B2. Spectral Estimation by the DFT

Objectives

- Define artifacts of the DFT of a complex exponential: main lobe and sidelobes
- Show how the artifacts depend on the data length
- Define the frequency resolution of the DFT
- Show how the frequency resolution improves with data length
- Show the use of data windowing to improve the frequency estimate

1. Introduction.

In the previous chapter we introduced the Discrete Fourier Transform (DFT) of sequences with a finite duration. In particular we have seen that the frequencies present in the signal give rise to peaks in the magnitude of the DFT.

Since all signals can always be modeled as superpositions of complex exponentials, we can determine a one to one correspondence between a signal and its DFT.

A very important application is the numerical computation of the frequency spectrum of a given signal. In particular we want to find out what are the dominant frequencies of a signal so that we can infer whether (say) a target with a certain signature is present or not.

The typical problem is shown in the figure below.

Typical problem: collect a finite data set and estimate its frequency spectrum.
We have a signal \( x(t) \), sampled at \( F_s \) samples per second and we collect \( N \) samples. Based on the DFT of these \( N \) samples we want to estimate the frequency content of the given data.

The kinds of questions we try to answer:
- Which range of frequencies can we estimate;
- What would be the accuracy of the estimates;
- Something else ...

### 2. The DFT (and the FFT) for Spectral Estimation.

Recall the Discrete Fourier Transform. Given a signal \( x[n], n = 0, ..., N - 1 \) of finite duration, we define the DFT and Inverse DFT as follows

\[
X[k] = DFT\{x[n]\} = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn}
\]

\[
x[n] = IDFT\{X[k]\} = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j \frac{2\pi}{N} kn}
\]

with both \( k, n = 0, ..., N - 1 \).

In particular we saw that a complex exponential of frequency \( F_0 \) Hz, sampled at \( F_s \) samples per second, yields a discrete time complex exponential with digital frequency \( \omega_0 = \frac{2\pi F_0}{F_s} \) radians (it has no dimensions).

We have seen that the DFT of a sinusoid \( x[n] = A \cos(\omega_0 n + \alpha) \), for \( n = 0, ..., N - 1 \) is of the form

\[
X[k] = \left( \frac{A}{2} e^{j\alpha} \right) W_N(\omega - \omega_0) + \left( \frac{A}{2} e^{-j\alpha} \right) W_N(\omega - 2\pi + \omega_0)
\]

where (again refer to the previous chapter)

\[
|W_N(\omega)| = \left| \frac{\sin(\omega N / 2)}{\sin(\omega / 2)} \right|
\]

depends only on the length of the data. Its plot is shown in the figure below.
\[
\frac{W_N(\omega)}{N} = \frac{1}{N} \frac{\sin(\omega N/2)}{\sin(\omega/2)}
\]

If the signal \(x[n]\) of length \(N\) has a number of frequency components \(\omega_1, \omega_2, \ldots\) and so on, the contribution of each frequency component is in terms of \(W_N(\omega - \omega_1)\), \(W_N(\omega - \omega_2)\) \ldots and so on. Therefore the estimates of the spectral components strictly depend on the shape of the function \(W_N(\omega)\).

In order to better quantify \(W_N(\omega)\) let us look at its plot in dB’s, as shown in the figure below.

We see two important features:
• The Main Lobe, centered around $\omega = 0$, where the maximum is. The width of the main lobe, between the two zero crossings, is

$$\Delta \omega = \frac{4\pi}{N}$$

This represents the spread of the estimated spectrum around the maximum value. The narrower the main lobe is, the better the estimate is, since the plot gets more localized around the frequency of the signal. Fortunately the width of the main lobe is inversely proportional to the data length $N$, so the longer the data set, the “sharper” is the peak spectrum around the peak.

• The Side Lobes. These represent artifacts of the DFT caused exclusively by the finite data length and do not belong to the signal itself.

In what follows we see the effects of these two features on the ability of resolving different frequency components on the basis of the DFT alone.

### 3. Main Lobe and Frequency Resolution

In order to see the effect on the frequency resolution, suppose we have a signal with two frequency components $\omega_1, \omega_2$ as

$$x[n] = A_1e^{j\omega_1n} + A_2e^{j\omega_2n}, \quad n = 0, ..., N-1$$

If we take the DFT of the signal $X[k] = DFT\{x[n]\}$ we obtain, as usual

$$X[k] = A_1W_N(\omega - \omega_1) + A_2W_N(\omega - \omega_2)|_{\omega = k2\pi/N}$$

Now the issue of “resolution” is whether we can resolve the two frequencies based on the DFT. In other words, when we plot $|X[k]|$ with a good resolution we see two peaks, one per frequency. As long as the two main lobes of $|W_N(\omega - \omega_1)|$ and $|W_N(\omega - \omega_2)|$ are
distinct the two frequencies can be resolved. Since each main lobe has a width 
\[ \Delta \omega = \frac{4\pi}{N}, \]
we can say that we can resolve the two frequencies provided
\[ |\omega_1 - \omega_2| > \frac{4\pi}{N} \]
as shown in the figure below.

**Example.** Consider the following signal 
\[ x[n] = 3.0e^{j0.1n} + 2.0e^{j0.2n}, \ n = 0, \ldots, 127. \] Can we resolve the two frequencies by the DFT? We need to look at the difference
\[ |\omega_2 - \omega_1| = 0.1 \text{ radians}. \]
Since
\[ |\omega_1 - \omega_2| = 0.1 > \frac{4\pi}{128} = 0.0982 \]
the DFT of the signal has two distinct peaks as shown in the figure below.
If we zoom into the plot, around the two peaks we can estimate the frequencies, as shown in following figure. The maxima occurs at $k = 2, 4$ which yields an estimate of the two frequencies directly from the plot.

**Example.** Now consider the signal $x[n] = 3.0e^{j0.1n} + 2.0e^{j0.15n}$, $n = 0, ..., 127$. The two frequencies only $|\omega_2 - \omega_1| = 0.05$ radians apart, which is below the resolution threshold of 0.0982 radians. The result is only one maximum, and clearly we cannot resolve the two frequencies. This is shown in the plot which follows.

The resolution of the frequency spectrum from the DFT can be summarized as follows:

$$|\omega_1 - \omega_2| > \frac{4\pi}{N} = 0.0491$$

The resolution of the frequency spectrum from the DFT can be summarized as follows:
Digital Frequency Resolution. Given a signal \( x[n], n = 0, \ldots, N - 1 \) of length \( N \) with DFT \( X[k], k = 0, \ldots, N - 1 \), two digital frequencies \( \omega_1 \) and \( \omega_2 \) can be resolved from the DFT, provided

\[
|\omega_1 - \omega_2| > \frac{4\pi}{N}
\]

Using the relation \( \omega = 2\pi F / F_s \) between analog and digital frequencies, we can determine the resolution in terms of the frequencies of the analog signal as

\[
|F_1 - F_2| > \frac{2F_s}{N}
\]

Since we can write the sampling frequency in terms of the sampling interval as \( F_s = 1/T_s \), and \( N \times T_s = T_{\text{MAX}} \) is the total data length of the analog signal (say in seconds), we can relate the frequency resolution with total data length as follows.

Analog Frequency Resolution. Given a continuous time signal \( x(t) \) with \( 0 \leq t \leq T_{\text{MAX}} \), regardless of its sampling frequency, using the DFT we can resolve two frequencies \( F_1, F_2 \) provided

\[
|F_1 - F_2| > \frac{2}{T_{\text{MAX}}}
\]

We can see this with a simple example.

Example. Consider a signal of length \( T_{\text{MAX}} = 2\text{ sec} \). Then, using the DFT, we can resolve two frequencies provided they are separate by at least \( \Delta F = 2/(2 \times 10^{-3}) = 1,000 \text{ Hz} \).

So we can see that the limitation in the frequency resolution is not the sampling frequency, but the actual data length. For the example above, no matter how fast we sample, the resolution will always be the same.

This does not mean that the sampling frequency is irrelevant. Recall that the maximum frequency we can see is, after sampling, half the sampling frequency \( F_s / 2 \) which shows that \( F_s \) always has to be larger than twice the bandwidth of the signal.

Example. We have a signal with bandwidth \( B = 10kHz \) and we want to estimate its frequency spectrum with a resolution \( |\Delta F| \leq 100Hz \). How do we choose the sampling frequency and what is the minimum number of data points? From the frequency resolution we determine the minimum length of the analog signal, as

\[
T_{\text{MAX}} > \frac{2}{|\Delta F|} \geq 0.02 \text{ sec}
\]
From the bandwidth of the signal we determine the requirement on the sampling frequency

\[ F_s > 2 \times B = 20kHz \]

Since \( T_{\text{MAX}} = N \times T_S = N / F_S \), we obtain

\[ N = F_s \times T_{\text{MAX}} > (20 \times 10^3)(0.02) = 400 \]

so we need at least 400 samples to obtain the desired resolution in the frequency spectrum.

The bottom line, is that, to improve the resolution in frequency we need to take a longer data set.

4. Sidelobes and Windowing

VIDEO: Frequency Artifacts (10:03)
http://faculty.nps.edu/rcristi/eo3404/b-discrete-fourier-transform/videos/chapter2-seg2_media/chapter2-seg2-1.wmv

The other issue is the presence of sidelobes. As stated before, these are artifacts due to the fact that the data has a finite length. However, as we see from the figure below, which shows cases of different data length, it does not improve with increasing data length.

![Plot of \( 20 \log_{10}(\left| W_N(\omega) \right| / N) \) around the origin, for various values of the data length \( N \).](image)

Although in each case (\( N = 64, 256, 1024 \)) the amplitudes of the sidelobes decrease as the frequency increases, an important fact to notice is that the largest sidelobes (next to the mainlobe) have the same amplitude regardless the data length \( N \). It is easy to
check from the figure that the attenuation with respect to the peak is about 13dB’s. This is illustrated in the figure below and, again, it is independent of the data length $N$.

![Figure showing attenuation with respect to the peak](image)

**The largest sidelobes are 13dB lower than the main lobe, regardless the data length.**

The reason why we are concerned about the sidelobes is the fact that they show frequencies which do not belong to the signal we are analyzing. In other words the sidelobes in the DFT plot are artifacts due to the fact that we are taking a finite set of data. This can be seen in the following example.

**Example.** Consider the signal $x[n] = 2.0e^{j0.3n} + 0.01e^{j0.4n}$, $n = 0,...,255$, of length $N = 256$ data points. It has two complex exponentials, one of them with much lower energy. The magnitude of its DFT is shown in the figure below. Clearly we see only one peak due to the stronger complex exponential at $\omega = 0.3$ radians, since its sidelobes completely obscures the other weaker component at $\omega = 0.4$ radians.

![DFT magnitude plot](image)

**Magnitude of the DFT of two complex exponentials: the weaker component is buried under the sidelobes of the stronger component.**

Now that we know the problem, we need to find a remedy. Unlike the previous issue of resolution, increasing the data length has no effect on the sidelobes.

The reason at the origin of this problem goes back to the very definition of the DFT, being the same as the Discrete Fourier Series (DFS). When we defined a signal of finite duration we actually looked at it as one period of a period sequence. As shown in the
figure below, when we repeat a signal periodically, at the edges we are likely to have discontinuities, since it is very unlikely that a signal begins and ends with the same value.

![Discontinuity!!!](image1.png)

A signal of finite length as one period of a periodic signal: at the edges it is likely that there discontinuities.

VIDEO: Windowing (06:08)
http://faculty.nps.edu/rcristi/eo3404/b-discrete-fourier-transform/videos/chapter2-seg2_media/chapter2-seg2-2.wmv

The presence of discontinuities causes the high frequency components in the sidelobes. From this argument, a simple remedy is to “smooth the edges” at both ends of the signal, by multiplying by a window sequence, of the same length as the data, which is zero (or close to zero) at the two ends. Then we can take the DFT and observe that the sidelobes are much lower, thus reducing the effects of artifacts in the frequency domain. This is showing in the following figure.

Multiply the data set by a window to reduce discontinuities at the edges.

A number of windows exist, with different properties. In particular the “hamming” window provides a particularly good attenuation of the sidelobes, by lowering to about 40 dB’s. This is a good improvement with the respect to the 13dB attenuation of the sidelobes without any window.
The following example shows the effect of the window.

**Example.** Let us revisit the previous example, with the signal 

$$x[n] = 2.0e^{j0.3n} + 0.01e^{j0.4n}, \ n = 0,...,255.$$ Now compute 

$$X[k] = DFT\{w[n]x[n]\}$$ with 

$$w[n], n = 0,...,255$$ a hamming window of the same length as the data. The plot of its DFT is shown in the figure below, where the peak due to the waker component is can be seen.

![DFT of data with hamming window. The sidelobes are lower and the weaker component now can be detected.](image)

By zooming into the the two peaks, we can determine the indeces of the DFT and therefore the two frequencies.

This is shown in the figure below. The two peaks occur at \( k = 12 \) and \( k = 17 \), which yields the estimates of the two frequencies

\[
\omega_1 \equiv 12 \times \frac{2\pi}{256} = 0.2945\text{rad}
\]

\[
\omega_2 \equiv 17 \times \frac{2\pi}{256} = 0.4172\text{rad}
\]

![Zoom into the plot of the example](image)