Introduction to Spectral Estimation

Yi-Wen Liu

25 March 2013

Musical signals primarily comprise of time-varying sinusoidal components plus noise. To encode musical signals, one may calculate the amplitude, phase, and frequency of all sinusoidal components for each frame (i.e., a windowed block of samples). When doing so, we need to consider the theoretic limitation that prevents us from achieving perfect estimation. There are two related but different concepts: *resolution* vs. *accuracy*.

1 The time-frequency resolution trade-off

1.1 A stationary two-tones complex

Let us consider the case of two tones whose frequencies f_1, f_2 are nearby. For the simplest case, assume that the tones sustain forever and their frequencies never change. We are asked to estimate their frequencies without waiting till the end of time. Now, the discrete-time signal can be written as

$$x[n] = A_1 e^{j(2\pi f_1 nT + \phi_1)} + A_2 e^{j(2\pi f_2 nT + \phi_2)},$$

where $T = 1/f_s$ is the sampling period. From what we've studied previously, we can multiply $x[n]$ with a certain window function $w[n]$, and then look for peaks in the magnitude spectrum of $x[n]w[n]$. The DTFT of $w(n)x(n)$ can be written as

$$X_w(\omega) = A_1 e^{j\phi_1} W(\omega - 2\pi f_1 T) + A_2 e^{j\phi_2} W(\omega - 2\pi f_2 T), \tag{1}$$

where $W(\omega)$ is the DTFT of $w[n]$.

Exercise: If $A_1 = 2A_2$ and $\phi_1 = \phi_2 = 0$, draw a sketch of $X_w(\omega)$ in Eq. 1 assuming $w(n)$ is the rectangular window.

To estimate the frequencies (and amplitudes) of two tones, we can look for the two highest peaks in the magnitude spectrum. Figure 1 shows that the two peaks are well-resolved only if the window is sufficiently long, and the minimum peak-resolving length depends on the type of window — whether it is rectangular or Hann in this case. Further, Fig. 2 shows that how well the two peaks are resolved also depends on the relative phase of the two sinusoids.

1.2 Practical considerations

1.2.1 Choosing the window length

The stationarity assumption is an idealization of audio signals in the real world. Signal characteristics always change in time. Therefore, we are always facing this dilemma when choosing the size of the window: on one hand, a shorter window is preferred in favor of following rapid changes in time. On the other hand, a longer window is preferred in favor of resolving neighboring sinusoids in frequency. This is the classical trade-off in time-frequency analysis. The rule of thumb is: *Resolution in time Δt is inversely proportional to resolution in frequency $\Delta\omega$.*

1.2.2 Spectral splattering

Note that in Fig. 1 we have two magnitude peaks representing two spectral components. Unfortunately, we also have sidelobes extended to the left and to the right. These sidelobes make it difficult to judge how many tones there actually are. This is called *spectral splattering*, meaning that energy is spread spectrally when the signal is windowed in time. The rectangular window has the best spectral resolution but the worst spectral splattering — the height of the sidelobe is just about 13 dB lower than the mainlobe. Other types of windows, such as Hann, Blackman, or the Kaiser family, have better sidelobe suppression ratios. In fact, the Kaiser family of windows are parameterized so that you can fine-tune to trade spectral resolution for sidelobe suppression. However, Kaiser windows do not satisfy the constant overlap-add constraints. More interested readers can refer to [4].

2 The accuracy limit due to the presence of noise

2.1 Modeling a single tone in noise

In certain situations we desire not only to resolve multiple peaks in the spectrum, but also to estimate the frequencies (and amplitudes) of them *accurately*. Let us consider the simplest case first: a single tone in noise.

The problem of tone estimation is stated as follows. We are looking for a tone that might be contaminated by noise. The tone is a pure tone, so we write

$$s(n; A, \omega) = Ae^{j\omega n}, \tag{2}$$

where A is an arbitrary complex number. Were there no noise, the amplitude of this tone would simply be computed as the absolute value at any time n, and the frequency

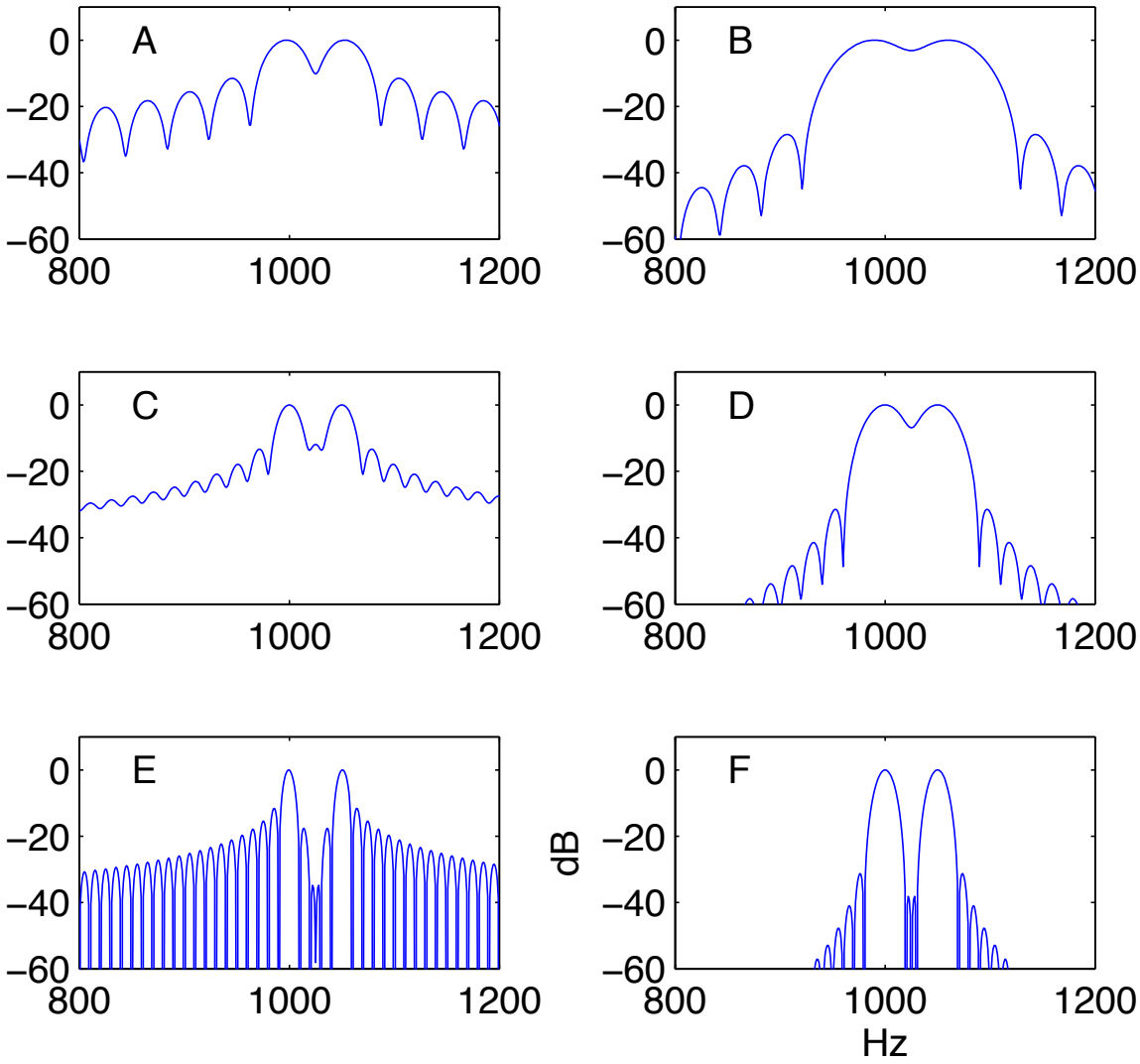

Figure 1: Short vs. long windows in spectral analysis. Panels show summed spectra of two equal-intensity tones at 1000 and 1050 Hz, respectively. **A, B:** window length = 25 ms. **C, D:** 50 ms. **E, F:** 100 ms. Panels A, C, E are obtained using the rectangular window while B, D, F are obtained using the Hann window.

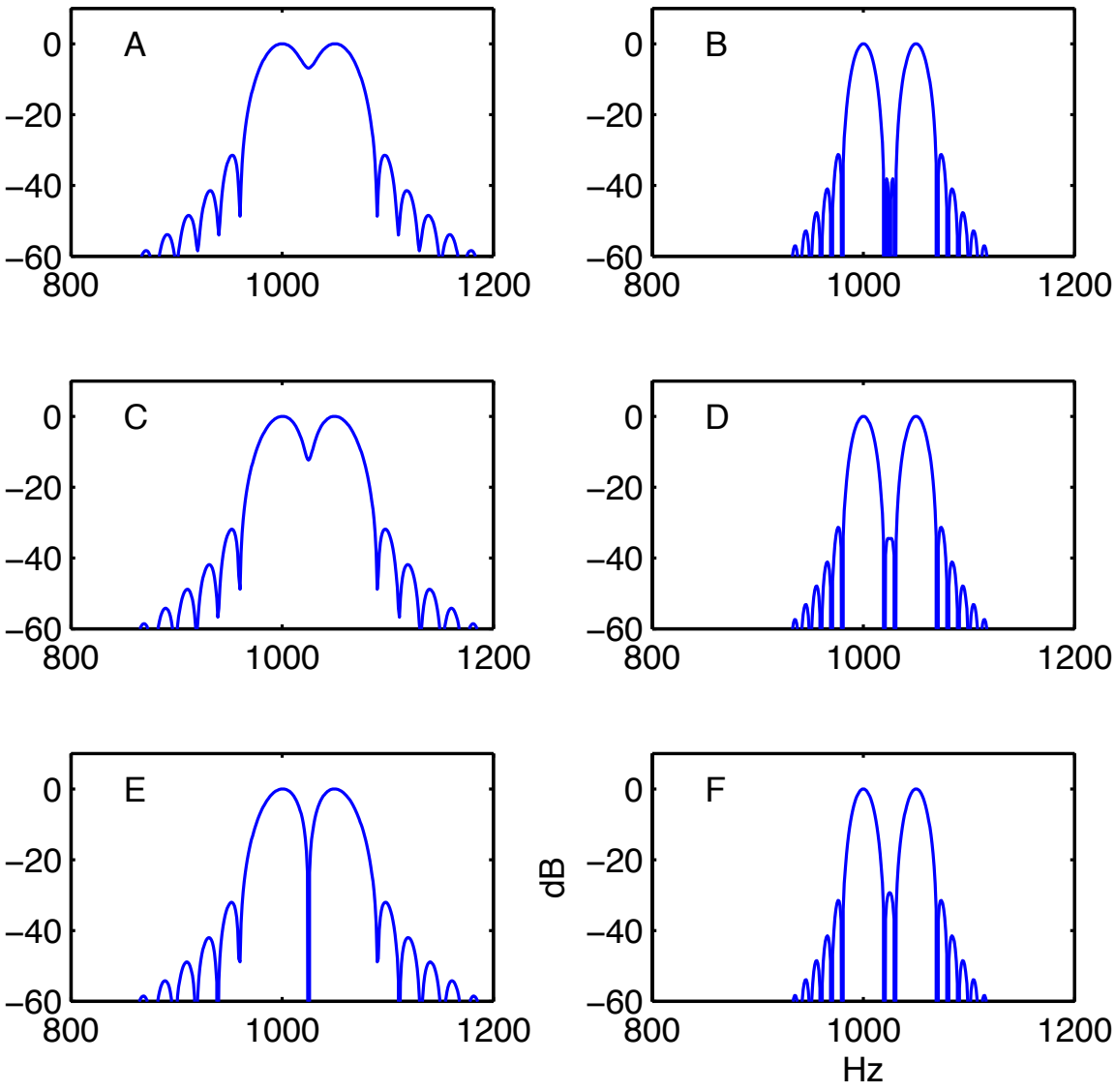

Figure 2: Spectral separation of two tones depends on their relative phase. Using Eq. (1), assume that $f_1 = 1000$ Hz, $f_2 = 1050$ Hz, and $A_1 = A_2$. Panels on the left are obtained by Hann window of length 50 ms, and on the right, 100 ms. Each row was obtained by a different relative phase. **A, B**: $\phi_2 - \phi_1 = 0$. **C, D**: $\phi_2 - \phi_1 = \pi/4$. **E, F**: $\phi_2 - \phi_1 = \pi/2$.

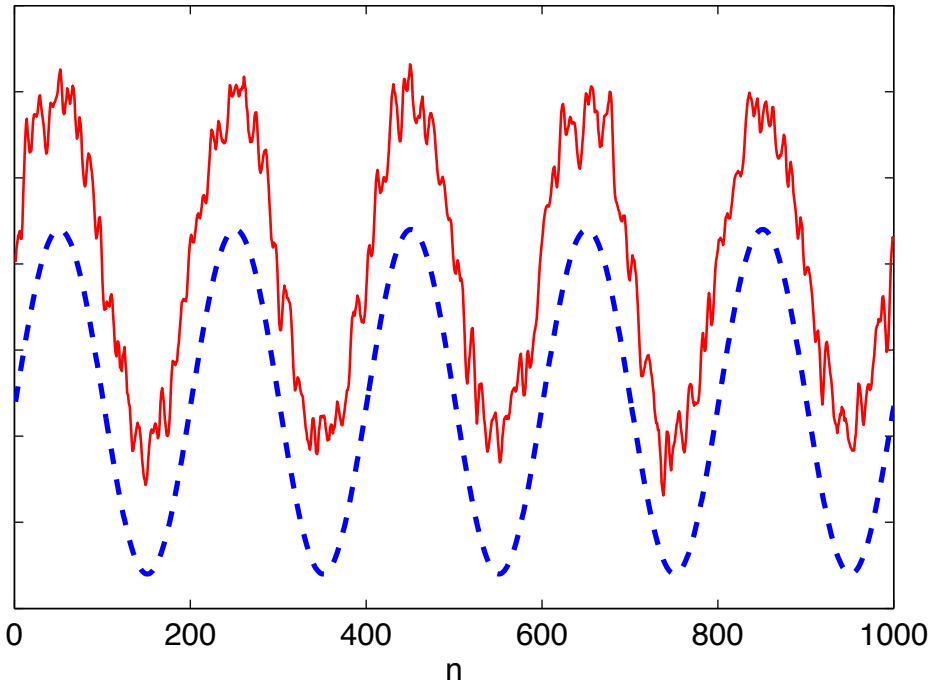

Figure 3: A tone interfered by noise. Red: the observed noisy signal. Blue: the ideal sinusoid to extract.

could be figured out in a straightforward manner (how?). However, when the signal is added with noise, we cannot directly observe $s(n)$. Instead, we might be looking at $x(n) = s(n) + u(n)$, where $u(n)$ is a noise signal. Our task is to estimate its frequency (and amplitude). Figure 3 illustrate the situation we are facing. We look at the noisy signal (in red) and hope to recover a pure signal underlying there (in blue). How should we do it?

To formulate this estimation problem rigorously, we need to assume that the noise obeys certain statistics. A simple example is the Gaussian noise. Let us assume that $u(n)$ is i.i.d.[1] Gaussian with zero mean and variance σ^2. Next section presents a simple method for frequency estimation of a single tone in noise.

2.2 Quadratic interpolation in magnitude spectrum

The QI-FFT (quadratic interpolation of FFT) method involves four steps:

- Windowing
- Zero-padding to about 8–16 times the window length
- FFT
- Quadratic interpolation near the peak of the logarithmic magnitude spectrum

[1] independent identically distributed

We've covered the first three steps in class. To complete the fourth step, we need to identify the maximum in the magnitude spectrum. Let us denote the maximum as $L = 20\log_{10}|X_k|$, i.e., the peak is located at the kth frequency. Then we fit a parabola near the peak (f_k, L); that is, assuming that the peak and the point to its left and to its right are on a parabola $L(f) = -a(f - \hat{f})^2 + L_{\max}$. Then, $L_{\max}$ is an estimate of the tone level (intensity) and $\hat{f}$ is an estimate of the tone frequency.

Exercise: Derive a formula for $\hat{f}$ and $L_{\max}$.

The QI-FFT method is intuitive. To evaluate its performance, we need to compare against a limit in theory. The limit is the Cramér-Rao lower bound (CRB).

2.3 Fisher information and the Cramér-Rao bound (CRB)

Assume that we know the signal comes from a family of signals and the family of signal is parameterized by a real-valued index θ. For example, the amplitude A is known, then Eq. (2) describes a family of signal parameterized by frequency ω. Assume that the signal is contaminated with noise of known statistics; i.e., $x(n) = s(n) + u(n)$ and we know the multi-variate probability density function of $u(n)$. Also, assume that we observe $x(n)$ from time $n = 0$ to $N - 1$. Define

$$V = \frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta) = \frac{\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}$$

where $f(\mathbf{x};\theta)$ is the probability density function of $\mathbf{x} = [x(0), x(1), ..., x(n-1)]^T$.

The *Fisher information* $J(\theta)$ is defined as the variance of V [2]:

$$J(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta)\right]^2$$

It can be shown that, for any *unbiased* estimator $\hat{\theta} = T(\mathbf{x})$, the mean-square error (MSE) of parameter estimation is lower bounded,

$$E(\hat{\theta} - \theta)^2 \geq \frac{1}{J(\theta)}. \tag{3}$$

For each θ, the quantity $1/J(\theta)$ is called the CRB. It gives a theoretic limit for any method that estimates θ unbiasedly.

Under this framework, it can be derived that the Fisher information of frequency estimation is proportional to N^3A^2/σ^2 [3, 5]. Note that A^2/σ^2 is the signal to noise ratio, and the fact that the CRB decreases by the factor N^3 tells us that the *accuracy* of frequency estimation improves beyond linearly in time.

To compare, the frequency *resolution* of FFT improves linearly with respect to time. It is important to tell the difference. Often, a frequency estimating scheme that achieve the time dependence of mean-square error at N^{-3} is called a "coherent frequency estimator".

It has been shown that the QI-FFT method achieves coherent estimation for a reasonable range of SNR (say SNR $\geq$ 5 dB) at a reasonably chosen window length. More interested reader can refer to [1] for details of performance analysis.

References

[1] M. Abe and Julius O. Smith. Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of fft magnitude peaks. In *AES 117th Convention*, San Francisco, Oct. 2004.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[3] Yi-Wen Liu and III Smith, J.O. Watermarking parametric representations for synthetic audio. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages 660–3, April 2003.

[4] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing, 3rd Ed.* Pearson, New York, 2010.

[5] B. Porat. *Digital Processing of Random Signals: Theory and Methods*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1994.