Supplementary Notes for ELEN 4810 Lecture 6
Windowing and the Short-Time Fourier Transform

John Wright
Columbia University

October 20, 2025

Disclaimer: These notes are intended to be an accessible introduction to the subject, with no pretense at completeness. In general, you can find more thorough discussions in Oppenheim's book. Please let me know if you find any typos.

Reading suggestions: Oppenheim-Schafer Chapter 10

In this lecture, we will first finish material from Lecture 5 on DFT and its fast computation via FFT. Time permitting, we will then move on to discuss spectral analysis based on finite time windows, and spectral analysis for signals with time-varying frequency content, i.e., the content of this lecture note.

With efficient tools for computing the Fourier transform in hand, in this lecture we briefly describe issues that arise when we try to use this tool to analyze signals. We will discuss two major issues. The first is the effect of only collecting a finite number of samples $x[0], \ldots, x[L-1]$, rather than observing $x[n]$ for all n. The second is what happens when the signal of interest has frequency content that varies with time.

1 Windowing

Consider a continuous time signal $x_c(t)$. We obtain samples $x[n] = x_c(nT)$, and wish to infer something about the frequency content of the original signal $x_c(t)$. From the sampling theorem, we know that if x_c is bandlimited, the sequence $x[n]$ contains all the information needed to reconstruct $x_c(t)$. For example, if our signal is a superposition of two sinusoids

$$x_c(t) = \cos(\Omega_0 t) + \cos(\Omega_1 t), \tag{1.1}$$

we know that the Fourier transform

$$X_c(j\Omega) = \pi\delta(\Omega - \Omega_0) + \pi\delta(\Omega + \Omega_0) + \pi\delta(\Omega - \Omega_1) + \pi\delta(\Omega + \Omega_1) \tag{1.2}$$

is a superposition of four spikes. Moreover, $x[n] = \cos(\omega_0 n) + \cos(\omega_1 n)$, with $\omega_i = \Omega_i T$; assuming that $T < \pi / \max\{|\Omega_0|, |\Omega_1|\}$, there is no aliasing, and so $X(e^{j\omega})$ contains four spikes in the interval

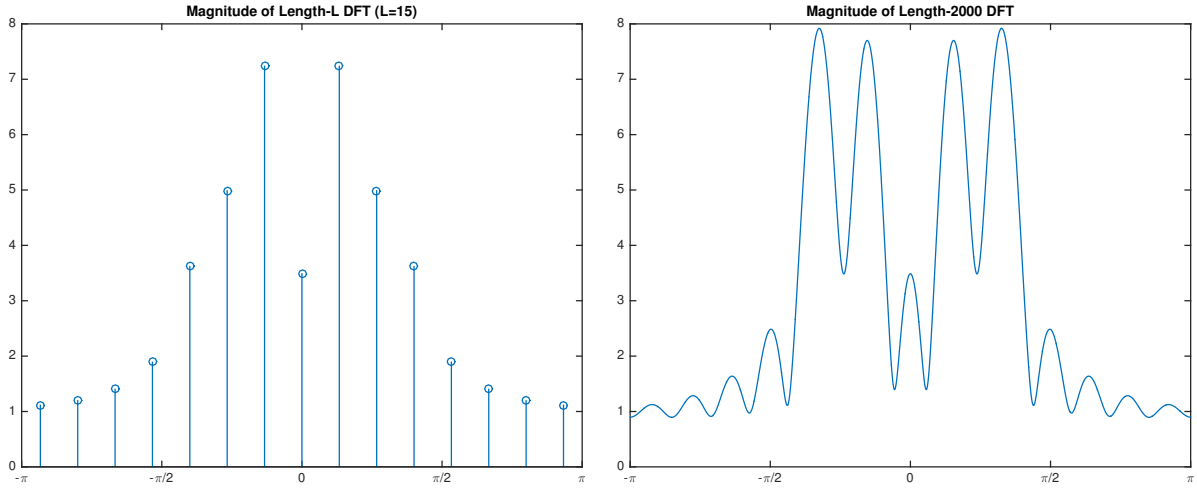

Figure 1: **Spectral analysis of a signal, Part I.** Starting from $x[n] = \cos(\pi n/3.3) + \cos(\pi n/5.4)$, we extract the first $L = 15$ samples, $\bar{x}[n] = x[n]$, $n = 0, \ldots, 14$. At left, we display the magnitude of the length-L DFT of this sequence. Although $X(e^{j\omega})$ contains four Dirac δ measures in the interval $(-\pi, \pi]$, this is not obvious from the DFT of $\bar{x}$. At right, we plot the length-N DFT of $\bar{x}$ for $N = 2000$.

$-\pi < \omega \leq \pi$. In principle, we can identify the frequency Ω_0 simply by looking for the spikes in the DTFT $X(e^{j\omega})$.

However, in practice there is another challenge: *we can only sample for a finite length of time*. We have to start sampling at some time, and then stop sampling at some other time! So, we cannot actually observe the entire sequence $x[n]$. For simplicity, suppose assume that we start sampling at $n = 0$, and obtain samples until $n = L - 1$, so the total number of samples is L. Let

$$\bar{x}[n] = \begin{cases} x[n] & n = 0, \ldots, L-1, \\ 0 & \text{else.} \end{cases} \tag{1.3}$$

That is to say, $\bar{x}$ contains the samples that we have observed, with zeros filled in everywhere else.

Since $\bar{x}$ is a length-L signal, we can apply the length-L Discrete Fourier Transform. Figure 1 shows this, for the particular choices of $\omega_0 = \pi/3.3$, $\omega_1 = \pi/5.4$, and $L = 15$. In Figure 1, you may notice a strange phenomenon: although we started out with two pure tones (and four spikes in frequency domain), the DFT $\bar{X}[k]$ is nonzero everywhere. Looking at the picture, it is quite challenging to tell that the original signal consisted of two tones!

What happened? There are essentially two phenomena that we have to be careful about here: *sampling in frequency* and *windowing in time*. The first effect arises because $\mathrm{DFT}_N\{x\}$ samples the DTFT $X(e^{j\omega})$:

$$\mathrm{DFT}_N\{\bar{x}\}[k] = \bar{X}(e^{j\omega})\big|_{\omega = \frac{2\pi k}{N}} \tag{1.4}$$

Although the length-L DFT of $\bar{x}$ contains all of the information needed to reconstruct $\bar{x}$, if our goal is to understand the properties of $\bar{X}(e^{j\omega})$, it is better to compute the length N DFT of $\bar{x}$, for some larger $N \gg L$. This produces denser samples in frequency domain. Figure 1 (right) shows an approximation to the DTFT of $\bar{x}$, computed using the length-N DFT with $N = 2,000 \gg L$.

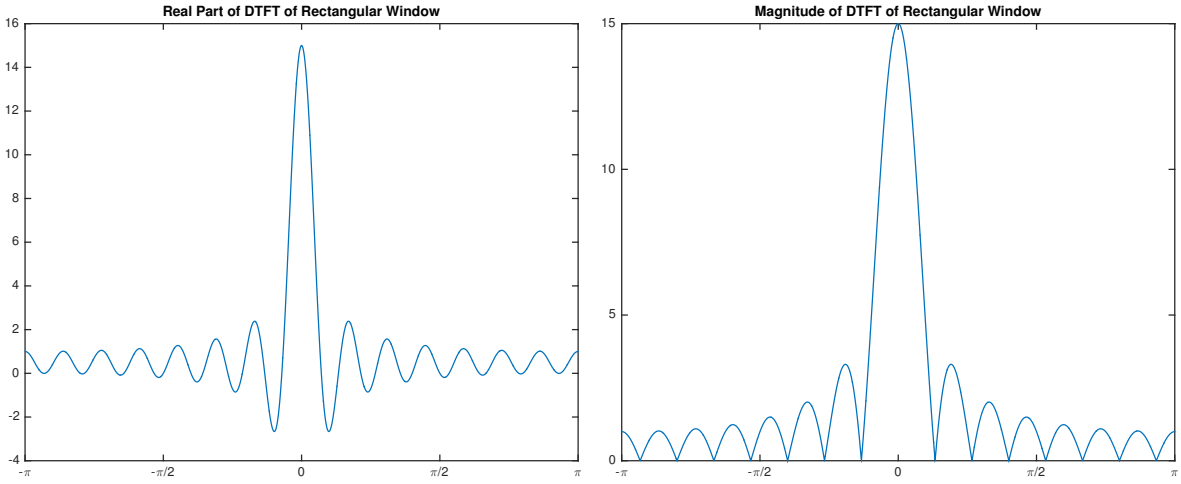

Figure 2: **DTFT for a Rectangular Window of Length $L = 15$.** Left: Real part of the DTFT. Right: DTFT magnitude. The DTFT magnitude exhibits a *mainlobe* about $\omega = 0$, and a number of *sidelobes* of slowly decreasing magnitude.

Figure 1 (right) displays a much denser sampling of the DTFT in frequency. However, it is still very difficult to tell what is going on – it is certainly not obvious that the original signal consisted of two pure tones! To understand why this is the case, we need to relate the DTFT of $\bar{x}$ to the DTFT of x. We can view $\bar{x}$ as a *windowed* version of x. Define a *window*

$$w[n] = \begin{cases} 1 & 0 \leq n \leq L-1, \\ 0 & \text{else} \end{cases} \tag{1.5}$$

We can write

$$\bar{x}[n] = w[n]x[n]. \tag{1.6}$$

That is to say, the finite length signal $\bar{x}$ is a windowed version of the original, infinite-length signal x. We know that multiplication in the time domain becomes convolution in the Fourier domain, and so

$$\bar{X}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\theta}) X(e^{j(\omega-\theta)}) d\theta. \tag{1.7}$$

The DTFT of the window $w[n]$ is a *periodic sinc*

$$W(e^{j\omega}) = \frac{\sin(\omega L/2)}{\sin(\omega/2)} e^{-j\omega(L-1)/2}. \tag{1.8}$$

Figure 2 shows the real part and magnitude of $W(e^{j\omega})$ in the principal interval $(-\pi, \pi]$.

Multiplication with w in time is equivalent to convolution with $W(e^{j\omega})$ in frequency. The effect of this convolution can be understood in terms of the magnitude response in Figure 2 (right). The magnitude response exhibits a *mainlobe* about $\omega = 0$, and a number of *sidelobes*.

(i) **Loss of resolution due to mainlobe width.** The effect of the mainlobe is a loss of resolution in frequency – even if $X(e^{j\omega})$ is very concentrated about some frequency ω_0, convolving with $W(e^{j\omega})$ will spread it out. This effect is different from the effect of sampling in frequency in the DFT, which can be mitigated by increasing the number of DFT samples, N. The loss of resolution due to the mainlobe width is a direct consequence of windowing in time. It cannot be completely eliminated, but it can be mitigated by increasing the window length L. Indeed, for the rectangular window w of length L, the width of the mainlobe is $4\pi/L$, which diminishes as L increases.

(ii) **Spurious peaks due to sidelobes.** The effect of the sidelobes is to crease "spurious" peaks in $\bar{X}(e^{j\omega})$ which occur at frequencies ω which may not correspond to sinusoidal components that are actually present in the original signal $x[n]$. These spurious peaks can be observed in Figure 1 (right). They can be viewed as a consequence of the slow decay of the magnitude response of $W(e^{j\omega})$: $w[n]$ transitions abruptly from 1 to 0 at $n = L$. This abrupt transition creates high frequency components.

We can substantially reduce the effect of the sidelobes, by choosing a window which tapers gradually toward zero at the endpoints $n = 0$ and $n = L - 1$. For example, we could choose a Hamming window

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n < L \\ 0 & \text{else.} \end{cases} \tag{1.9}$$

Figure 3 visualizes this window in both the time domain and frequency domain, in comparison to the rectangular window. Notice that the sidelobes are substantially reduced compared to the rectangular window. However, the mainlobe is somewhat wider. Indeed, the Hamming window has a mainlobe width of about $8\pi/(L-1)$ – almost twice as wide as the mainlobe of the rectangular window. In practice, this is often an acceptable price to pay for reducing the sidelobes.

Figure 4 shows an analysis of the same sinusoidal signal $x[n] = \cos(\pi n/3.3) + \cos(\pi n/5.4)$ with a Hamming window of length $L = 15$, in comparison with a rectangular window of the same length. Because the Hamming window has smaller sidelobes, there are fewer spurious frequency components. Figure 5 examines the effect of the window length. As L increases, both windows exhibit a narrower mainlobe, and hence improved spectral resolution.

The Hamming window is just one of many windows $w[n]$ which tradeoff between mainlobe width and sidelobe height in useful ways. The text describes several other families of windows. When we discuss FIR filter design, we will encounter the Kaiser windows, which are a parametric family of windows that allow use to adjust this tradeoff to meet a desired specification.

2 The Short-Time Fourier Transform: Motivation and Definitions

The *Short-Time Fourier Transform* (*STFT*),[1] is a basic tool for spectral analysis of time-varying signals. The STFT is widely used in audio and communications applications. A canonical application is in analyzing music or speech, where the signal may contain different frequencies (notes) at different times.

[1]The text calls this the "time-dependent Fourier transform" (c.f., Section 10.3).

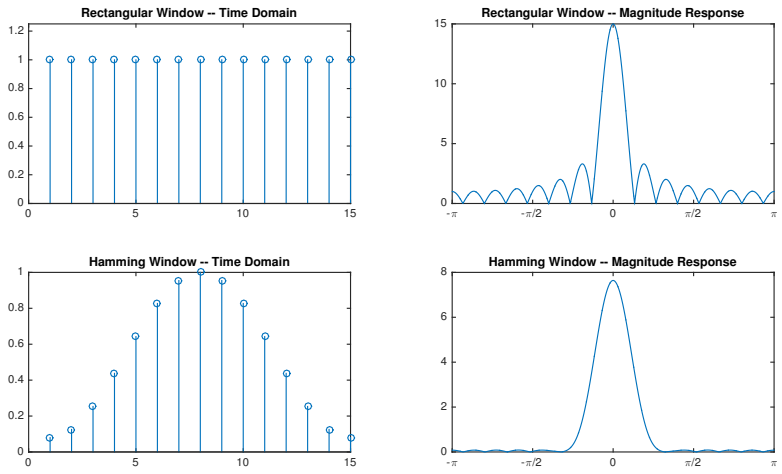

Figure 3: **Rectangular and Hamming windows.** Top: rectangular window of length $L = 15$ in both time and frequency domain. Bottom: Hamming window of length $L = 15$. Notice that in time domain, the Hamming window tapers much more gradually to zero. In frequency domain, this results in smaller sidelobes, at the expense of a somewhat wider mainlobe.

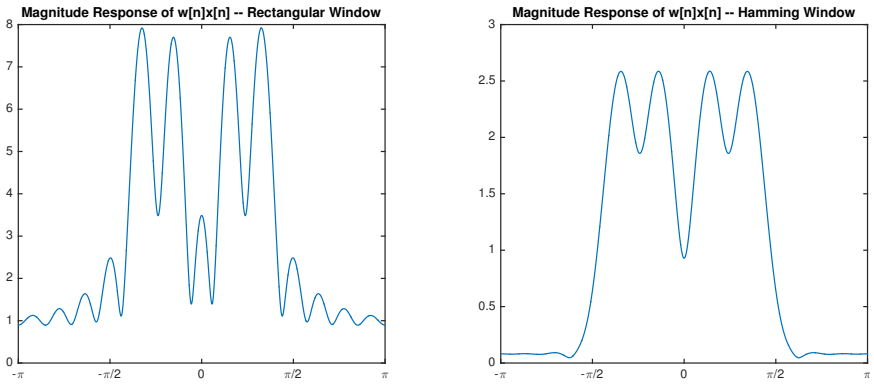

Figure 4: **Spectral analysis with rectangular and Hamming windows.** Left: magnitude response of $w[n]x[n]$ with w a rectangular window. Right: magnitude response of $w[n]x[n]$ with w a Hamming window.

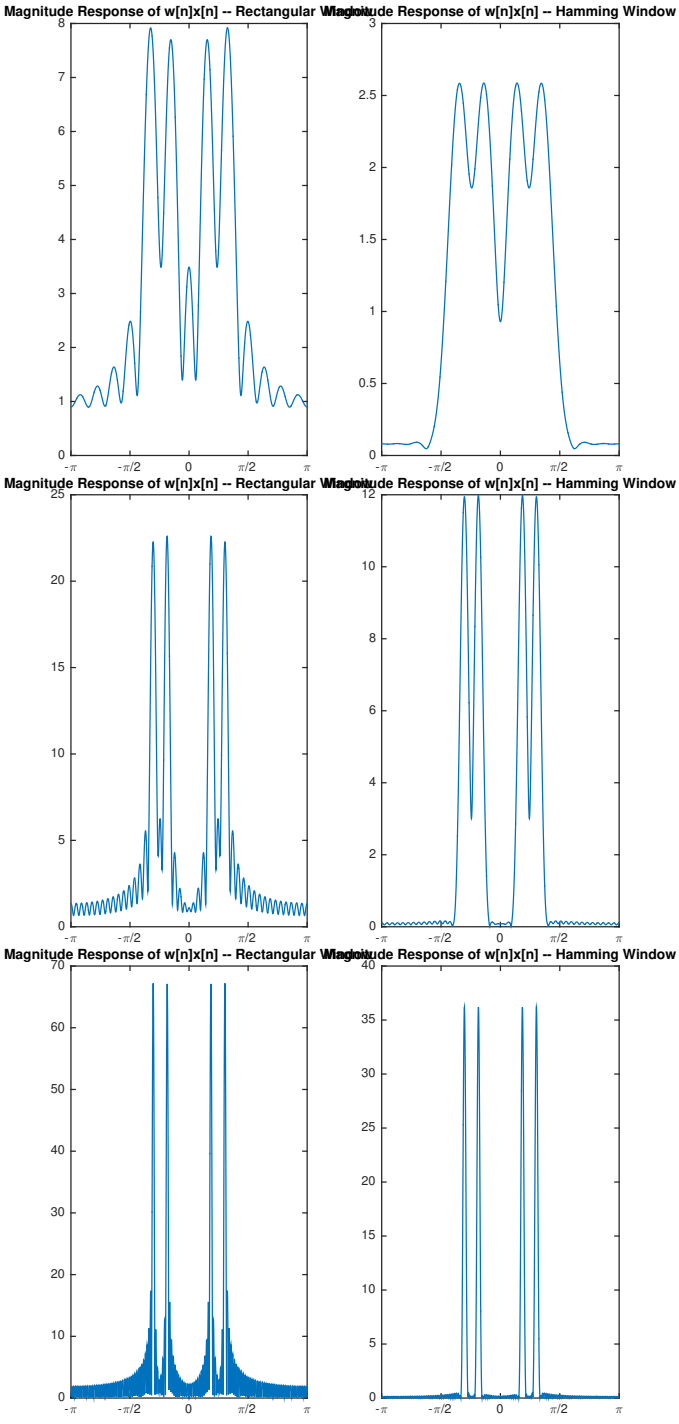

Figure 5: **Effect of window length.** Left: Rectangular window. Right: Hamming window. Top: $L = 15$. Middle: $L = 45$. Bottom: $L = 135$.

The approach is simple: we use a moving window to select different finite-length blocks of the signal, and for each block, compute its DTFT. Let $x[n]$ denote the signal of interest, and $w[m]$ the windowing function. The STFT is a complex-valued function of two variables – a time $n \in \mathbb{Z}$ and a frequency $\lambda \in \mathbb{R}$ – defined by the relationship

$$\bar{X}[n, \lambda) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m}. \tag{2.1}$$

The peculiar notation $\bar{X}[n, \lambda)$ is intended to remind us that n is a discrete variable, while λ is continuous. The STFT has the following interpretation. Suppose that $w[m]$ has length L, and is supported on $0, \ldots, L-1$. $\bar{X}[n, \lambda)$ is produced by first shifting the signal x to the left by n samples, then multiplying by w, and then taking the discrete-time Fourier transform. Thus, $X[n, \lambda)$ describes the "frequency content" of $x[n], x[n+1], \ldots, x[n+L-1]$. The window $w[m]$ could simply be chosen as a rectangular window

$$w[m] = \begin{cases} 1 & 0 \leq m \leq L-1, \\ 0 & \text{else}, \end{cases} \tag{2.2}$$

in which case the STFT can be interpreted as taking the DTFT of (shifted) pieces of the signal. Unfortunately, the rectangular window has all the same downsides that we saw above, and so more popular (and effective) choices include the Bartlett, Hann and Hamming windows, all of which taper towards zero at the edges of the interval $0 \leq m \leq L-1$. We will discuss the tradeoffs involved in windowing in more detail below.

Sampling in frequency. The STFT described in (2.1) is a useful conceptual tool, but it is not directly computable – it is defined over a continuous frequency variable λ. To obtain a more practical representation, we can sample the STFT at N frequencies

$$0, \frac{2\pi}{N}, \frac{2\pi \times 2}{N}, \ldots, \frac{2\pi(N-1)}{N}, \tag{2.3}$$

Write

$$\begin{aligned}
\bar{X}[n, k] &= \bar{X}[n, \lambda)\Big|_{\lambda = \frac{2\pi k}{N}} && (2.4) \\
&= \sum_{m=0}^{L-1} x[n+m]w[m] \exp\left(-j\frac{2\pi km}{N}\right) && (2.5) \\
&= \sum_{m=0}^{N-1} x[n+m]w[m] \exp\left(-j\frac{2\pi km}{N}\right) \qquad \text{Assuming } N \geq L. && (2.6) \\
&= \mathrm{DFT}_N\left\{x[n+\cdot]w[\cdot]\right\}[k]. && (2.7)
\end{aligned}$$

The sampled version $\bar{X}[n, k]$ is defined over two discrete variables – n, which ranges over the integers, and k, which ranges from 0 to $N-1$. The above calculation makes it clear that $\bar{X}[n, k]$ is just the N-point DFT of the shifted and windowed sequence $x[n+m]w[m]$. For this interpretation to make sense, it is crucial that the number of sample points N be chosen to be at least as large as the window length L.

Sampling in time. The STFT $\bar{X}[n,k]$ can be computed using the Fast Fourier Transform algorithm. Typically, to reduce the computational burden, we can also subsample in the time dimension, by choosing a step length $R \in \mathbb{Z}_+$, and setting

$$\begin{aligned} X[r,k] &= \bar{X}[rR,k] && (2.8)\\ &= \sum_{m=0}^{N-1} x[rR+m]w[m]\exp\left(-j\frac{2\pi km}{N}\right) && (2.9)\\ &= \mathrm{DFT}_N\left\{x[rR+\cdot]w[\cdot]\right\}[k]. && (2.10)\end{aligned}$$

That is to say, to compute the r,k entry we shift the signal x to the left by rR samples, window, and then compute the DFT. $X[r,k]$ can be computed in Matlab using the `spectrogram`[2] command.[3]

3 Filter bank interpretation

Set

$$h_k[n] = w[-n]\exp\left(j\frac{2\pi kn}{N}\right). \tag{3.1}$$

Notice that

$$\begin{aligned} h_k * x[n] &= \sum_{m=-\infty}^{\infty} x[n-m]h_k[m] && (3.2)\\ &= \sum_{m=-\infty}^{\infty} x[n-m]w[-m]\exp\left(j\frac{2\pi km}{N}\right) && (3.3)\\ &= \sum_{m=-\infty}^{\infty} x[n+m]w[m]\exp\left(-j\frac{2\pi km}{N}\right) && (3.4)\\ &= \bar{X}[n,k] && (3.5)\end{aligned}$$

Thus, $\bar{X}[\cdot,k]$ can be viewed as the result of convolving the input x with a filter h_k which consists of a windowed complex exponential. The entire transform $\bar{X}[n,k]$ can be viewed as convolving $x[n]$ with a bank of N filters $h_0, h_1, \ldots, h_{N-1}$. Figure 6 plots the real part of $h_k[n]$ for several choices of k.

4 Windowing, and some fundamental limits

The filter bank interpretation in the previous section reveals an interesting tradeoff in constructing $\bar{X}[n,k]$. The filter $h_k[n]$ is a product of two terms – $w[-n]$, which restricts its spatial support to a finite interval, and $\exp\left(j\frac{2\pi kn}{N}\right)$, which is localized in frequency. The resulting filter $h_k[n]$ is

[2] In typical usage, the term *spectrogram* refers to the squared magnitude of the short-time Fourier transform – $|X[r,k]|^2$. The Matlab function `spectrogram` produces the short-time Fourier transform, which is complex valued and contains phase information in addition to the magnitude.

[3] **Notation.** The notation $\bar{X}$ is not standard – the text uses X for $\bar{X}[n,\lambda)$, $\bar{X}[n,k]$, and $X[r,k]$. I have reserved the notation X for $X[r,k]$, the object typically used in practical computations.

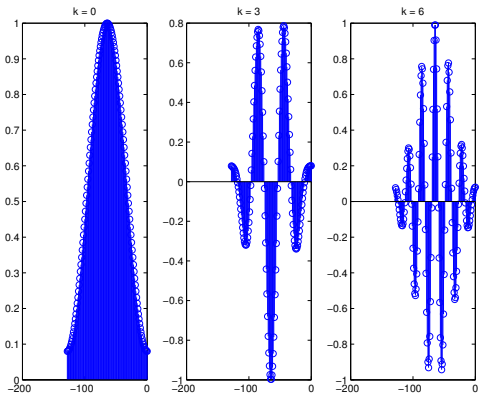

Figure 6: **Filterbank interpretation of the STFT.** Real part of the impulse response $h_k[n]$ for $k = 0, 3, 6$, with $L = 129$, $N = 129$, and $w[n]$ a Hamming window. The filters h_k consist of windowed complex exponentials, which are localized in both time and frequency.

thus localized in both space and frequency – a highly desirable property for frequency analysis of time-varying signals. Unfortunately, these two effects are in competition. As the window length L becomes shorter and shorter, $h_k[\cdot]$ becomes less concentrated around the target frequency $\lambda_k = \frac{2\pi k}{N}$. Conversely, as the window length becomes longer, $h_k[\cdot]$ becomes more concentrated in frequency, but less concentrated in time. This is a qualitative version of an *uncertainty principle*, which states that a function cannot be simultaneously localized in both frequency and time. The best-known uncertainty principles apply in continuous time.[4] However, it is possible to give discrete-time analogues. Figure 7 illustrates these tradeoffs for the special example of a windowed complex exponential – depending on the window size, we can either concentrate the resulting filter in time (short window) or in frequency (long window), but not both.

On a more practical level, for the STFT, it is important to use a window that has the best possible spectral properties. Even if the input is a pure sinusoid, the effect of windowing is to produce *spectral leakage*, in which the STFT of a pure sinusoid is diffused into multiple frequency bins, complicating the problem of detecting the active frequency.

[4] A function $f(t) \in L^2$ with Fourier transform $F(j\Omega)$ cannot be simultaneously concentrated in both time and frequency: if

$$\mu_t \doteq \frac{1}{\int_t |f(t)|^2 dt} \int_t t|f(t)|^2 dt, \tag{4.1}$$

$$\sigma_t^2 \doteq \frac{1}{\int_t |f(t)|^2 dt} \int_t (t-\mu_t)^2 |f(t)|^2 dt, \tag{4.2}$$

$$\mu_f \doteq \frac{1}{\int_\Omega |F(j\Omega)|^2 d\Omega} \int_\Omega \Omega |F(j\Omega)|^2 d\Omega \tag{4.3}$$

$$\sigma_\Omega^2 \doteq \frac{1}{\int_\Omega |F(j\Omega)|^2 d\Omega} \int_\Omega (\Omega-\mu_\Omega)^2 |F(j\Omega)|^2 d\Omega \tag{4.4}$$

then $\sigma_t^2 \sigma_f^2 \geq \frac{1}{4}$.

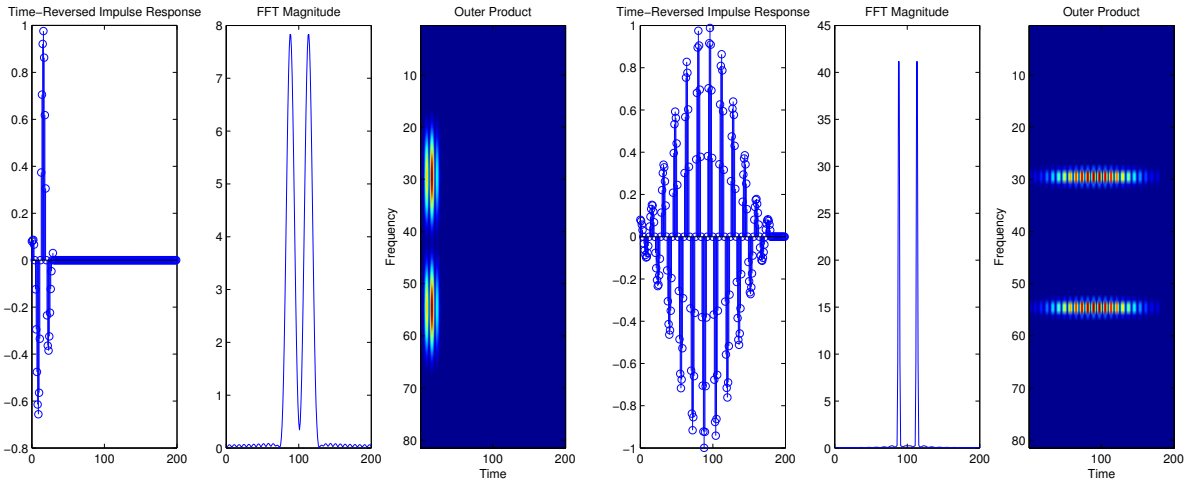

Figure 7: **Uncertainty in time-frequency localization**

5 Invertibility and practical inversion

In this section, we discuss issues around the inversion of the short-time Fourier transform. The STFT can be viewed as a linear map[5] $\mathcal{S}$, which takes a sequence $x \in \mathbb{C}^{\mathbb{Z}}$ to $X \in \mathbb{C}^{N \times \mathbb{Z}}$:

$$X = \mathcal{S}\{x\}. \tag{5.1}$$

Definition 5.1. *We say that the map $\mathcal{S}$ is invertible, if there is another mapping $\breve{\mathcal{S}} : \mathbb{C}^{N \times \mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}}$, which satisfies $\breve{\mathcal{S}}\{X\} = x$ whenever $X = \mathcal{S}\{x\}$.*

Depending on the window length L, the number of frequency samples N, and time step R, the STFT may not be invertible. In particular, if $R > L$, it is clear that the transform is not invertible, since there are some samples $x[n]$ that are not used in producing $X[r, k]$. However, we will see that under very mild circumstances, provided $R \leq L \leq N$, the transform *is* invertible.

After establishing this, we will talk about how to choose and implement the map $\breve{S}$. We will see that when $R < L$, the inverse map $\breve{S}$ is not unique. The reason for this non uniqueness is simple: *if $R < L$, there are many X which are not the transformation of any sequence x.* In practice, if we apply any nontrivial processing to X, we will likely produce an object which does not satisfy $X = \mathcal{S}\{x\}$ for any sequence x. We would like an inverse mapping $\breve{S}$ which still behaves sensibly in this situation. We will discuss two methods for doing this – the overlap and add method, and least squares inversion.

Invertibility. We begin by noting that transformation $x \mapsto \bar{X}[n, \lambda)$ is invertible, provided at least one entry of the window w is nonzero. For this, we simply note that the inverse DTFT formula gives

$$x[n+m]w[m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \bar{X}[n, \lambda) \exp(j\lambda m)\, d\lambda. \tag{5.2}$$

[5]The linearity of the STFT is straightforward to check.

Hence, provided $w[m] \neq 0$, this formula gives $x[n+m]$ as a function of $\bar{X}[\cdot,\cdot]$; by varying n we obtain the entire sequence $x[n]$.

A virtually identical argument shows that the mapping $x \mapsto \bar{X}[n,k]$ is invertible whenever $N \geq L$. Indeed, in this situation, we can obtain $x[n+m]w[m]$ by applying the inverse DFT to $\bar{X}[n,\cdot]$:

$$x[n+m]w[m] = \frac{1}{N}\sum_{k=0}^{N-1} \bar{X}[n,k]\exp\left(j\frac{2\pi km}{N}\right). \tag{5.3}$$

Again, provided $w[m] \neq 0$, this demonstrates $x[n+m]$ as a function of $\bar{X}[n,k]$. In fact, if $w[m] \neq 0$ for $0 \leq m \leq L-1$, we can obtain $x[n], x[n+1], \ldots, x[n+L-1]$ from the inverse DFT of $\bar{X}[n,\cdot]$.

If we consider the time-sampled version $X[r,k]$, by exactly the same argument we can obtain $x[rR], x[rR+1], \ldots, x[rR+L-1]$ from $X[r,\cdot]$. We can obtain $x[(r+1)R], x[(r+1)R+1], \ldots, x[(r+1)R+L-1]$ from $X[r+1,\cdot]$. Provided $R \leq L$, every sample $x[n]$ is contained in at least one of these reconstructed windows, and hence the sequence $x[n]$ can be recovered from the transformation $X[r,k]$:

Theorem 5.2. *If $R \leq L \leq N$, and $w[m]$ is nonzero for $m = 0, 1, \ldots, L-1$, then the short-time Fourier transform $x[n] \mapsto X[r,k]$ is invertible.*

Practical inversion: Problems with the naive approach. The procedure described in above proves that the STFT is invertible. It suggests a natural way of implementing the inverse: set

$$\hat{x}_r[n] = \begin{cases} \frac{1}{Nw[n-rR]}\sum_{k=0}^{N-1} X[r,k]\exp\left(j\frac{2\pi k(n-rR)}{N}\right) & rR \leq n \leq rR+L-1 \\ 0 & \text{else} \end{cases} \tag{5.4}$$

and

$$\hat{x}[n] = \hat{x}_r[n], \tag{5.5}$$

for some r such that $rR \leq n \leq rR+L-1$. If $R \leq L$, there exists at least one choice of r for which these inequalities are satisfied. This simply corresponds to choosing one of the windowed versions which contains the n-th sample. It is not difficult to show that if $R = L$, the inverse mapping is unique, and the (only) inverse is given by (5.4).

Unfortunately, (5.4) has limited practical value. To minimize the effect of spectral leakage, we usually choose a window $w[\cdot]$ which tapers near its edges, with $w[0]$ and $w[L-1]$ very close to zero. Small changes to $X[r,k]$ can produce extremely large changes in the reconstruction $\hat{x}_r[n]$. Since the main point of moving to the STFT domain is so that we can modify $X[r,k]$ in some useful way, this property is undesirable – the reconstruction is unstable, and the reconstructed x may bear little resemblance to a physically realizable signal.

If $R = L$, we are stuck using (5.4). However, if $R < L$, there are many choices of the inverse map $\breve{S}$, which satisfy the exact reconstruction property $\breve{S}\{S\{x\}\} = x \;\forall\, x$, but respond much more gracefully when X is *not* the image $S\{x\}$ of some signal x.

Practical inversion (I): Overlap and add. A classical approach is known as *overlap-and-add* (OLA). In this approach, we simply use the inverse DTFT to construct the windowed versions:

$$x_r[m] = x[rR+m]w[m] \qquad -\infty < m < \infty \tag{5.6}$$

and then add them up:

$$\hat{x}_{OLA}[n] \quad = \quad \sum_r x_r[n - rR]. \tag{5.7}$$

Does this approach produce an inverse map? Plugging in the definition of $x_r[\cdot]$, we obtain

$$\hat{x}_{OLA}[n] \quad = \quad \sum_r x[n]w[n - Rr] \tag{5.8}$$

$$= \quad x[n] \sum_r w[n - Rr]. \tag{5.9}$$

Hence, the OLA estimate $\hat{x}_{OLA}$ is an inverse if and only if

$$\sum_{r=-\infty}^{\infty} w[n - Rr] = 1 \tag{5.10}$$

for all n. In fact, as long as

$$\sum_{r=-\infty}^{\infty} w[n - Rr] = C \tag{5.11}$$

for some constant $C \neq 0$ and all n, the OLA estimate produces $Cx[n]$, from which $x[n]$ is easily obtained by multiplying by C^{-1}.

Two windows for which (5.11) holds are the Bartlett (triangular) window

$$w[n] = \begin{cases} 2n/(L-1) & 0 \leq n \leq (L-1)/2 \\ 2 - 2n/(L-1) & (L-1)/2 < n \leq L-1 \\ 0 \text{else} \end{cases} \tag{5.12}$$

and the Hann window

$$w[n] = \begin{cases} 0.5 - 0.5\cos(2\pi n/(L-1)) & 0 \leq n \leq L-1 \\ 0 & \text{else.} \end{cases} \tag{5.13}$$

For both of these windows, (5.11) holds whenever $L-1 = 2^p$ for some integer p, and $R = (L-1)/2^{p'}$ for some integer $p' < p$. So, $R = (L-1)/2, (L-1)/4, (L-1)/8, \ldots, 1$ all work. We can demonstrate this graphically – see Figure 8. It can be shown analytically using detailed properties of the Fourier transforms of the Bartlett and Hann window – see Chapters 7 and 10 of the text. However, it does not hold for general windows $w[n]$. In particular, the Hamming window, which modifies the Hann window to minimize the height of the side lobe, does not have property (5.11), which can be seen graphically in Figure 9.

Practical inversion (II): Minimum norm approaches. As mentioned above, when $R < L$, the inverse mapping $\breve{S}$ is not unique. This is a consequence of the fact that there are many X which are not the STFT of any sequence x. The definition of an inverse map $\breve{S}$ does not constrain its behavior on these X. When confronted with an X which is not the short-time fourier transform of any x, a

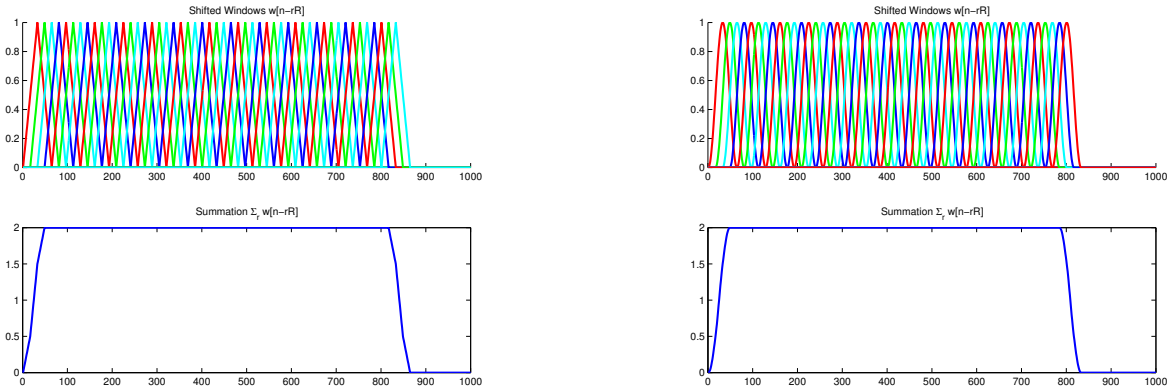

Figure 8: **Overlap and add succeeds with Bartlett and Hann windows.** Here, $L = 65$, $R = 16$. Top: shifted windows $w[n - rR]$. Bottom: summation $\sum_r w[n - rR]$. Left: Bartlett window. Right: Hann window. In both cases, $\sum_{r=-\infty}^{\infty} w[n - rR]$ is constant.

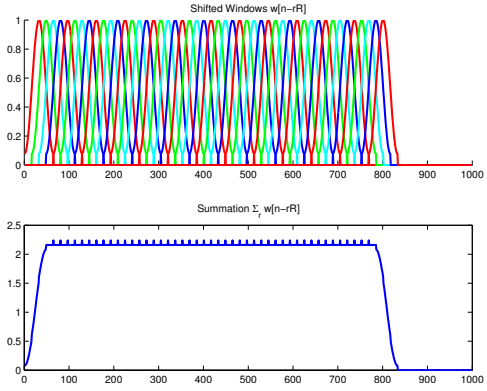

Figure 9: **Overlap and add does not provide an inverse with Hamming window.** Here $L = 65$, $R = 16$. Top: shifted windows $w[n - rR]$. Bottom: $\sum_r w[n - rR]$. Here, $\sum_{r=-\infty}^{\infty} w[n - rR]$ is not constant, and so overlap and add does not provide an exact inverse.

sensible approach to "inverting" the transform is to seek an $\hat{x}$ such that $\mathcal{S}\{\hat{x}\} \approx X$ – the STFT of $\hat{x}$ is as close as possible to X in some clearly defined sense. For example, we can use an energy or ℓ^2 measure of quality of approximation, vis

$$\min_x \|\mathcal{S}\{x\} - X\|_2^2, \tag{5.14}$$

where

$$\|E\|_2^2 = \sum_{rk} |E[r,k]|^2. \tag{5.15}$$

This also leads to relatively simple computations, which look visually similar to overlap-and-add:

Theorem 5.3 (Least Squares Reconstruction)**.** *Set* $q_r[m] = \mathrm{DFT}_N^{-1}\{X[r,\cdot]\}[m],\ m = 0,\dots,N-1$. *The problem* (5.14) *has a unique optimal solution* $\hat{x}_{LS}$*, given by*

$$\hat{x}_{LS}[n] = \frac{\sum_{r=-\infty}^{\infty} w[n-rR]q_r[n-rR]}{\sum_{r'=-\infty}^{\infty} w^2[n-r'R]}. \tag{5.16}$$

Proof. Our goal is to

$$\begin{aligned}
&\min_x \sum_{r,k} |\mathcal{S}\{x\}[r,k] - X[r,k]|^2 \\
&= \min_x \sum_r \sum_k \left|\mathrm{DFT}_N\{w[\cdot]x[rR+\cdot]\}[k] - \mathrm{DFT}_N\mathrm{DFT}_N^{-1}\{X[r,\cdot]\}\right|^2 \\
&= \min_x \sum_r N \sum_{m=0}^{N-1} \left|w[m]x[rR+m] - \mathrm{DFT}_N^{-1}\{X[r,\cdot]\}[m]\right|^2 \qquad \text{(by Parseval)} \\
&= \min_x \sum_r \sum_{m=0}^{N-1} |w[m]x[rR+m] - q_r[m]|^2
\end{aligned} \tag{5.17}$$

The above problem decouples into separate problems for each entry of $x[n]$:

$$\begin{aligned}
\hat{x}_{LS}[n] &= \arg\min_z \sum_{r=-\infty}^{\infty} |w[n-rR]z - q_r[n-rR]|^2 \\
&= \frac{\sum_{r=-\infty}^{\infty} w[n-rR]q_r[n-rR]}{\sum_{r'=-\infty}^{\infty} w^2[n-r'R]}.
\end{aligned} \tag{5.18}$$

□

The least squares approach yields an inverse map $\check{\mathcal{S}}$, whenever one exists.

Other norms. Depending on our goals and assumptions, we can choose other norms for measuring quality of approximation. For example the ℓ^∞ norm

$$\|E\|_\infty = \max_{r,k} |E[r,k]| \tag{5.19}$$

leads to the best uniform approximation

$$\min_x \|\mathcal{S}\{x\} - X\|_\infty , \tag{5.20}$$

while the ℓ^1 norm

$$\|E\|_1 = \sum_{r,k} |E[r,k]| \tag{5.21}$$

leads to "robust" approximations, which are allowed to be vastly different from X on a few entries $[r,k]$. There exist efficient[6] algorithms for finding best ℓ^∞ and ℓ^1 approximations,[7] although these algorithms are slower than least squares reconstruction.

An important issue, especially in audio processing, is how to properly handle the phase of the STFT, which may be garbled by processing, leading to aesthetically unpleasant reconstructions. One approach is to simply ignore the phase information, and try to reconstruct the time domain signal x from its STFT magnitude only. This is an example of a *phase retrieval* problem. Phase retrieval is a rich area, but is well beyond the scope of our discussion here.

[6] In the computer scientist's sense.

[7] These are examples of "convex" optimization problems, which can often be solved efficiently.