Project 0: Part 2
A second hands-on lab on Speech Processing
Frequency-domain processing

February 12, 2020

In this lab, you will familiarize yourself with frequency domain analysis of speech signals. You will explore the frequency domain structure of the most basic speech elements, such as vowels and consonants, using the Fast Fourier Transform (FFT). You will learn about time-frequency representation of speech signals, with the help of Short-Time Fourier Analysis (spectrogram), and you will estimate basic speech parameters, such as the fundamental frequency.

The final goal of this lab is to implement a simple system for speaker gender (male, female) and age (adult or children) detection.

1 Theoretical Background

You will first familiarize yourself with the time-frequency representation of speech, the so-called *spectrogram.* The spectrogram can be produced using wideband or narrowband analysis. Wideband analysis includes the use of short analysis windows in time, whereas narrowband analysis is performed using long analysis windows in time.

1.1 Short Time Fourier Analysis - Spectrogram

In the previous lab, you have seen that speech consists of a sequence of different events. These events are so radically different both in time and in frequency that a single Fourier transform over the whole speech signal cannot capture the time-varying frequency content of the waveform. In contrast, the *short-time Fourier Transform - STFT* consists of separate Fourier Transforms on pieces of the waveform under a sliding window - pretty much like we did in the previous lab, but in frequency domain this time. :-)

The Fourier transform of the windowed speech waveform (STFT) is given by

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x[n, \tau] e^{-j\omega n} \tag{1}$$

where

$$x[n, \tau] = w[n, \tau] x[n] \tag{2}$$

represents the windowed speech segment as a function of the center of the window, at time τ. The *spectrogram* is a graphical 2D display of the squared magnitude of the time-varying spectral characteristics and it can be described mathematically as

$$S(\omega, \tau) = |X(\omega, \tau)|^2 \tag{3}$$

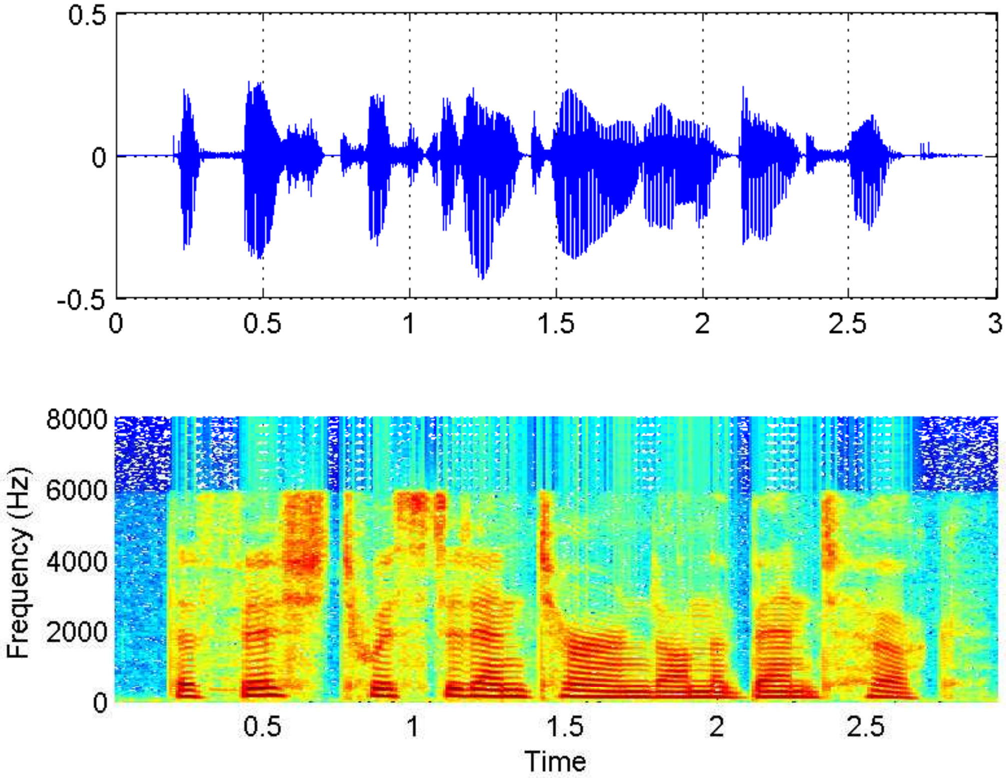


Figure 1: *Narrowband analysis of speech.*

For voiced speech, we can approximate the speech waveform as the output of a linear time-invariant system with impulse response $h[n]$ and with a glottal flow input given by the convolution of a series of periodically placed impulses, $p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP]$, with P being the pitch period, and a glottal flow over one cycle, $g[n]$:

$$x[n, \tau] = w[n, \tau](p[n] * g[n] * h[n]) \tag{4}$$

Thus, the spectrogram can be expressed as

$$S(\omega, \tau) = \frac{1}{P^2}\Big|\sum_{k=-\infty}^{\infty} H(\omega)G(\omega)W(\omega - \omega_k, \tau)\Big|^2 \tag{5}$$

Now based on that expression, there are two different types of STFT analysis, according to the window length that is used. A long window (i.e. up to 3 or 4 pitch periods) results in *narrowband* analysis, whereas a short window (i.e. a pitch period or even less) results in *wideband* analysis. You should already know that the length of the window affects its spectral characteristics, and mainly the size of the mainlobe (and thus, its bandwidth). Also, you should know that multiplying a window with a speech segment in time results in a frequency convolution between the corresponding spectra. Hence, simply speaking, the spectrum of the analysis window is "placed" around and on the harmonics of the underlying speech spectrum. Keeping this in mind, let us discuss the wideband and narrowband analysis.

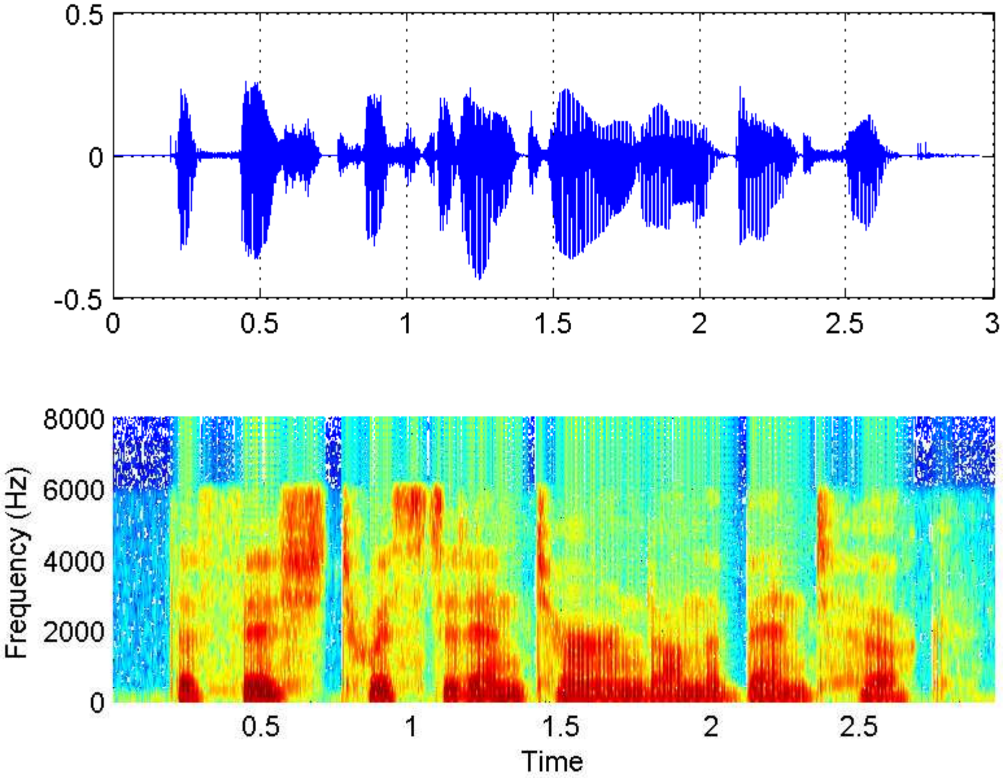


Figure 2: *Wideband analysis of speech.*

1.1.1 Narrowband analysis

As we said, the narrowband spectrogram, a ”long” - in time - analysis window is used, typically of duration of at least two pitch periods (more than 20 ms). Under the condition that the main lobes of the shifted window Fourier transforms are non-overlapping, and that the sidelobes of the window transform are negligible, we can approximately state that

$$S(\omega, \tau) \approx \sum_{k=-\infty}^{\infty} |G(\omega_k)H(\omega_k)|^2 |W(\omega - \omega_k, \tau)|^2 \tag{6}$$

A typical narrowband spectrogram is given in Figure 1. The code that generated it is given:

```
[s, fs] = wavread(’H.22.16k.wav’);
t = 0:1/fs:length(s)/fs - 1/fs;
% Window length of 30 ms and step of 10 ms
figure; subplot(211); plot(t, s); xlabel(’Time (s)’); subplot(212);
spectrogram(s, 30*10^(-3)*fs, 20*10^(-3)*fs, 1024, fs, ’yaxis’)
```

We can see that using a long window in time on a voiced segment gives a STFT that consists of a set of narrow ”harmonic” lines - whose width is determined by the Fourier transform of the window - which are shaped by the magnitude of the product of the Fourier transform of the glottal flow and the vocal tract transfer function. The narrowband spectrogram gives *good frequency resolution* because the harmonics are effectively resolved (horizontal striations on the spectrogram). However, it also gives *poor time resolution*, because the long analysis window covers several pitch periods and thus is unable to reveal fine periodicity changes over time. It should be noted that colors in spectrogram have a meaning: intense red or black color corresponds to high magnitude values (high energy), whereas yellow or blue color is for low magnitude areas (and thus, low energy regions).

1.1.2 Wideband analysis

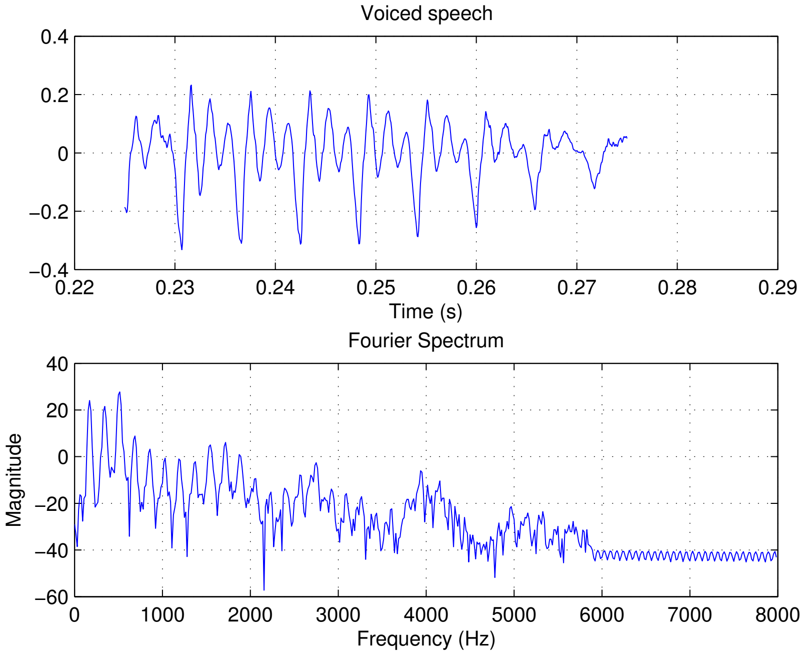
For the *wideband* spectrogram, a ”short” window is chosen with a duration of less than a single pitch period. By shortening the window length, its Fourier transform ”widens”. This ”wide” Fourier transform of the window, when ”placed” on the harmonics, will overlap and add with its neighbouring window transform and smear out the harmonic line structure, roughly revealing the spectral envelope $|H(\omega)G(\omega)|$ due to the vocal tract and glottal flow contributions. Thus, *poor frequency resolution* is provided by wideband analysis, but *good time resolution* is provided. For a steady-state voiced segment, the wideband spectrogram can be very roughly approximated as

$$S(\omega, \tau) \approx |H(\omega)G(\omega)|^2 E[\tau] \tag{7}$$

where $E[\tau]$ is the energy of the waveform under the sliding window. Thus, the spectrogram shows the formants of the vocal tract in frequency, but also gives vertical striations over time. These vertical striations arise because the short window is sliding through fluctuating energy regions of the speech waveform. A wideband spectrogram is depicted in Figure 2. The code is given below:

```
% Window length of 5 ms and step of 3 ms
figure; subplot(211); plot(t, s); xlabel(’Time (s)’); subplot(212);
spectrogram(s, 5*10^(-3)*fs, 2*10^(-3)*fs, 1024, fs, ’yaxis’);
```

Colours have the same meaning as in narrowband spectrogram.


Voiced speech

Fourier Spectrum

Figure 3: *FFT spectrum of voiced speech.*

1.2 Fourier Transform and Spectral Content of Speech

So, it is obvious that the STFTs are generated by ”concatenated slices” of Fourier spectra. According to the type of analysis, we get either the harmonic structure or an approximation of the vocal tract formants. However, a question should be: how is the spectral content of different speech elements? Let us find out! :-) For our purpose, a wideband analysis is not convenient, since it does not reveal the spectral content of the source, but rather the the envelope of speech. Thus, a narrowband analysis will be used.

If we select a voiced speech portion, long enough to resolve the harmonics in the spectrum, and apply a FFT on it, what we have is in Figure 3. The necessary MATLAB code is given:

```
% Loading the waveform
[s,fs] = wavread('H.22.16k.wav');
% Extracting a frame
frame1 = s(3600:4400);
L1 = length(frame1);
% Windowing it
frame_v = hamming(L1).*frame1;
% Apply FFT and then take the absolute value in 1024 points
NFFT = 1024;
X1 = abs(fft(frame_v, NFFT));
% Make frequency bins into frequencies
freq = [0:fs/NFFT:fs/2-1/fs];
% Plot
subplot(211); plot(frame1); xlabel('Time (samples)'); grid;
subplot(212); plot(freq, 20*log10(X1(1:NFFT/2)));
ylabel('FFT Magnitude'); xlabel('Frequency (Hz)'); grid;
```

It is clear that the horizontal striations that are seen in Figure 1 come from the harmonic peaks of the FFT spectra. It is also clear that the speech harmonics are up to 4 kHz, and the rest of the spectrum is mostly covered by noise[1]. The spectrum and its peaks are a means to build our gender and age detection system.

If we select an unvoiced speech portion, long enough to resolve any structure (surely not harmonic) in the spectrum, and apply a FFT on it, what we have is in Figure 3. The code that produces this figure is given below:

```
% Loading the waveform
[s,fs] = wavread('H.22.16k.wav');
% Extracting a frame
frame2 = s(4800:5500);
L2 = length(frame2);
% Windowing it
frame_unv = hamming(L2).*frame2;
% Apply FFT and then take the absolute value in 1024 points
NFFT = 1024;
X2 = abs(fft(frame_unv, NFFT));
% Make frequency bins into frequencies
```

[1] Recent studies, however, have shown that speech is (quasi-)harmonic up to 16 kHz!!

```
freq = [0:fs/NFFT:fs/2-1/fs];
% Plot
subplot(211); plot(frame2); xlabel('Time (samples)'); grid;
subplot(212); plot(freq, 20*log10(X2(1:NFFT/2)));
ylabel('FFT Magnitude'); xlabel('Frequency (Hz)'); grid;
```

We can see that the spectrum of unvoiced speech is almost flat and covers the whole spectrum. There is no harmonic structure. This representation is consistent with the approximation of unvoiced speech as white noise, and the spectrogram information that we get in unvoiced regions (see unvoiced parts in Figure 1).[2].

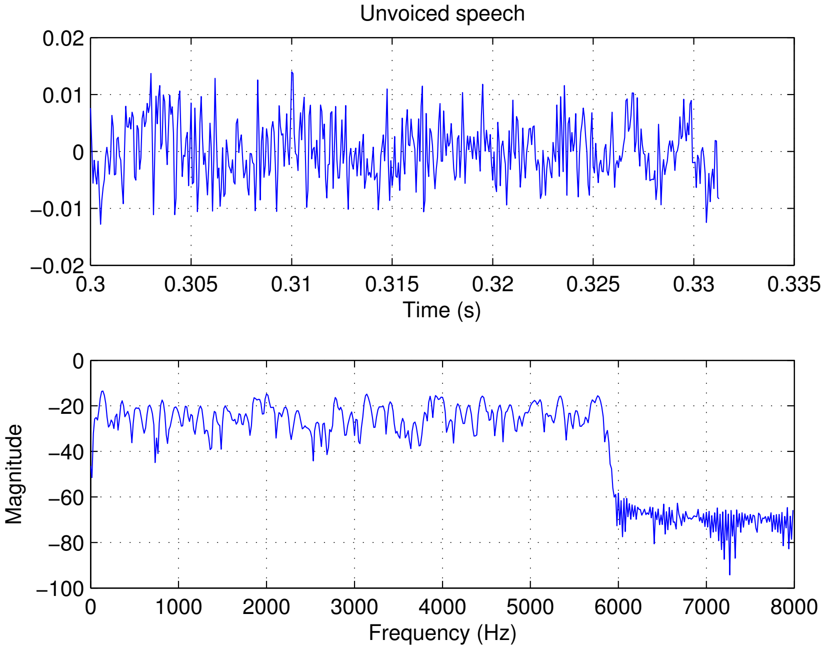


Figure 4: *FFT spectrum of unvoiced speech.*

1.3 Pitch

The periodic opening and closing of the vocal folds results in the harmonic structure in voiced speech signals. The inverse of the period is the *fundamental frequency* of speech. Pitch is the perceived sensation of the fundamental frequency of the pulses of airflow from the glottal folds. So, although they are not the same, sometimes the terms pitch and fundamental frequency are used interchangeably in literature.

The pitch is determined by four main factors. These include the length, tension, and mass of the vocal cords and the pressure of the forced expiration also called the sub-glottal pressure. The pitch variations carry most of the intonation signals associated with prosody (rhythms of speech), speaking manner, emotion, and accent. Figure 1 illustrates an example of the variations of the trajectory of pitch (and other harmonics) over time.

[2] In all speech sample waveform depicted in these figures, the signal is lowpass-filtered at 6 kHz, and that is why there is no spectral content -not even noise- above 6 kHz

Among others, the following information is contained in the pitch signal:
(a) Gender is conveyed in part by the vocal tract characteristics and in part by the pitch value. The average pitch for females is about 200 Hz whereas the average pitch for males is bout 110 Hz. Hence, pitch is the main indicator of gender.
(b) Age and state of health. Pitch can also signal age, weight and state of health. For example, children have a high-pitched speech signal of $300 - 400$ Hz.

Hence, we can detect the gender and the age of a speaker by tracking his/her pitch. :-) Thus, we should implement some simple techniques for pitch tracking. For this, we will describe and implement a simple time-domain and a simple frequency-domain method for estimating the pitch of voiced speech, and therefore a simple gender+age detection system can be implemented.

2 Pitch Tracking Techniques

Pitch tracking is still a very hot topic of research in speech signal engineering - even deep learning is utilized to this task! Although there are several algorithms in literature, the robust estimation of pitch is still a relatively open subject.

For our purpose, we will implement and compare a pair of rather simple (and for that, not very efficient :-)) methods for pitch estimation. Our pitch estimates can then give us an idea about the gender and the age of the speaker.

2.1 Short-time autocorrelation method

The autocorrelation function is (or should be :-)) known to you from Digital Signal Processing courses. We will remind you here the most basic notions of the autocorrelation theory.

The autocorrelation function of a discrete-time deterministic signal is defined as

$$\phi(k) = \sum_{m=-\infty}^{\infty} x[m]x[m+k] \tag{8}$$

The autocorrelation is a measure of similarity between signals. For example, if the signal is periodic with period P samples, then it can be shown that

$$\phi(k) = \phi(k+P) \tag{9}$$

i.e. the autocorrelation function of a periodic signal is also periodic with the same period. It can also be easily shown that for periodic signals, the autocorrelation function attains a maximum at samples $0, \pm P, \pm 2P, ...$ That is, the pitch period can be estimated by finding the location of the first maximum in the autocorrelation function.

If we apply the autocorrelation function in the voiced speech segment that is presented in the examples above, we get the result of Figure 5. As you can see, the first peak of the autocorrelation function is at time $t_0 = 0.005875$ sec, which corresponds to $f_0 = 1/t_0 = 170.2128$ Hz. If we measure the distance of the highest peaks in the waveform, we can see that it is $D = 0.2434 - 0.2376 = 0.0058$, which is the same result, and thus the pitch is $f_0 = 1/0.0058 = 170.2128$ Hz. :-)

For your convenince, MATLAB has its own function for correlation measurements. It is *xcorr* and it was this function that generated the result in Figure 5.

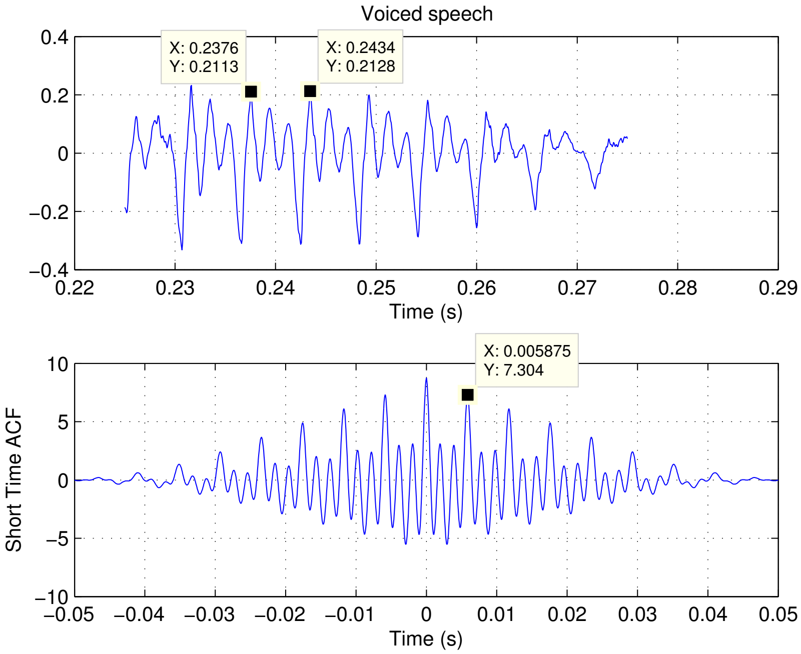


Figure 5: *Upper panel: Voiced speech waveform. Lower panel: Autocorrelation function of voiced speech*

2.2 Peak picking

As it is shown in the previous sections, voiced speech has a certain structure in frequency domain: it is dominated by sharp peaks at frequency locations that are nearly harmonically related to the fundamental frequency. Since the first significant peak of the spectrum is related to the fundamental frequency (and thus, the pitch), we can develop an algorithm that can perform peak-picking on an FFT spectrum and reveal not only the pitch but the whole harmonic structure a voiced speech segment! :-)

For example, let us take a look at the magnitude spectrum of the usual voiced speech spectrum, and select the first significant peak, we will see the result of Figure 6. The first peak is located at frequency $f_0 = 171.9$ Hz, which is very close to 170.2 Hz. However, the mismatch can be due to the fact that the signal is not strictly periodic, or due to the resolution of the FFT (1024 points). Of course, the actual pitch is unknown, so we cannot validate our result, unless we create a synthetic signal that has known parameters. :-)

3 Age+Gender Detection System Implementation

You will use the pitch trackers described above in order to design your age+gender detection system. For your convenience, follow the next steps:

1. Load one of the provided waveforms that end in *-pout.wav*. These signals are purely voiced, synthetic speech, with known f_0 and sampling frequency $F_s = 8$ kHz. Perform pitch estimation using an approach similar to the one used in VUS discriminator:
   - Do a frame-by-frame analysis, with an analysis window of 30 ms and a frame rate of 10 ms.
   - Estimate the pitch for each frame using both algorithms - FFT peak-picking and ACF. Use MATLAB's built-in functions *fft* and *xcorr*. Do not forget to apply a Hamming window on your speech segment! You also have to write your own peak picking algorithm (not so

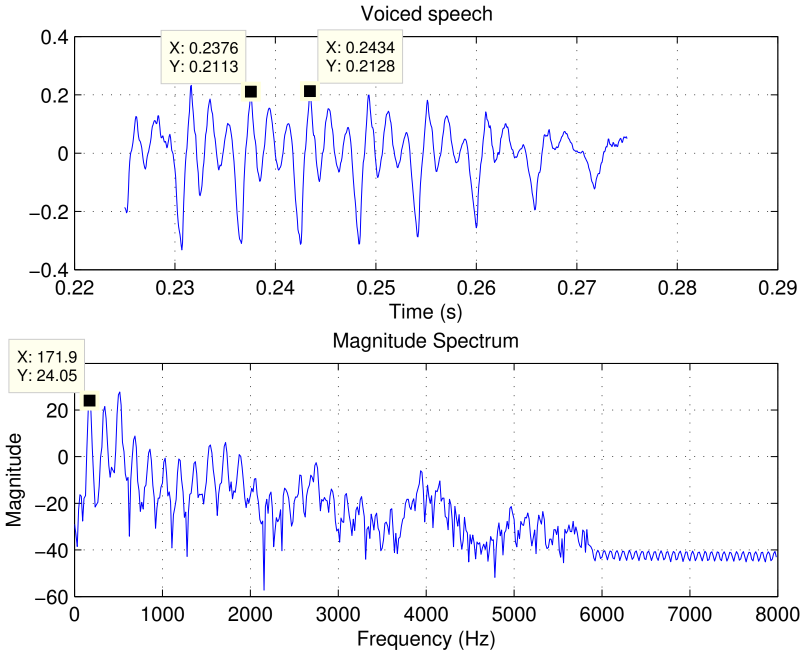


Figure 6: *Upper panel: Voiced speech waveform. Lower panel: Magnitude Spectrum and first peak*

difficult - a simple first derivative criterion is enough). A MATLAB function can be written like this:

```
function [out1, out2] = function_that_does_something(in1, in2, in3)
% Comments
% FUNCTION_THAT_DOES_THAT takes in1, in2, in3 arguments and returns out1, out2

%CODE
%CODE
%CODE

out1 = %CODE
out2 = %CODE
% End of function
```

Then you can save it as `function\_that\_does\_something.m` file and call it whenever you like.

- Use an FFT resolution of 2048 points.
- Interpolate your pitch estimates using splines in order to obtain a *pitch contour*. The *interp1* function of MATLAB will become useful.
- Optional: perform peak picking in ALL peaks of the spectrum and construct an estimate of the *frequency grid* of the voiced speech waveform.

2. Which contour is closer to the true frequency given in the name of the *-pout.wav* files?

3. Which method performs better? Why?

4. For gender+age detection, you are given that an adult has a pitch ranging from 70 to 250 Hz, whereas a child has a pitch range from 300 to 500 Hz. A male adult ranges from 70 to 150 Hz, and the pitch of a female adult lies in the range $160 - 250$ Hz.

5. According to the previous note, the output of your system should be a plot of the speech waveform, a plot of the pitch contour, and a text string, *'adult male', 'adult female', 'child'*.

6. **Optional**: Use the VUS discriminator of the previous lab and the pitch tracker of your choice, and build the pitch contour for a full speech waveform! :-) (Care should be taken for the non-voiced parts: since the ACF and the peaks do not correspond to any pitch, you can pre-detect non-voiced parts with your VUS discriminator and set the pitch to zero in these time intervals).

7. **Delivery deadline: 24 February 2020**

If you have ANY questions on this lab, please send an e-mail to : hy578-list@csd.uoc.gr