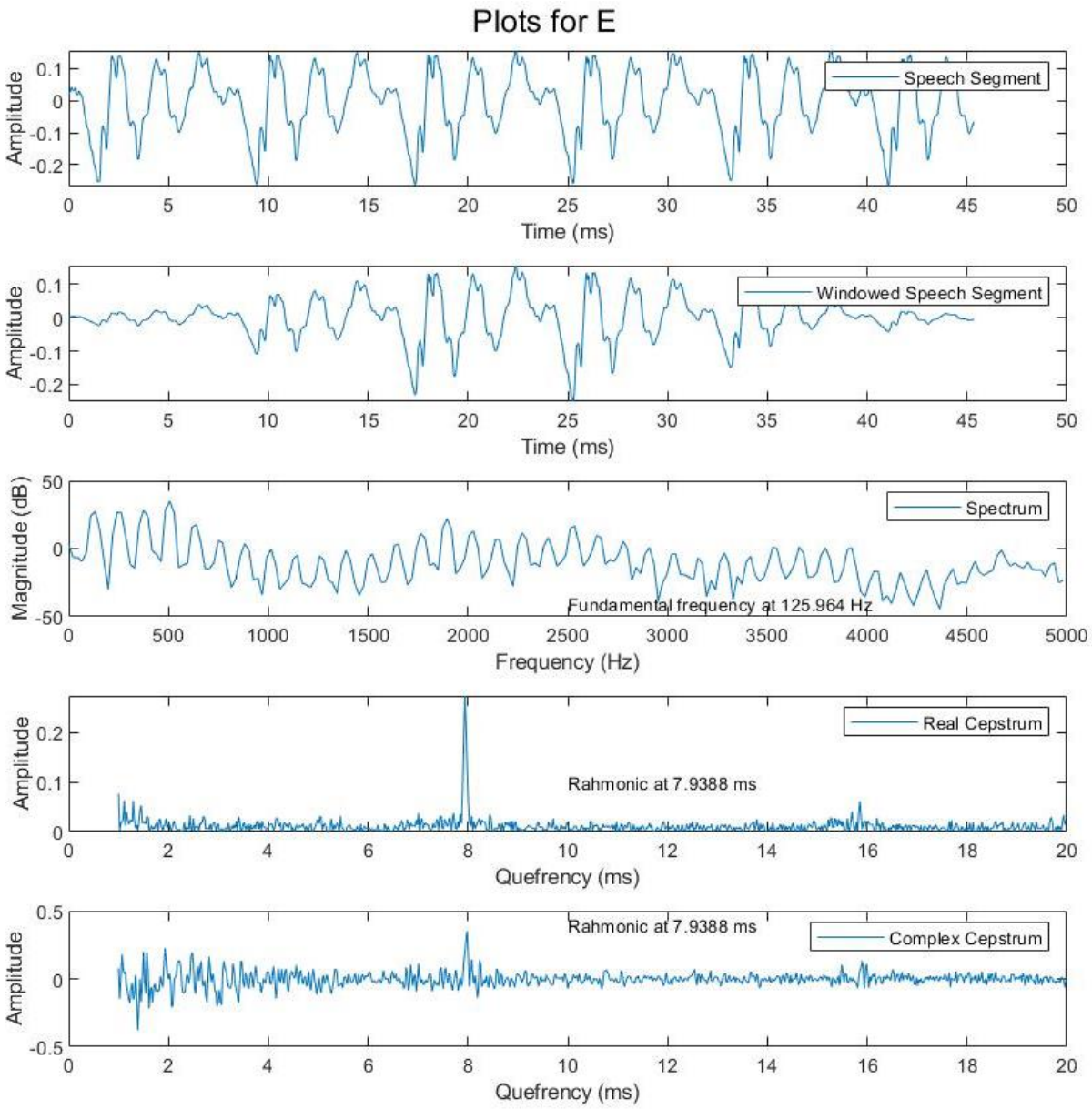


Fig. 4 – Analysis of vowel 'E' / Male

Note that Fourier spectral density of voiced signals has a form of an envelope that modulates a periodic function of frequency. The envelope is described by the impulse response, which is where the interest of estimation is.

The so called "rahmonic" is the peak in the cepstrum, indicating the pitch of the speaker. Now the comparison with the female 'E' vowel:

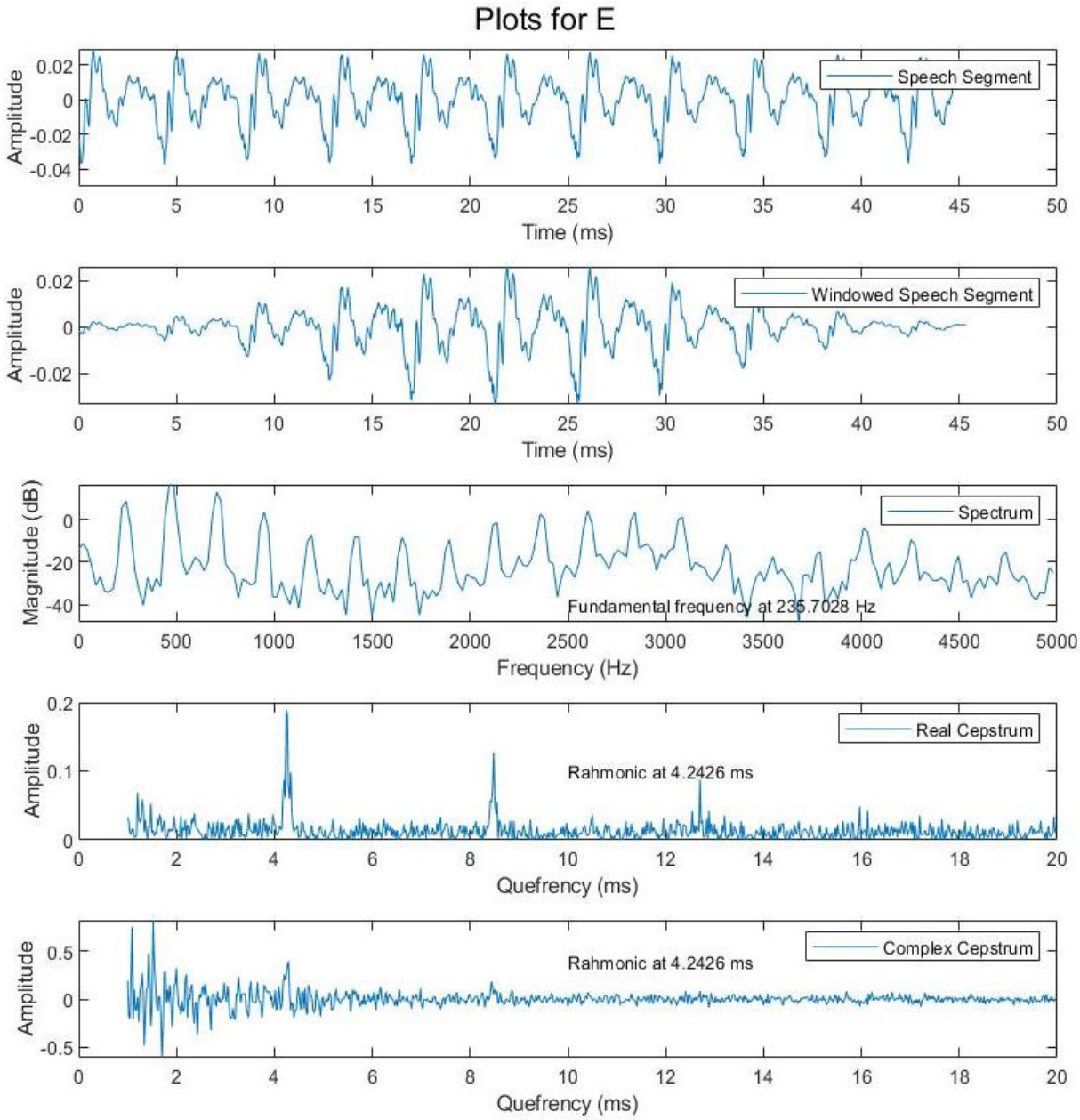


Fig. 5 – Analysis of vowel 'E' / Female / Same time as the corresponding male segment for comparison

We can see that there are more peaks present in the cepstrum of a female vowel. Also, it seems that the male fundamental peaks are of higher amplitude, but the rest of the peaks have a higher amplitude in female vowel, indicating stronger/louder echo/harmonics. As mentioned, the fundamental frequency is higher, yielding to a rahmonic at a lower quefrency.

More plots:

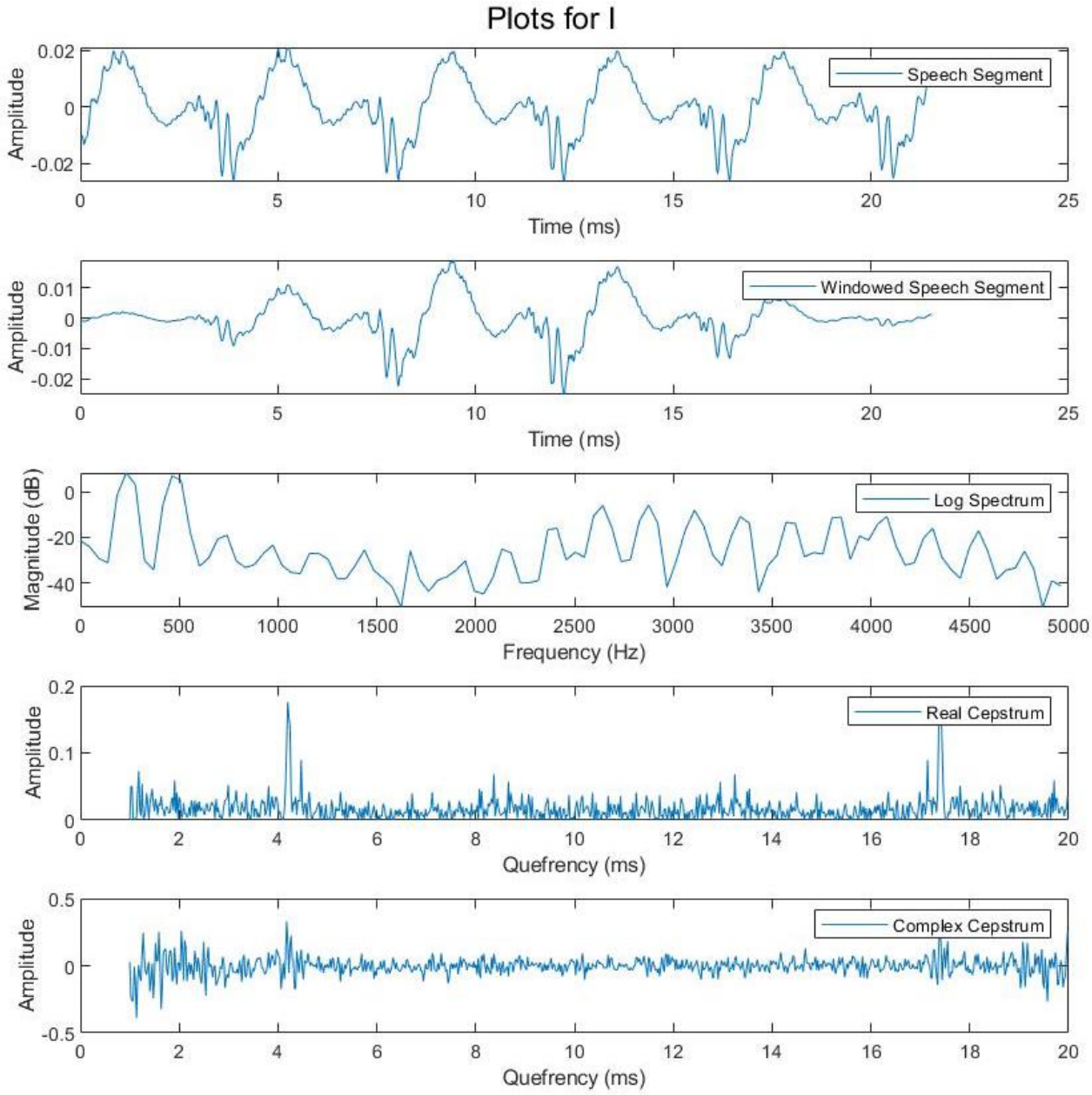


Fig. 6 – Analysis of vowel 'I' / Female

One more difference is that female vowels are "cleaner", meaning that they approach a sinusoid sum form, while male vowels are "raspier", bringing this kind of interference in the sinusoids. This also holds among the different vowels. For example, we see that 'U' is pretty clean, almost representing the sound of a sinusoid, while 'E' and 'I' are not. Also, in every vowel, male or female, there is a fall and rise from peak to peak on each period.

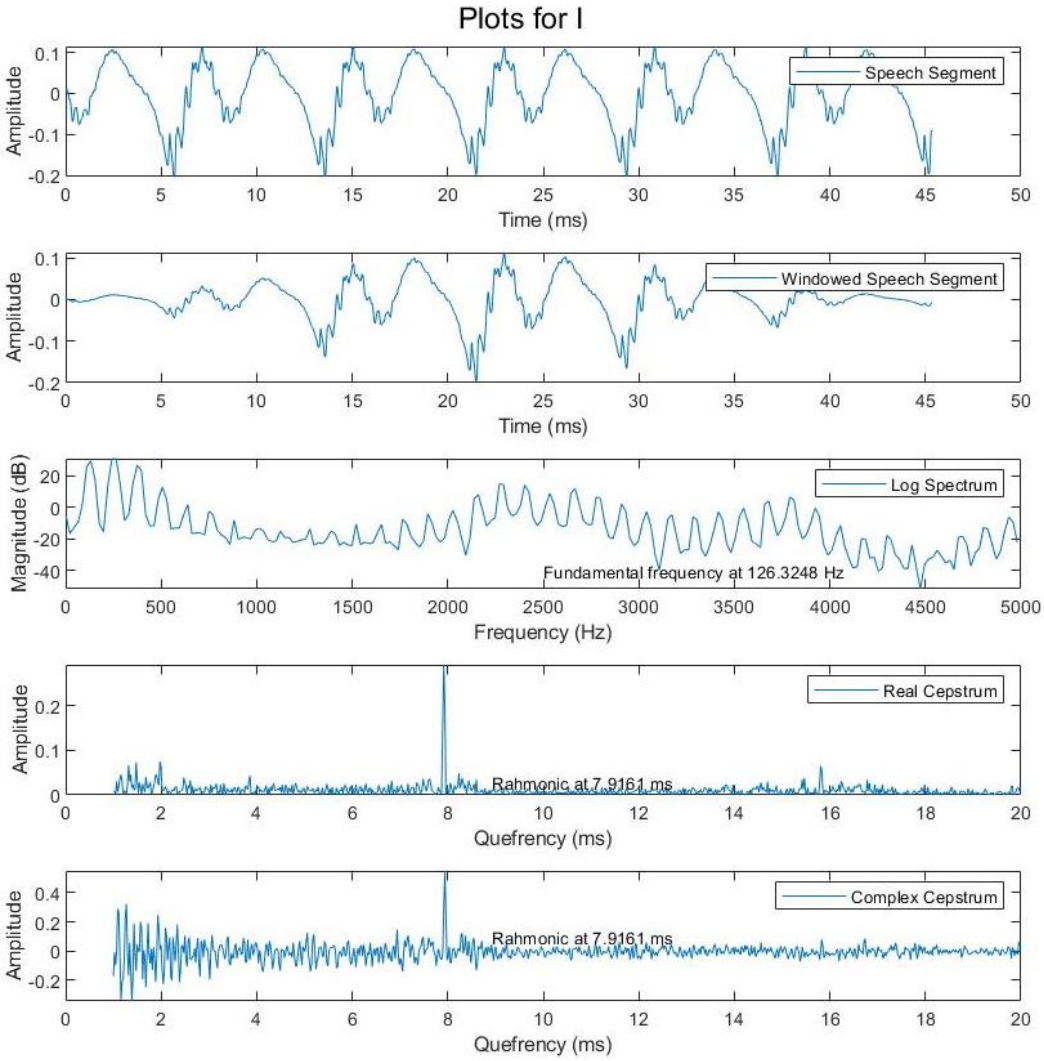


Fig. 7 – Analysis of vowel 'I' / Male

Also, some of the Mixed Phase Complex Cepstrum plots:

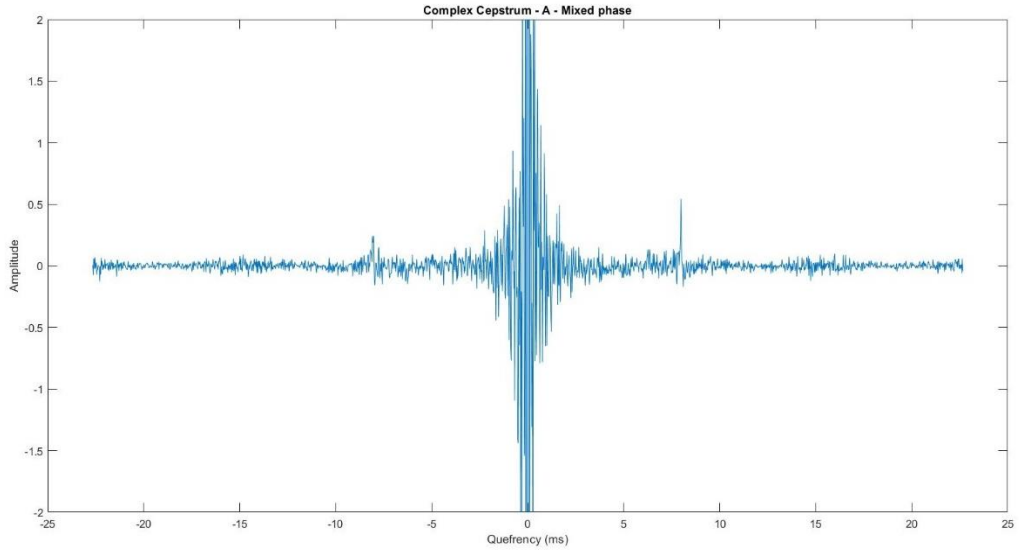


Fig. 8 – Complex Cepstrum / Mixed Phase/ 'A' / Male

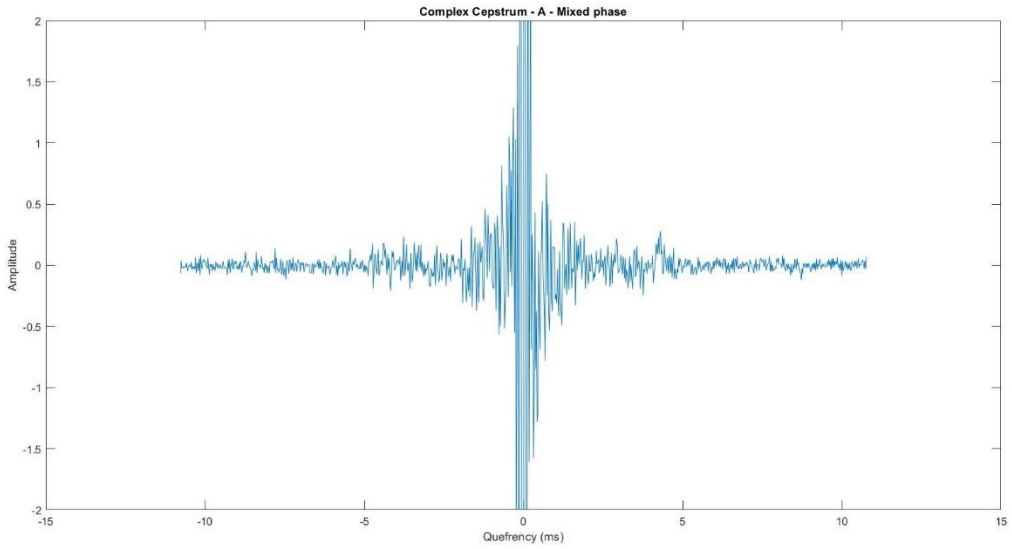


Fig. 9 – Complex Cepstrum / Mixed Phase/ 'A' / Female

3-4) For each of the vowels we:

i) Estimate the pitch period,
ii) Take a segment of the signal with length of about 2-5 pitch periods,
iii) Multiply the signal with a hamming window,
iv) Compute the cepstrum,
v) Liftering with low cut and a high cut filter to separate the impulse response from the source signal. The filter was not adaptive but rather the cut quefrency was estimated by trial and error / evaluation of the reconstructed sample.
vi) Synthesis of the new sample by convolute the estimated impulse response and the excitation signal of the full-time original signal.

Samples of the original vowel signals used for windowing:

Data Points / Gender	A	E	I	O	U
Female	**950**	**930**	**950**	**950**	**940**
Male	**1200**	**1200**	**1200**	**1200**	**1200**

Low-cut filter based on "cut -off" sample:

Vowel / Gender	A	E	I	O	U
Female	**40**	**54**	**50**	**58**	**61**
Male	**42**	**90**	**41**	**65**	**65**

Lowpass window used: $w(n) = \begin{cases} 1, & |n| < n_c \\ 0, & |n| > n_c \end{cases}$, where n_c is the "cut-off" sample.

Some of the results:

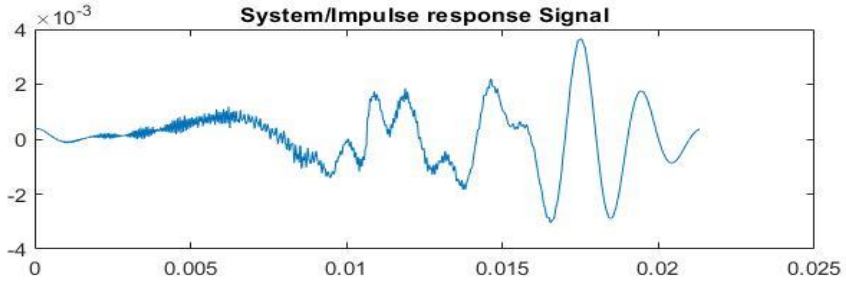


Fig. 10 – Impulse Response Signal / 'U' / Female

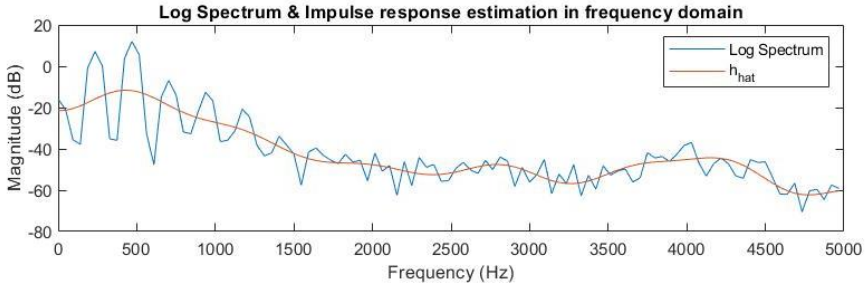


Fig. 11 – Impulse Response Envelope / Log Spectrum / 'U' / Female

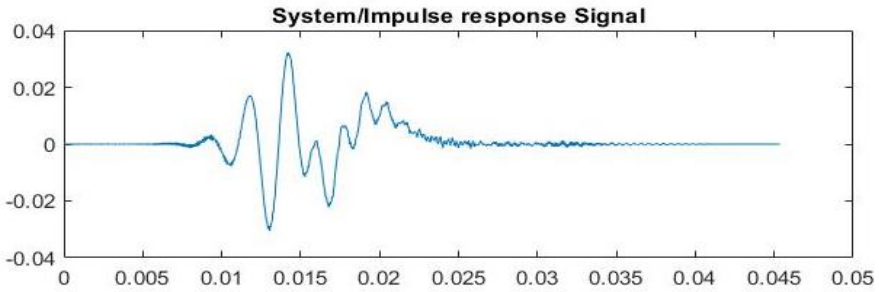


Fig. 12 – Impulse Response Signal / 'U' / Male

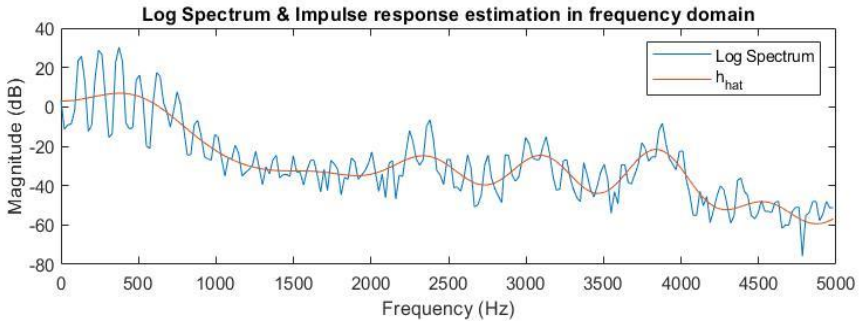


Fig. 13 – Impulse Response Envelope/ Log Spectrum / 'U' / Male

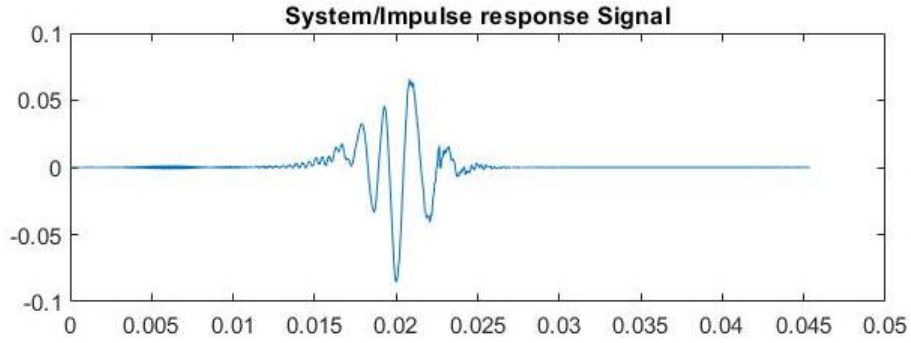


Fig. 14 – Impulse Response Signal / 'O' / Male

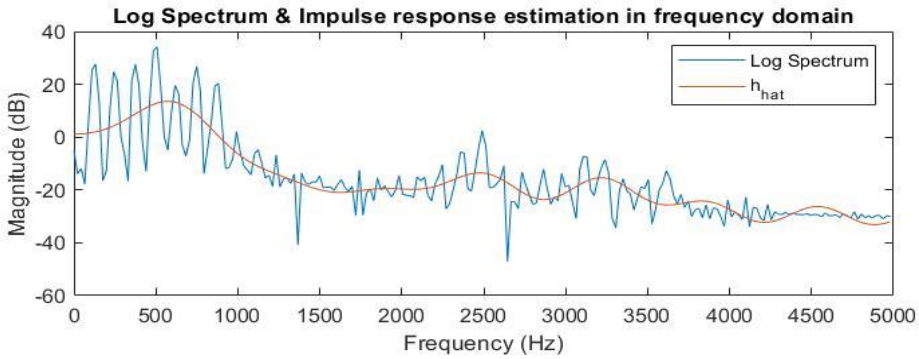


Fig. 15 – Impulse Response Envelope/ Log Spectrum / 'U' / Male

However, with bad choice of window length, the separation of the two signals may not be entirely possible because of overlap. Also, a bad choice of filter can lead to odd results and to "muffled", "robotic", "buzzed" or even sweep sounds.
In Fig. 14 we see a signal that is similar to a fraction of the windowed segment before the cepstral analysis. That is because the impulse response describes the system, meaning that when convoluted with an excitation signal that describes the pitch, it is supposed to give a new signal describing the same vowel, just in different tone, while giving the signal the form of the impulse response.

Conclusions:

Generally, the impulse response on female vowels was more difficult to be estimated. This is because the shorter pitch period required less samples to work on (because as mentioned, window must be applied in 2-5 periods), thus the changes among the different cut quefrencies were larger and therefore more it was more challenging to find an optimal one.

After choosing right length for window, the cut quefrency parameters were chosen based on

i) Hearing the reconstructed sample.
ii) Evaluate the plot of the Impulse response signal.
iii) Evaluate the form of the envelope on the log Spectrum domain.

The rest of the plots can be shown when running the Main_4.m file of the code.

Future Work/Goals:

- Creation of adaptive filter to automate the separation of the signals.
- Better estimation impulse response of the current samples.
- Enrich the database of samples with more people for a thorough understanding.
- Implementation of a Cepstrogram using STFT.

References:

[1] G. Ravindran, S. Shenbagadevi, V. Salai Selvam, "Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech", JBiSE, January 2010