Practice I. Signal analysis with DFT: Application to speech signals.

1 Objectives

- Review basic notions about the Fourier Transform, the DFT and the window effect on discrete signals.
- Model speech signals and understand their characteristics in the time and frequency domains.
- Relate the signal features obtained in the frequency domain with physical features of the human voice production system.

2 Previous Study

2.1 The Discrete Fourier Transform (DFT)

The discrete Fourier transform (DFT) of a sequence x[n] is defined as

$$X[k] = DFT_N\{x[n]\} = \sum_{n=0}^{N-1} x[n]e^{-j2\pi\frac{k}{N}n}, \qquad 0 \le k \le N-1$$

which generates N frequency samples.

As it can be seen, the DFT observes only samples of $x[n]$ in the interval $n \in [0, N-1]$. That is, the signal is implicitly windowed with a rectangular window of length N. In addition to this implicit windowing, we can apply an additional window $v[n]$ of length L≤N (e.g., rectangular, Hamming, Hanning, etc.), obtaining the windowed signal:

$$x_v[n] = x[n] \cdot v[n]$$

It is known that the **DFT of $x_v[n]$ is equal to the sampling of the Fourier Transform of $x_v[n]$**, that is:

$$X_v[k] = DFT_N\{x_v[n]\} = X_v(F)|_{F=k/N} = X(F) \otimes V(F)|_{F=k/N}, \qquad 0 \le k \le N-1 \qquad \textbf{\textit{(Eq. 1)}}$$

where $\otimes$ denotes the periodic convolution and $X_v(F)$, $X(F)$ and $V(F)$ are the Fourier transform of $x_v[n]$, $x[n]$ and $v[n]$, respectively. In particular, if $v[n]$ is a rectangular window of length L, we have that

$$V(F) = e^{-j\pi F(L-1)} \frac{sin(\pi FL)}{sin(\pi F)}$$

The election of the window v[n] is determined by the requirements in terms of resolution and sensitivity:

- **Resolution:** (spectral) resolution is the capacity of distinguishing spectral components of similar frequency. More specifically, the spectral resolution is given by the minimum frequency separation of two spectral components of x[n] that can be distinguished from each other. The

spectral resolution is determined by the width of main lobe of V(f). The resolution can be improved by increasing the window length L. An important theoretical result is that, for a given window length L, the maximum resolution is achieved by the rectangular window.

- **Sensitivity:** (spectral) sensitivity is the capacity of discriminating weak spectral components in the presence of stronger spectral components of rather distinct frequency. The spectral sensitivity is determined by the (relative) amplitude of the sidelobes of V(F). Sensitivity is a characteristic of the selected window (e.g., rectangular, Hamming, Barlett, etc.) but it does not depend on the window length L.

When selecting the window, there is a fundamental trade-off between resolution and sensitivity; the better the sensitivity, the worse the resolution. This trade-off is important in spectral analysis and will be studied further in the laboratory.

1. The following figures correspond to the magnitude of the Fourier transform of two windows of length L=8 samples. Indicate the window having the best resolution and the window having the best sensitivity. Justify your anwsers.

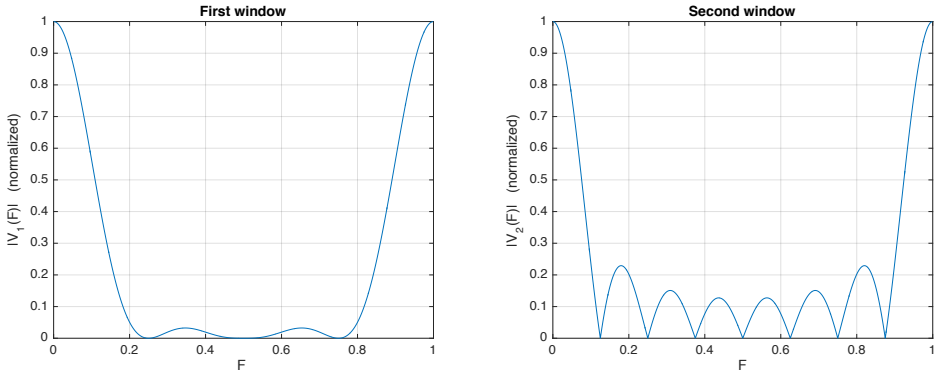


Figure 2.1: Magnitude of the Fourier transform of two windows (L=8)

The inverse DFT is given by

$$x_v[n] = DFT_N^{-1}\{X_v[k]\} = \frac{1}{N}\sum_{k=0}^{N-1} X_v[k]e^{j2\pi\frac{k}{N}n}, \qquad 0 \le n \le N-1$$

which generates N temporal samples. Note that the last N-L samples of $x_v[n]$ (i.e., L≤n≤N-1) are equal to zero due to the window $v[n]$.

Let us first characterize a sinusoid in the frequency domain:

2. What is the Fourier transform of the sinusoid x[n]=cos(2πF$_0$n)? Take the frequency in the interval (0,1).
3. Consider that we have only L=8 samples of the previous sinusoid (x[0], ..., x[7]) . Obtain the expression of the DFT of N=16 samples.

2.2 Frequency analysis of periodic sequences

A sequence x[n] is periodic if there exists an integer P such that x[n+P]=x[n] for any value of n. The lowest value of P holding the above identity is the *period* of x[n] and its inverse F_0=1/P the (discrete) *fundamental frequency* of x[n] (or, alternatively, $f_0=F_0 \cdot f_m$ its analog counterpart).

A periodic sequence x[n] can be expressed as the convolution of a basic sequence $x_b[n]$ with an infinite train of deltas:

$$x[n] = \sum_{i=-\infty}^{\infty} x_b[n - iP] = x_b[n] * \sum_{i=-\infty}^{\infty} \delta[n - iP] \qquad (Eq. 2)$$

The Fourier Transform (FT) of a periodic train of discrete deltas in the time domain is a periodic train of Dirac deltas in the frequency domain:

$$\sum_{i=-\infty}^{\infty} \delta[n - iP] \xrightarrow{FT} \frac{1}{P} \sum_{l=-\infty}^{\infty} \delta\left(F - \frac{l}{P}\right) = \frac{1}{P} \sum_{i=-\infty}^{\infty} \sum_{m=0}^{P-1} \delta\left(F - \frac{m}{P} - i\right) \qquad (Eq. 3)$$

Note that in Eq. 3 we make explicit that, for a fixed *i*, we have one period of the Fourier Transform, and the sum over *i* expands the periodicity (as any Discrete Fourier Transform).

Using Eq. 3, the Fourier Transform of x[n] is given by

$$X(F) = X_b(F) \frac{1}{P} \sum_{i=-\infty}^{\infty} \sum_{m=0}^{P-1} \delta\left(F - \frac{m}{P} - i\right) = \sum_{i=-\infty}^{\infty} \sum_{m=0}^{P-1} \frac{1}{P} X_b\left(\frac{m}{P}\right) \delta\left(F - \frac{m}{P} - i\right) \qquad (Eq. 4)$$

That is, the Fourier Transform of a periodic signal is composed of Dirac deltas at harmonic frequencies of F_0=1/P.

Given a window v[n] of length L samples, the windowed periodic signal $x_v[n]$=x[n]·v[n] has the following Fourier transform:

$$\begin{aligned} X_v(F) = X(F) \otimes V(F) &= V(F) * \sum_{m=0}^{P-1} \frac{1}{P} X_b\left(\frac{m}{P}\right) \delta\left(F - \frac{m}{P}\right) \\ &= \sum_{m=0}^{P-1} \frac{1}{P} X_b\left(\frac{m}{P}\right) V\left(F - \frac{m}{P}\right) \end{aligned} \qquad (Eq. 5)$$

The Fourier Transform of a windowed periodic signal is composed of a set of replicas of V(F) at harmonic frequencies of F_0=1/P. The complex amplitude of each replica is proportional to the value of the Fourier Transform of the basic signal $x_b[n]$ at the corresponding harmonic.

The N points DFT of the windowed periodic signal, with N≥L, is obtained by sampling (Eq. 5) as follows:

$$X_v[k] = X_v(F)|_{F=\frac{k}{N}} = \sum_{m=0}^{P-1} \frac{1}{P} X_b\left(\frac{m}{P}\right) V\left(\frac{k}{N} - \frac{m}{P}\right) \qquad (Eq.\ 6)$$

Now, we analyze a periodic sequence that is fundamental to characterize the speech signal in this practice. This sequence is a periodic train of unit impulses and it can be written as follows:

$$t_L[n] = \begin{cases} t[n] & 0 \le n \le L-1 \\ 0 & other\ n \end{cases} \qquad \text{with} \qquad t[n] = \sum_{i=-\infty}^{\infty} \delta[n - iP]$$

4. Write analytically the DFT of length N of this signal $t_L[n]$ (P << L ≤ N).

Finally, we consider an arbitrary periodic signal of the form:

$$x[n] = t[n] * h[n] = \sum_{i=-\infty}^{\infty} h[n - iP]$$

5. Check that x[n] is periodic, with period P, whatever the selected sequence h[n]. As an example, you are suggested to plot x[n] for h[n]=[1,1,1,1] and P=3 and P=5.
6. Find the Fourier transform of x[n] as a function of P and the Fourier transform of h[n], say H(F). Hint: Use Eq 5.

2.3 The speech signal

Fig. 2.2 shows an example of a speech signal that has been sampled at 8kHz. Two types of segments can be clearly identified. The first one, known as *voiced sounds*, corresponds to high energy and almost periodic segments of the signal whereas the second type corresponds to low energy segments that are called *voiceless sounds*.

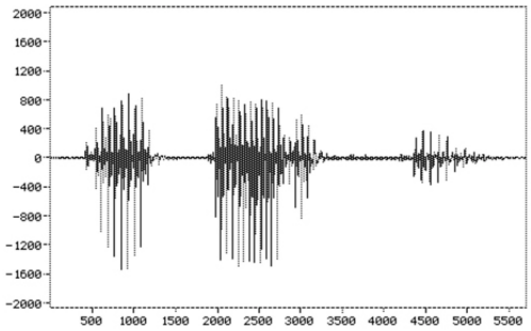


Figure 2.2: example of speech signal sampled at 8kHz.

Speech signal is produced by expulsing air from the lungs through the vocal tract. Voiced sounds are *periodic* sounds produced by the vibration of the vocal folds. Their fundamental frequency is determined by the frequency of the vocal folds vibration. It can be modified and forms the basis of the prosodic intonation and the singing ability. In average, the fundamental frequency (pitch) of voiced sounds is around 130Hz for men and 220Hz for women.

The resulting periodic signal propagates through the vocal tract that amplifies or attenuates the harmonic frequencies of the fundamental frequency. The amplified frequencies (resonances) are known as “formants” and are characteristic of specific sounds. They are essentially defined by the shape of the pharyngeal and oral resonance cavities. See Fig. 2.3 for a detailed view of the human speech production system.

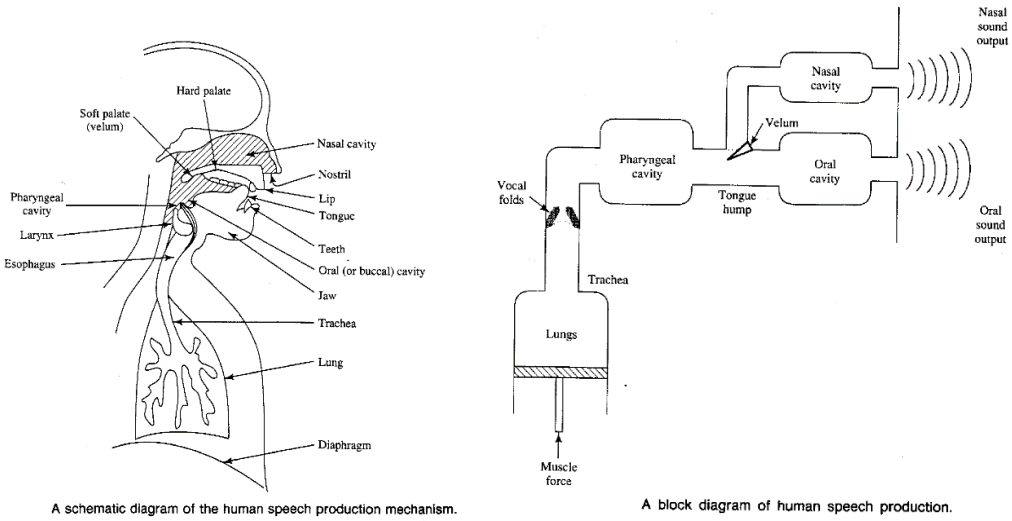


Figure 2.3: Illustration of the human speech production system.

As seen in the course, the speech production system can be modeled in terms of signals and systems as described in Fig. 2.4. The vocal tract can be modeled by a linear invariant filter $h[n]$ whose function is to amplify or to attenuate some frequencies (depending on the resonances of the cavities). The filter input signal can be either a periodic train of impulses $t[n]$ in the case of voiced sound or a random noise $r[n]$ in the case of voiceless sounds.

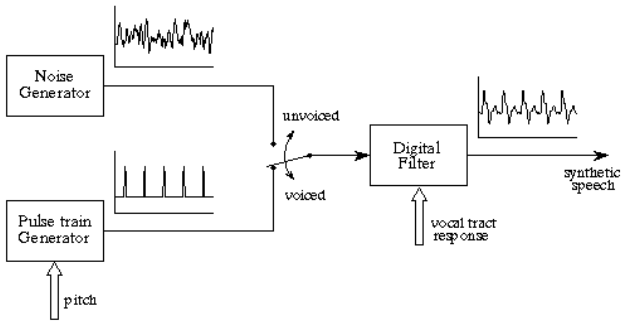


Figure 2.4: Model of the human speech production system.

7. Write the output of the system described by Fig 2.4 in both time and frequency domains for voiced sounds, in terms of the response of the digital filter (model of the vocal tract) and the signal period (pitch).

Vowels, which are voiced sounds, can be characterized by two formants. These formants depend on the specific configuration of the vocal tract (articulation) that produces the vowel. The articulation can be described in terms of the following three factors

a- The ***height***, which is the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw. The height controls the size of the constriction generated by the tongue hump.

b- The ***backness***, which is the position of the tongue during the articulation of a vowel relative to the back of the mouth. The backness controls the size of the resonant chambers.

c- The ***roundness*** of the lips.

Fig. 2.5 shows the articulation of the five Spanish vowels and the associated height and backness.

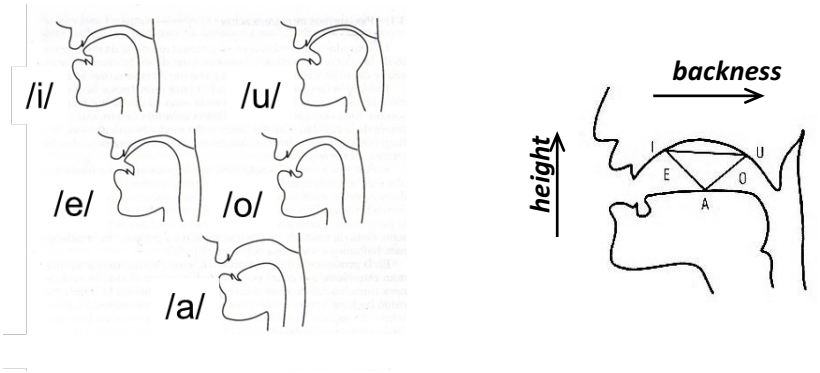


Figure 2.5: Pictures showing the articulation of the five Spanish vowels

The Fourier transform of a vowel shows two resonance frequencies (formants), which will be denoted by f_1 and f_2. Surprisingly, the two formants characterizing a vowel fall into a triangle as shown in the vowel diagram of Fig. 2.6.

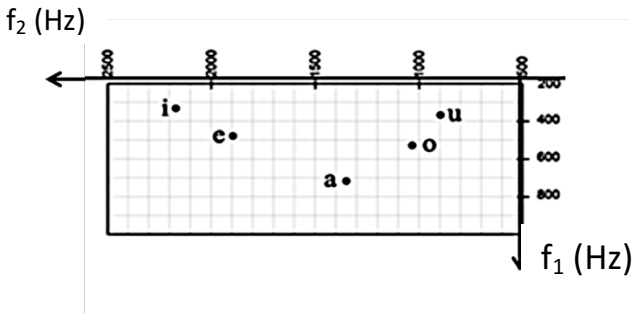


Figure 2.6: Vowel diagram providing the lower and higher resonance frequencies (formants) for the five Spanish vowels.

Finally, Fig. 2.7 shows an example of vowel segment sampled at 8 kHz that has been windowed (upper plot), and its DFT in logarithmic scale (lower plot).

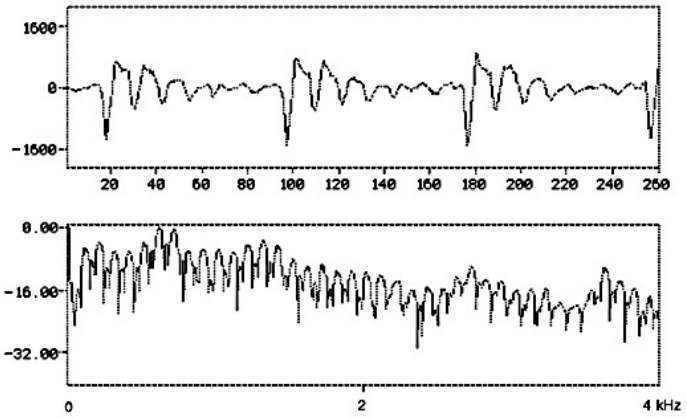


Figure 2.7: Example of a windowed vowel and its DFT (lower plot)

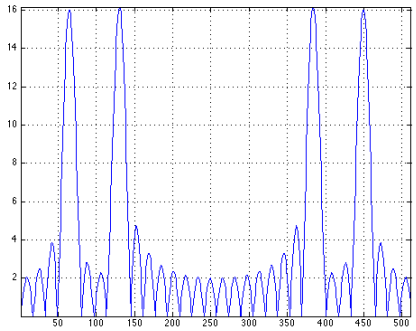
8. What is the fundamental frequency of the vocal folds? (Explain how you can estimate it from both the time-domain signal and its DFT in Figure 2.7)
9. Looking at figure 2.7, how can you see the windowing effect on the DFT of the signal?
10. In view of figure 2.7, what are approximately the two most important formants of this vowel?

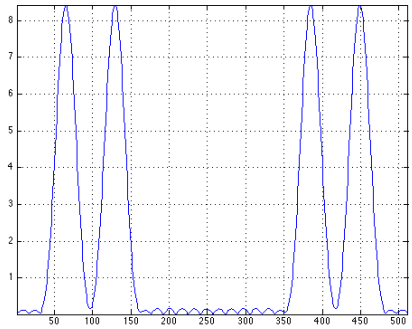
3 Laboratory

3.1 Window effect

The objective of this section is to understand the effect of a window function applied to a discrete signal both in the temporal and frequency domains. First of all, we are going to characterize two common window functions: a rectangular window $v_r[n]$ and a Hamming window $v_h[n]$. For both of them, we will study their resolution and sensitivity to detect spectral components. For this, we are going to use an input signal of L=32 samples composed of two separate tones F_1 and F_2 with amplitudes A_1 and A_2.

```
% Sinusoid parameters
A1 = 1; F1 = 0.125;
A2 = 1;
F2 = 0.25;
% Segment of 32 samples and its DFT
n = transpose(0:31);
xr = A1*cos(2*pi*F1*n) + A2*cos(2*pi*F2*n);
XR = fft(xr,512);
figure(1), plot(abs(XR)); axis tight; grid on;
% Applying the hamming window
vh = hamming(32);
xh = xr.*vh;
XH = fft(xh,512);
figure(2), plot(abs(XH)); axis tight; grid on;
```



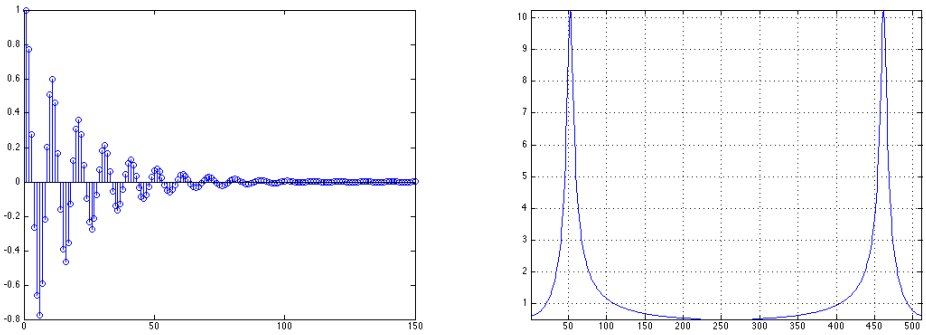


Figures 3.1 and 3.2: Two windowed tones using a rectangular window (left) and a Hamming window (right)

1. Fixing the values of A_1=1, A_2=1, F_1=0.125 and **F_2=0.148**, repeat the MATLAB code above and represent the DFT of 512 samples of the two windowed tones **xr[n]** and **xh[n]** (using a rectangular window and a hamming window respectively). Try to distinguish the two tones in the DFT representation. Can you easily distinguish the two tones F_1 and F_2 using both windows? From this result, which window has better ***resolution*** in frequency?
2. Now, using the values F_1=0.125, F_2=0.25, A_1=1 and **A_2=0.1** represent again the DFT of the two tones using the rectangular and the Hamming window and try to distinguish the two tones F_1 and F_2. Observing the resulting figures, which window has the better ***sensitivity***?

Next, we are going to study the structure of speech signals by means of synthetic signals and the effect of windowing in the spectral analysis of speech signals. For this, we are going to filter a train of unit impulses with a filter **h[n]**, that models the impulse response of the vocal tract. The adopted filter is defined next:

```
n = transpose(0:149);
h = (0.95).^n.*cos(2*pi*0.1.*n);
H = fft(h,512);
figure(3);stem(h);
figure(4);plot(abs(H));axis tight; grid on;
```



Figures 3.3 and 3.4: Ideal filter h[n] (left) and its DFT of N=512 samples (right)

We are going to filter the signal **e[n]**, that consists of a train of impulses separated by 60 samples:

```
e = zeros(1000,1); e(1:60:end) = 1;
figure(5),stem(e), axis tight;
```

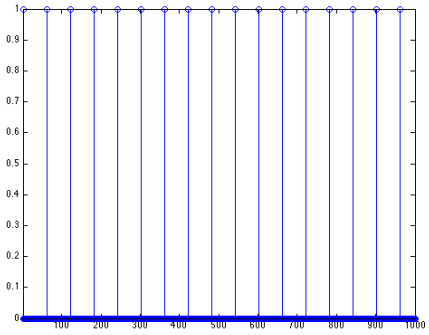


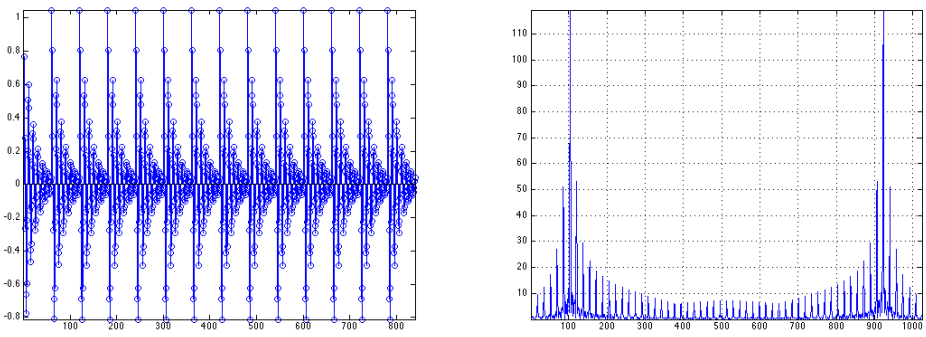
Figure 3.5: Signal e[n] composed of several unit impulses separated 60 samples

to obtain the resulting signal **y[n]**:

```
y = conv(e,h);
```

Finally, we apply a rectangular window of length L=840 so that we only observe the first L=840 samples of the sequence **y[n]**:

```
y = y(1:840);                          % windowing (rectangular window)
Y = fft(y,1024);
figure(6),stem(y); axis tight;
figure(7),plot(abs(Y)); axis tight; grid on;
```



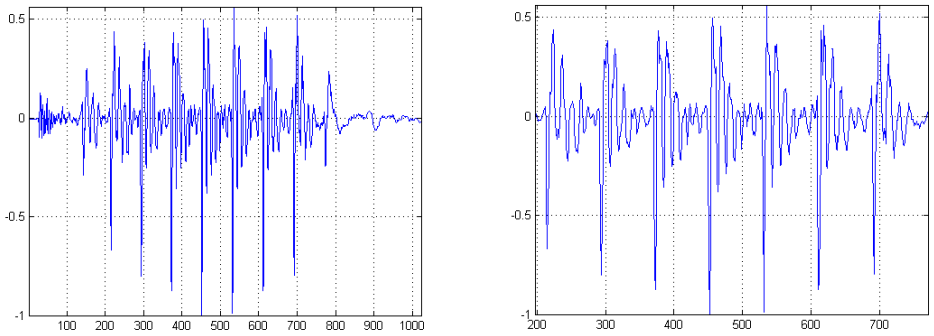
Figures 3.6 and 3.7: Signal y[n] (left) and its DFT of N=1024 samples (right)

3. Explain the differences between Fig. 3.7 and 3.4 and justify why there are 'spectral lines' in the DFT of **y[n]**. Relate the distance between the spectral lines in Fig. 3.7 with some feature of the signal **e[n]** (note that the DFT of **y[n]** in Fig. 3.7 has N=1024 samples). Explain which the effect of the window is.

3.2 Speech generation model

The objective of this section is to study the Fourier transform of vowels, and see the relationship of the module with the production of the signal.
The first exercise is to load the signal, inspect its shape, and periodicity properties. We are going to load a speech signal, in this case the sound 'can', and inspect its characteristics. The signal was recorded at a sample frequency of 8000 samples/second. In the following we plot the whole signal and the segment that contains the vowel (samples between 197 and 771).

```
FileName = 'can.wav';
[a,fs] = audioread(FileName);
figure(8); plot(a); axis tight; grid on;
interval = [197:771];
figure(9); plot(interval,a(interval)); axis tight; grid on;
```



Figures 3.8 and 3.9: Representation of the signal with the sound 'can' (left) and a segment with the vowel 'a' (right)

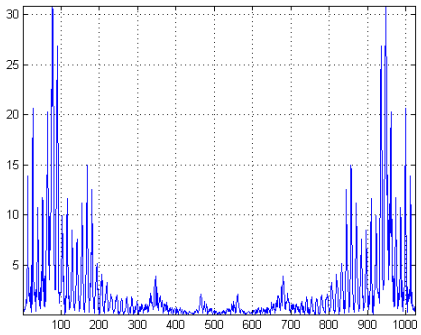
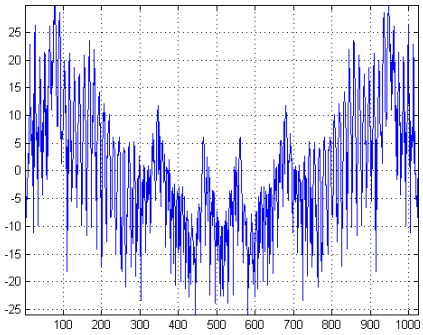
4. Select a new `interval` as a period of the vowel segment (an approximate version of the impulse response of the vocal tract for the vowel 'a') and estimate the fundamental frequency of the vocal chords.

As can be seen in Fig. 3.8 and 3.9, the vowel is practically a periodic signal. We are going to explore the frequency characterization. First, we are going to compute the DFT of the windowed signal. Note that we have selected 1024 samples for the DFT. This has been done because we use the FFT algorithm, that has been designed to work with signals of length equal to powers of two.

```
a_f=fft(a(interval),1024);
figure(10), plot(abs(a_f)); axis tight; grid on;
```

However, the plot of the absolute value of the FFT of the signal does not capture features of the frequency representation of the signal that are of low amplitude. It is known that the subjective perception of the loudness of sound is a logarithmic function of the intensity of the sound. Therefore a natural representation of the FFT is a representation that reflects the perception of the sound, i.e. a representation in a logarithmic scale.

```
figure(11), plot(20*log10(abs(a_f))); axis tight; grid on;
```

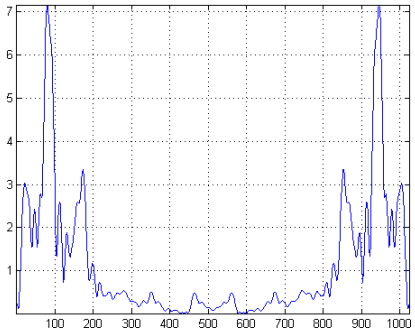
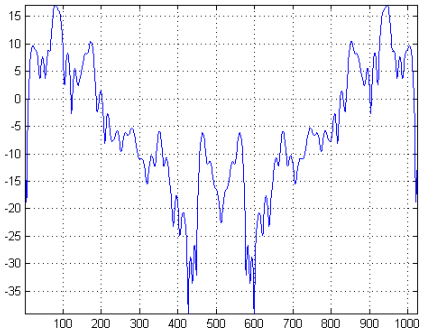



Figures 3.10 and 3.11: Plot of the absolute value of the FFT of a segment of the vowel 'a' (left) and in logarithmic scale (right)

5. Explain the spectral lines that appear in Fig. 3.10 and relate them to the fundamental frequency (pitch) of the vowel (similarly to question 3).

Plot the DFT of one period of the vowel 'a', which will be approximately equal to the impulse response of the vocal tract and therefore will capture the frequency characteristics of the vocal tract.

```
ha = a(370:449);
HA = fft(ha,1024);
figure(12); plot(abs(HA)); axis tight; grid on;
figure(13); plot(20*log10(abs(HA))); axis tight; grid on;
```

Figures 3.12 and 3.13: Absolute value of the DFT of one cycle of the vowel 'a' (left) and with logarithmic scale (right)

6. Identify the two most important formants and calculate their analogue frequencies. Do they have the expected values for a vowel 'a', as explained in the vowel triangle of Figure 2.5?
7. Load any of the WAV files 'unknown_x.wav' and, following a similar procedure as in the last two questions, relate the formants observed in the DFT plot to the vowels diagram of Figure 2.5 and try to guess which vowel is. *Note: do not forget to readjust manually the selected interval (variable 'interval').*