CS578 Speech Processing
Laboratory 0: Part 2
Frequency Domain Speech Processing

George Manos
csd4333@csd.uoc.gr

Alexandros Angelakis
csd4334@csd.uoc.gr

13 December 2022

1 Implementation

Using Matlab, we created a pitch estimation system that analyzes with 2 different ways the input speech signal (for the first part, that will be a purely voiced synthetic speech), the short-time auto-correlation method and the short-time Fourier transform method, as described by the project's description. Again, once you run the code you select a folder that contains the .wav files you want to test and applies the method iteratively to all of them, creating a figure of 2 plots where the upper represents the pitch estimation and the lower one represents the input speech signal. Also, it prints the classification result for age+gender discrimination.

2 Method Comparison

First of all, let's talk about the FFT peak picking method. What we simply did was to take the FFT of each frame of the waveform and then founnd all the peaks that are in the range [70hz,500hz] (frequencies). In this method, the fundamental frequency/voice pitch f_0 corresponds to the first positive peak of the frequency spectrum of that frame*.

The auto correlation method is pretty similar with the FFT method, but in the time domain of the frame. More specifically, after we find the auto correlation of the frame, we consider only the peaks that are in the range $[\frac{1}{500}, \frac{1}{70}]$ (seconds) and we choose the time index of the maximum peak within that range. Finally, $f_0 = \frac{1}{t_0}$, where t_0 the aforementioned time index. As described by the project's description, this peak corresponds to the voice pitch as it is the point where the signal is most similar to itself, hence closer to the period P of that frame.

We tested both methods in all of our given voiced speech files, we will see that the FFT peak picking method performs better. Let's take a look at some examples and comment on them.

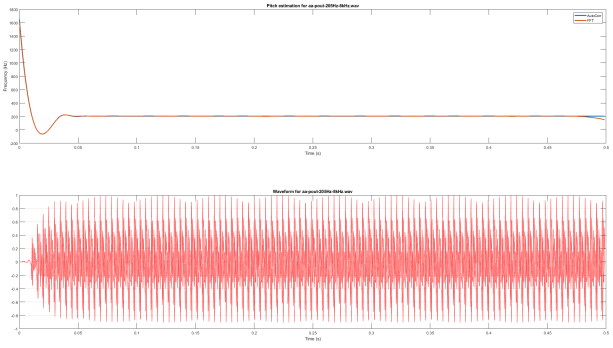

Figure 1: /aa/ speech signal with $f_0 = 205$ Hz. Classified as Adult Female.

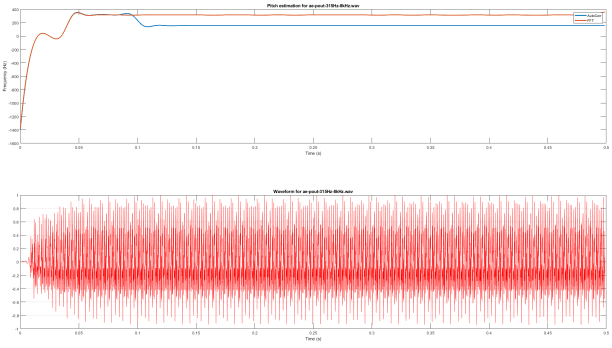

Figure 2: /ae/ speech signal with $f_0 = 315$ Hz. Classified as Child.

*if there are no positive peaks in that range then we assume the f_0 is 0 and the speech segment is not voiced.

First of all, notice how the plotted line extends to really low/high numbers for low time indexes. This is due to how the spline interpolation method works. The vector's content itself is pretty much stable and near to a specific frequency value while only the first couple items are further away from that frequency value (more specifically, the first 2 elements are around 400hz - almost double the actual frequency! This is due to the fact that at the first frame the signal behaves differently as the speech signal starts).

We can see that for the Figure 1, the two methods are performing the same, with the FFT doing a miss calculation at the end. Both of them correctly found that the fundamental frequency of this speech waveform is 205 Hz.

For the Figure 2, we can see differences between the two methods, the FFT method has correctly found that the pitch of this speech waveform is 315 Hz, unlike the auto correlation method that has correctly found a pitch of approximately 180 Hz.

The rest of the speech signals behave in a similar manner. The FFT method usually performs better and achieves a closer result to the ground truth.

3 Gender and Age Detection Accuracy

The age and gender detection algorithm is based on fixed fundamental frequencies for all of genders and ages, and since the FFT peak picking method performs better than the auto correlation one, we are using all the f_0 we found for each frame with the FFT method.

We then classify all the f_0 values we found for each frame in adult male, adult female or child, depending on the f_0 in a specific frame, resulting in 3 different vectors, where Adult Male range was [70Hz, 160Hz], the Adult Female was [160Hz, 275Hz] and finally the child range was [275Hz, 500Hz]. Then, we pick the category who's vector has the most elements (maximum length vector). We also considered other methods for classification, such as taking the maximum, mean or median f_0 value, but they appeared to be more vulnerable to outliers (e.g. if there is a high-pitched frame due to the speaker's voice fry or even lombard speech).

We tested it in all of our given voiced (-pout) speech files, and for all of them, it correctly detect the age and gender of the speaker. To decide which technique is more accurate, instead of visually assessing the results we could probably have used a Loss function and established an automated comparison between the 2 methods (e.g. Mean Squared Error from the ground truth).

4 Full Speech Waveform using VUS

We extended our system to work with not strictly voiced speech signals, using our VUS discriminator. The gender+age detection was applied only to speech segments classified as voiced, and considered 0 frequency for the rest. Also, we increased the Energy threshold for voiced speech segments to minimize the false voiced classification rate of our discriminator.

Let's take a look at some sentence speech signals and how the detector works with the VUS. Greetings from Gibbs' Phenomenon! ☺

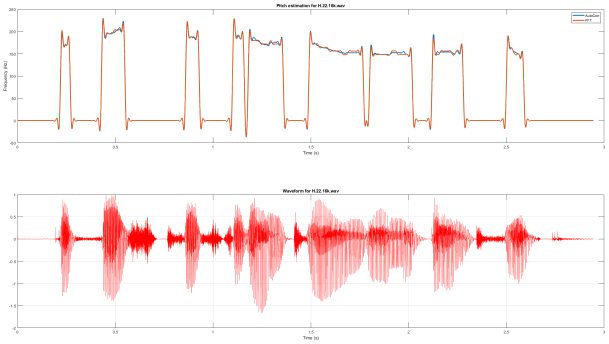

Figure 3: An adult female saying "The fish twisted and turned on the bent hook". Classified as Adult Female.

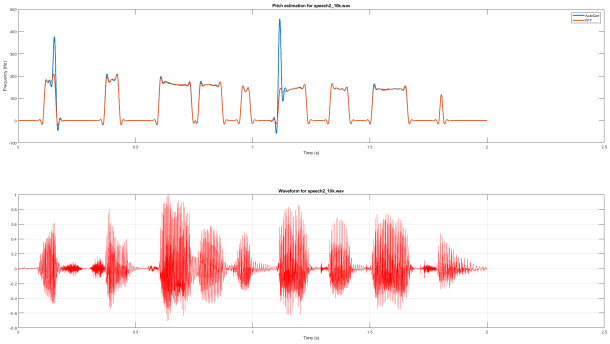

Figure 4: An adult male saying "Which tea party did Baker go to?". Classified as Adult Male.

For the left figure, we can see that both the FFT and the auto correlation method work the same, correctly classifying the speaker as an adult female.

For the right figure, we can some spikes that are erroneously detected by the auto correlation method. Since we use the FFT method for the classification, it correctly classified the speaker as an adult male.

5 Other Plots

In the last four figures, there are a lot of differences between the two methods, with the FFT again performing better than the auto correlation method. The spline interpolation method early big slope is also common on all signals. Also, notice how both speech wave forms in figure 13 and 14 - the Lombard speech signals - are mistakenly classified as adult female and child respectively. The voice itself is high pitched, but apparently the stressed voice also plays a key role in speech, affecting the classification result.

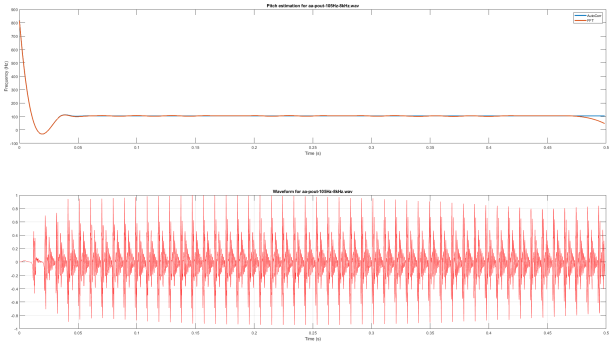

Figure 5: /aa/ speech signal with f0 = 105 Hz.
Classified as Adult Male.

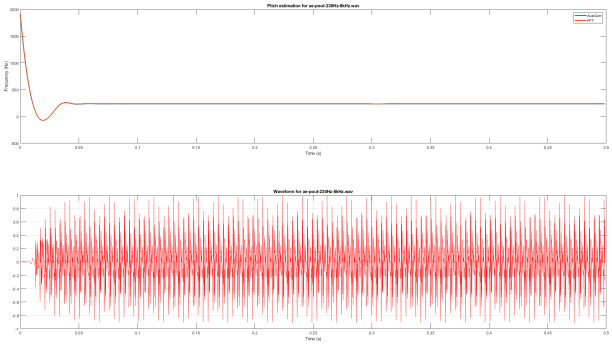

Figure 6: /ae/ speech signal with f0 = 230 Hz.
Classified as Adult Female.

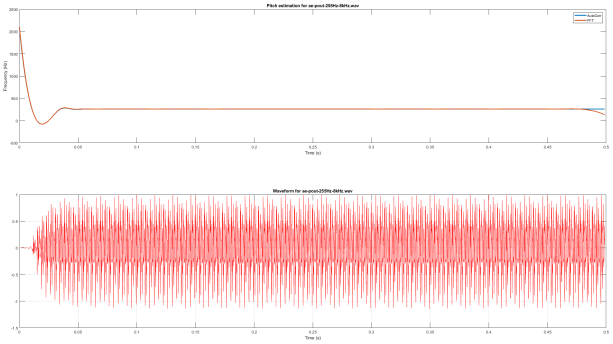

Figure 7: /ae/ speech signal with f0 = 255 Hz.
Classified as Adult Female.

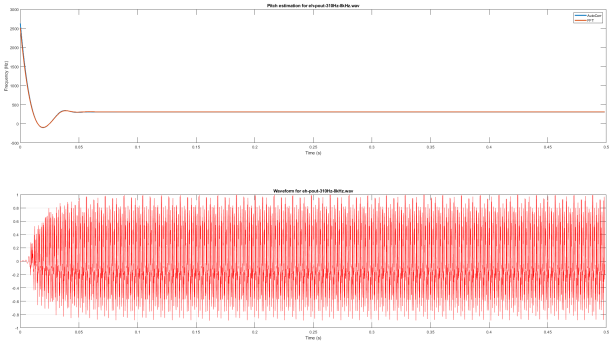

Figure 8: /eh/ speech signal with f0 = 310 Hz.
Classified as Child.

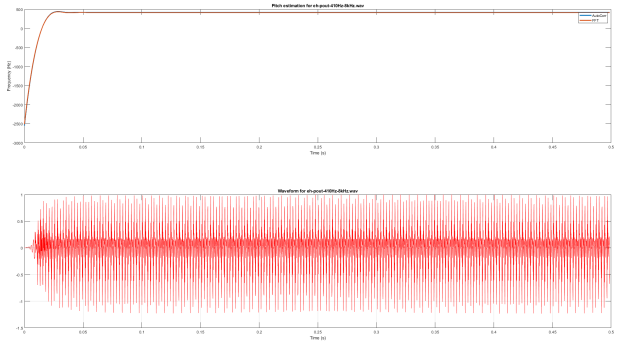

Figure 9: /eh/ speech signal with f0 = 410 Hz. Classified as Child.

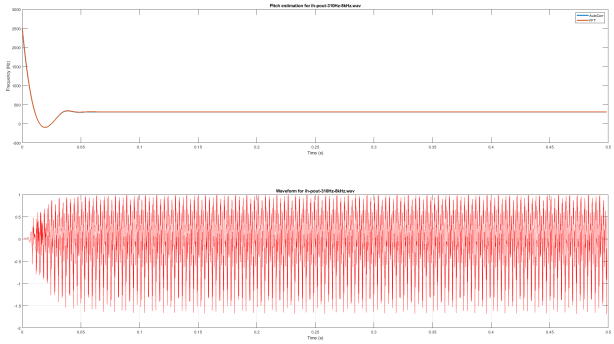

Figure 10: /ih/ speech signal with f0 = 310 Hz. Classified as Child.

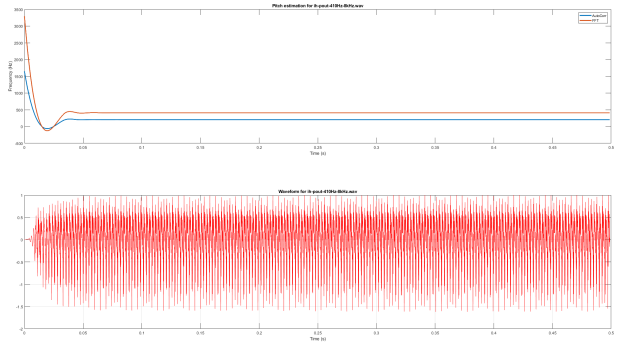

Figure 11: /ih/ speech signal with f0 = 410 Hz. Classified as Child.

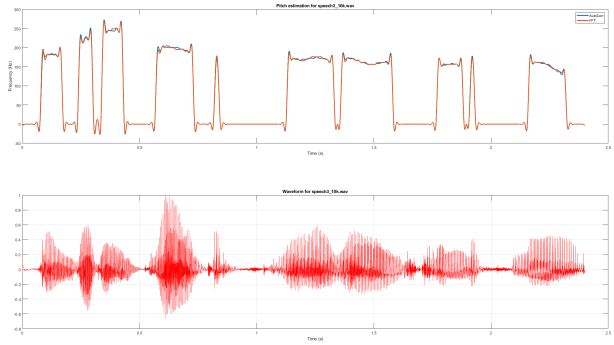

Figure 12: An adult female saying "A little blanket laid around on the floor". Classified as Adult Female.

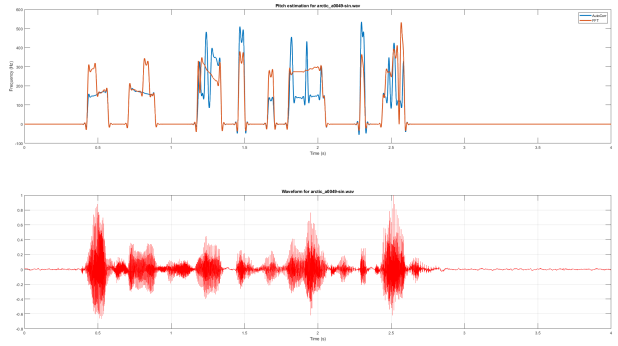

Figure 13: An adult male saying "Gregson was asleep when he rentered the cabin". Classified as Adult Female.

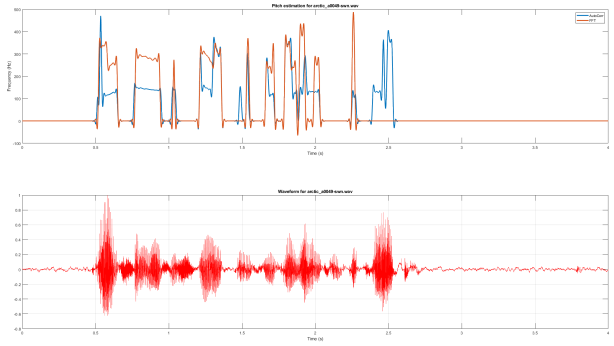

Figure 14: An adult male saying "Gregson was asleep when he rentered the cabin". Classified as Child.

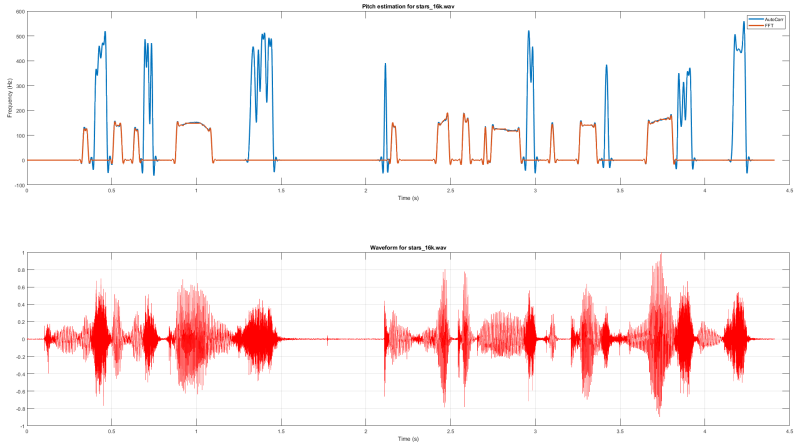

Figure 15: George Manos speaking. Classified as Adult Male.

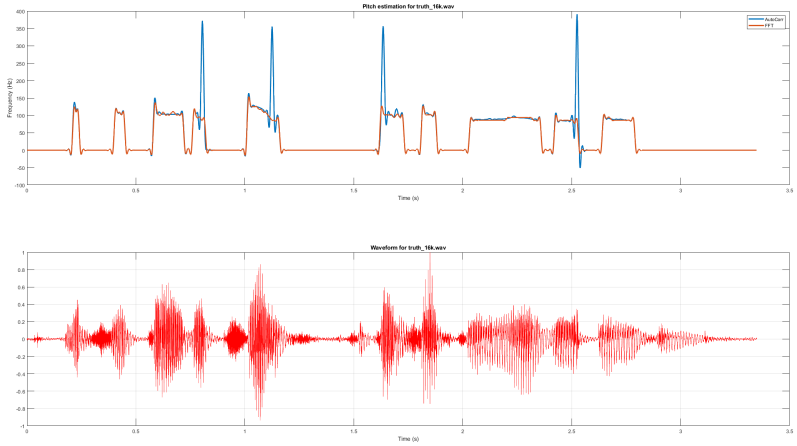

Figure 16: Alexandros Angelakis speaking. Classified as Adult Male.

6 Appendix

a.k.a. ELTA transform ☺