Elec4621
Advanced Digital Signal Processing
Chapter 11: Time-Frequency Analysis

Dr. D. S. Taubman

May 23, 2011

In this last chapter of your notes, we are interested in the problem of finding the "instantaneous" power spectrum of a signal. Earlier we encountered various methods for estimating the power spectrum of a random process, $X[n]$, based on a finite number of observations of the signal, $x[0]$ through $x[N-1]$. The assumption was that the random process is stationary, so that the power spectrum does not change over time.

In practice, signals generally have time varying statistics and tracking the power spectrum as it changes over time is fundamental to many practical applications. Consider, for example, an audio recording from a music concert. As each instrument sounds a note, the power spectrum will contain a peak at the corresponding frequency. Tracking the peaks over time allows us to find the point at which each instrument plays each note and the duration of those notes. In practice, of course, the sounding of a single note from a musical instrument generally involves a complex pattern of harmonics. To detect and track harmonic signatures requires some kind of "short time" spectrum, rather than a "whole signal" spectrum. This kind of tracking can be used for many things, including to distinguish between instruments in an ensemble. If multiple microphones are available, cross correlation of the separated instrument tracks can reveal the locations of the sound sources, which opens up the possibility of "auditory scene analysis" and potentially the synthesis of a new composition with a subset of instruments at different locations.

1 DFT and Filter Banks

One obvious way to keep track of changes in the power spectrum over time is to divide the signal $x[n]$ into blocks, each of length N. Within each block, we may use the periodogram as our estimate of the power spectrum within that block. Specifically, the q^{th} block's periodogram is written

$$S_q(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[Nq+n]\, e^{-j\omega n}\right|^2$$

As discussed previously, $S_q(\omega)$ is completely represented by its samples at the frequencies $\omega = \frac{2\pi}{N}k$, which may be found by taking the N-point DFT of the q^{th} block of samples, and squaring their magnitudes. Specifically,

$$\begin{aligned} S_q[k] &\triangleq S_q(\omega)|_{\omega=\frac{2\pi}{N}k} = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[Nq+n]\, e^{-j\frac{2\pi}{N}nk}\right|^2 \\ &= \frac{1}{N}|X_q[k]|^2 \end{aligned}$$

where $X_q[k]$ is the DFT of the q^{th} block of samples, $x[qN]$ through $x[qN+N-1]$. That is,

$$\begin{aligned} X_q[k] &= \sum_{n=0}^{N-1} x[Nq+n]\, e^{-j\frac{2\pi}{N}nk} \\ &= \sum_{n=-\infty}^{\infty} x[n]\, h_k[Nq-n] \\ &= \left[\sum_{n=-\infty}^{\infty} x[n]\, h_k[p-n]\right]_{p=Nq} \end{aligned} \tag{1}$$

where

$$h_k[n] = \begin{cases} e^{j\frac{2\pi}{N}nk} & 0 \le n < N \\ 0 & \text{otherwise} \end{cases}$$

The reader will recognize equation (1) as convolution of $x[n]$ by a filter with complex valued impulse response $h_k[n]$, followed by decimation of the result by the factor N, as shown below:

$$x[n] \longrightarrow \boxed{\star\, h_k[n]} \longrightarrow \boxed{\downarrow N} \longrightarrow X_q[k]$$

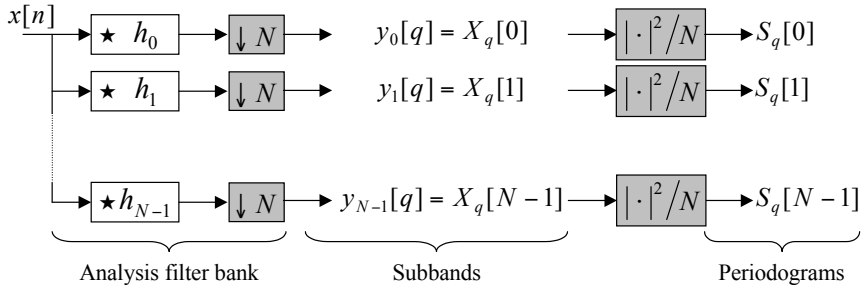

Figure 1: *Block-based power spectrum analysis interpreted as a filter bank system.*

The complete power spectrum analysis system is illustrated in Figure 1. Note that the system is essentially a filter bank, having N complex-valued filters, $h_k[n]$. Half of the filter outputs are redundant, due to conjugate symmetry of the DFT. As a result, the power spectrum analyzer could alternatively be implemented using a bank of N real-valued filters, corresponding to the real and imaginary parts of $h_k[n]$. Although we could pursue these matters further, it is convenient to stick with our complex valued filter bank for the purpose of analysis and interpretation.

The role of the decimators in Figure 1 is to ensure that only one estimated power spectrum is produced for each disjoint block of N samples. There is no reason why we cannot sample the filter outputs at a higher rate, to obtain more frequent updates of the power spectrum. Of course, this will produce redundant information, since the signal can already be reconstructed perfectly from the DFT of each of its blocks – i.e., from the outputs of the maximally decimated filter bank. In the extreme case, we could do away with the decimators altogether and get a new power spectrum for each new sample. We might write these power spectra as

$$\begin{aligned} S[p,k] &= \frac{1}{N}\left|\sum_n x[n]\, h_k[p-n]\right|^2 \\ &= \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n+p]\, e^{-j\frac{2\pi}{N}nk}\right|^2 \end{aligned}$$

the p^{th} power spectrum, $S[p,k]$, being based upon input samples $x[p]$ through $x[p+N-1]$. Of course, we would expect $S[p,k]$ to change only slowly as a function of p.

2 Short Time Fourier Transform

We have seen that block-based application of the periodogram is equivalent to implementing a bank of N filters, squaring the magnitudes of their outputs, and sub-sampling the result. Moreover, the sub-sampling is only a device for controlling the amount of overlap between successive blocks; if we sub-sample by N, the blocks are disjoint.

The properties of the individual spectral estimates may be understood by considering the properties of the filters in the filter bank. In particular, observe that each filter's impulse response is a windowed sinusoid,

$$h_k[n] = w[n] \cdot e^{j\frac{2\pi}{N}nk} \tag{2}$$

where $w[n]$ is the all-or-nothing window,

$$w[n] = \begin{cases} 1 & 0 \le n < N \\ 0 & \text{otherwise} \end{cases}$$

It follows that

$$\begin{aligned} \left|\hat{h}_k(\omega)\right| &= \left|\hat{w}\left(\omega - \frac{2\pi}{N}k\right)\right| \\ &= \left|\frac{\sin\frac{N}{2}\left(\omega - \frac{2\pi}{N}k\right)}{\sin\frac{1}{2}\left(\omega + \frac{2\pi}{N}k\right)}\right| \\ &= \left|\sum_j \operatorname{sinc}\left(\frac{N}{2\pi}\left(\omega - \frac{2\pi}{N}k + 2\pi j\right)\right)\right| \end{aligned}$$

That is, $\hat{h}_k(\omega)$ is a sum of the aliasing (wrap around) components produced by a sinc function, centred at $\omega = \frac{2\pi}{N}k$.

In order to improve the spectral selectivity of the individual channels in the filter bank, we will need to use a window function with a larger region of support and smoother decay. This leads to the so-called Short Time Fourier Transform (STFT). The STFT is a type of "modulated filter bank," in which the filters are obtained by modulating a basic prototype function, $w[n]$ (the window), with complex sinusoids, exactly as in equation (2).

2.1 Properties of the Gaussian Window Function

The most common choices of window function for the STFT are the Hanning (raised cosine) window and a Gaussian window. We will concentrate on the

properties of the Gaussian window shape, which are most revealing. We begin by considering the continuous time window function,

$$w(t) = e^{-\frac{1}{2}\frac{t^2}{\sigma^2}}$$

The Fourier transform of $w(t)$ is found from

$$\begin{aligned}
\hat{w}(\omega) &= \int e^{-\frac{1}{2}\frac{t^2}{\sigma^2}} e^{-j\omega t} dt \\
&= \int \exp\left(-\frac{1}{2}\frac{t^2 + 2j\omega\sigma^2}{\sigma^2}\right) dt \\
&= \int \exp\left(-\frac{1}{2}\frac{\left[t + j\omega\sigma^2\right]^2 + \omega^2\sigma^4}{\sigma^2}\right) dt \\
&= \exp\left(-\frac{1}{2}\frac{\omega^2}{\sigma^{-2}}\right) \int \exp\left(-\frac{1}{2}\frac{\left[t + j\omega\sigma^2\right]^2}{\sigma^2}\right) dt \\
&= \sqrt{2\pi\sigma^2} \exp\left(-\frac{1}{2}\frac{\omega^2}{\sigma^{-2}}\right)
\end{aligned}$$

The last line follows from the fact that the integral on the second last line is identical to the integral of the Gaussian function $w(t)$, with a time displacement of $j\omega\sigma^2$. The time displacement has no impact on the indefinite Gaussian integral, which is equal to $\sqrt{2\pi\sigma^2}$.

The key point to observe here is that the Fourier transform of the Gaussian window function is itself a Gaussian function, whose variance is the reciprocal of that of the time-domain function. We may summarize this as

$$\sigma_t = \frac{1}{\sigma_\omega}$$

Turning our attention back now to the discrete case, we set

$$w[n] = w(t)|_{t=n} = e^{-\frac{1}{2}\frac{n^2}{\sigma^2}}$$

To obtain the DTFT of $w[n]$, we should strictly take into account the aliasing components from the true Fourier transform of $w(t)$, but so long as

$$\sigma_w \lesssim 1$$

there will be no substantial aliasing contributions, due to the doubly exponential decay of $\hat{w}(\omega)$. For this reason, we will generally take

$$\hat{w}(\omega) = \sqrt{2\pi\sigma_t^2} \exp\left(-\frac{1}{2}\frac{\omega^2}{\sigma_t^{-2}}\right) = \frac{2\pi}{\sqrt{2\pi\sigma_\omega^2}} \exp\left(-\frac{1}{2}\frac{\omega^2}{\sigma_\omega^2}\right)$$

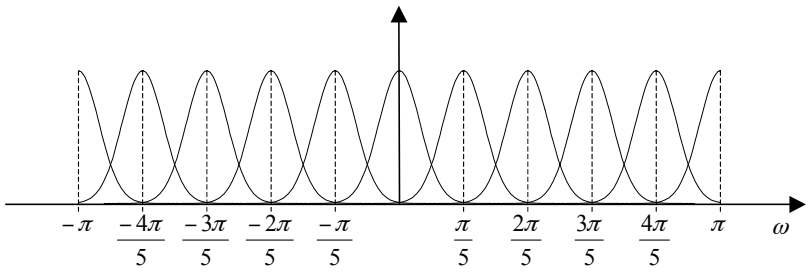

Figure 2: *Frequency responses of the modulated Gaussian filters associated with the STFT.*

to be both the Fourier transform of $w(t)$ and the DTFT of $w[n]$.

Similarly, although the Gaussian window function does not strictly decay to zero at any point, its rapid decay ensures that we need only keep samples in the range

$$|n| \lesssim 4\sigma_t = \frac{4}{\sigma_\omega}$$

Thus, for all practical purposes, the Gaussian window function may be understood as having finite support in both time and frequency.

Using this window function, the STFT filter bank involves filters with the modulated impulse responses,

$$h_k[n] = e^{-\frac{1}{2}\frac{n^2}{\sigma^2}} e^{j\frac{2\pi}{N}kn}$$

whose frequency responses are

$$\hat{h}_k(\omega) = \hat{w}\left(\omega - \frac{2\pi}{N}k\right) = \frac{2\pi}{\sqrt{2\pi\sigma_\omega^2}} \exp\left(-\frac{1}{2}\frac{\left(\omega - \frac{2\pi}{N}k\right)^2}{\sigma_\omega^2}\right)$$

These frequency responses are sketched stylistically in Figure **??**.

2.2 Time and Frequency Resolution of the STFT

Regardless of whether or not we choose to decimate the outputs of the STFT filter bank, the ability of any individual frequency channel to respond to rapid changes in the underlying signal statistics is limited by the extent of the window function, $w(t)$. We may formalize the concept of the filter bank's time resolution by considering an input signal consisting of two impulses, separated by time τ. That is,

$$x(t) = \delta(t) + \delta(t - \tau)$$

Considering the DC frequency band (the others respond in a similar way, with modulated copies of the same signal), we have

$$\begin{aligned} y_0(t) &= (h_0 \star x)(t) \\ &= \exp\left(-\frac{1}{2}\frac{t^2}{\sigma_t^2}\right) + \exp\left(-\frac{1}{2}\frac{(t-\tau)^2}{\sigma_t^2}\right) \end{aligned}$$

Clearly, the response is symmetric about $t = \frac{\tau}{2}$. If τ is large, there will be two clear peaks in the response; one at $t = 0$ and one at $t = \tau$. If τ is sufficiently small, there will be only one peak at $t = \frac{\tau}{2}$.

A reasonable definition for the time resolution Δt of a particular window function is the value of τ at which the two peaks merge into one. At this point, the second derivative of $y_0(t)$ must be exactly equal to 0 at $t = \frac{\tau}{2}$, since a positive value implies two peaks, while a negative value implies only one peak in the response. We find that

$$\frac{\partial^2 y_0(t)}{\partial t^2} = \left(\frac{t^2}{\sigma_t^4} - \frac{1}{\sigma_t^2}\right)\exp\left(-\frac{1}{2}\frac{t^2}{\sigma_t^2}\right) + \left(\frac{(t-\tau)^2}{\sigma_t^4} - \frac{1}{\sigma_t^2}\right)\exp\left(-\frac{1}{2}\frac{(t-\tau)^2}{\sigma_t^2}\right)$$

so that our condition becomes

$$0 = \left.\frac{\partial^2 y_0(t)}{\partial t^2}\right|_{t=\frac{\tau}{2}} = 2\left(\frac{(\tau/2)^2}{\sigma_t^4} - \frac{1}{\sigma_t^2}\right)\exp\left(-\frac{1}{2}\frac{(\tau/2)^2}{\sigma_t^2}\right)$$

and hence we must have

$$\Delta t = 2\sigma_t$$

Applying exactly the same argument in the frequency domain, we find that regardless of the number of frequency channels, N, the ability of the STFT to distinguish between sinusoids with two different frequencies is limited by a frequency resolution, $\Delta\omega$, which is equal to

$$\Delta\omega = 2\sigma_\omega = \frac{2}{\sigma_t}$$

Putting these results together, we find that the time-frequency resolution product satisfies

$$\Delta t \cdot \Delta\omega = 4$$

or, noting that $\omega = 2\pi f$,

$$\Delta t \cdot \Delta f = \frac{2}{\pi}$$

This fundamental result is known as the "uncertainty principle." It tells us that there is no way to improve the time resolution without sacrificing frequency resolution, or vice-versa. We can reliably observe a signal with high frequency resolution only by using very long window functions, which lead to very poor time resolution. Conversely, to observe rapid temporal fluctuations in the signal, we require very short window functions, which results in poor frequency resolution. It turns out that out of all possible window functions, the Gaussian window is the one which minimizes the time-frequency uncertainty product, $\Delta t \cdot \Delta f$, which provides theoretical justification for the use of Gaussian windows. This minimum time-frequency uncertainty product is analogous to Heisenberg's famous uncertainty principle in Physics.

2.3 Fast FFT Implementation

Our filter bank produces k outputs at each sampling point, p, given by

$$\begin{aligned} y[p,k] &= \sum_n h_k[n]\, x[p-n] \\ &= \sum_n h_k[-n]\, x[p+n] \\ &= \sum_{n=-L}^{L} x[n+p]\, w[n]\, e^{-j\frac{2\pi}{N}kn} \end{aligned}$$

where $w[n] = w[-n]$ is the symmetric window function, having support $|n| \leq L$. Evidently, $y[p,k]$ is the N-point DFT of the signal,

$$u_p[n] = w[n]\, x[n+p]$$

so long as $N \geq 2L+1$, so that the entire support of $u_p[n]$ is captured by the DFT.

This provides us with an alternate view of the STFT, in which the window function, $w[n]$, is applied to input samples in the range $p-L$ through $p+L$. The DFT of the windowed input samples is taken and the squared magnitudes of the DFT coefficients are usually taken to form a local spectrum of the signal. To form the spectrum at the next instant, $p+1$, the window is moved one sample to the right and the process is repeated. This procedure has obvious computational advantages over directly filtering the input sequence, since the FFT algorithm may be used to implement the DFT operation.

We have seen that the STFT is a generalization of the block-based DFT from two perspectives. From one perspective, we are replacing the

all-or-nothing window associated with the DFT filter bank, with a longer, smoother window (e.g., a Gaussian). From another perspective, we are applying a moving window to the data prior to taking the DFT.

2.4 Design Parameters

From the above discussion, there appear to be four design parameters of interest for the STFT. The number of frequency channels, N, and the temporal sub-sampling factor, say M, are two parameters. In the extreme case, we may keep the outputs from each frequency channel at each time instant, p, in which case $M = 1$, while a maximally decimated filter bank would have $M = N$. We have seen, however, that there are two much more fundamental parameters, Δt and $\Delta\omega$, identifying the time and frequency resolution of the filter bank, respectively. These are related through the uncertainty principle, and their product can be minimized by selecting a Gaussian window function.

Starting with the assumption of a Gaussian window function, one first selects σ_t in such as way as to ensure that the time resolution,

$$\Delta t = 2\sigma_t$$

and the frequency resolution

$$\Delta\omega = \frac{2}{\sigma_t}$$

are sufficient for the application at hand, bearing in mind that the discrete implementation requires

$$\sigma_t \gtrsim 1$$

to minimize aliasing effects.

The window size parameter may then be set to

$$L = 4\sigma_t$$

so that the window, and hence all of the filters, have $2L + 1 = 8\sigma_t + 1$ taps (for most applications, one could also away with $L = 3\sigma_t$, but not much less). An FFT-based implementation is most natural, for which we would set N to the nearest power of 2, such that

$$N \geq 2L + 1 = 8\sigma_t + 1$$

This gives us a frequency spacing of

$$\frac{2\pi}{N} \leq \frac{2\pi}{8\sigma_t + 1} \approx \frac{2\pi}{8}\sigma_\omega = \frac{\pi}{8}\Delta\omega$$

so that two sinusoids with the minimum resolvable frequency separation, are separated by more than 2 frequency channels. This gives our frequency sampling structure a modest level of redundancy.

For the same level of redundancy in the temporal sampling structure, we may select

$$M \gtrsim \frac{\pi}{8}\Delta t = \frac{2\pi}{8}\sigma_t \approx \sigma_t$$

suggesting that we take samples roughly every σ_t input samples. Depending upon the application, we may opt for a finer or coarser temporal sampling, since the temporal sampling rate directly affects the computational complexity of the STFT.

3 Wavelet Transforms

The principle weakness of the STFT is that every frequency channel has exactly the same frequency uncertainty, $\Delta\omega$, and exactly the same temporal uncertainty, Δt. Motivated by the human auditory system, it seems more natural to build time-frequency transforms whose frequency resolution varies logarithmically with the frequency itself. That is, we would like the frequency channel whose centre frequency is ω_0 to have a frequency uncertainty $\Delta\omega$, which is roughly proportional to ω_0. This means that the low frequency filters in our filter bank should have narrow bandwidth and hence very poor temporal uncertainty (Δt will be very large in the low frequency bands). On the other hand, the high frequency bands will have small Δt and large $\Delta\omega$. In this way, the high frequency bands will enable us to accurately locate sharp transients, while the low frequency bands will enable us to recover the most interesting elements of the fine spectral structure.

Noting that the STFT is essentially a uniform filter bank, the most natural strategy for building time-frequency transforms with a logarithmic pass band structure is to use the tree-structured filter banks introduced in connection with subband transforms in the previous chapter of your notes. To this end, we reproduce the DWT (Discrete Wavelet Transform) structure in Figure 3, together with its passband configuration.

As with the STFT, we will normally reduce or even eliminate the decimation, so as to facilitate access to the evolving spectrum. If we completely

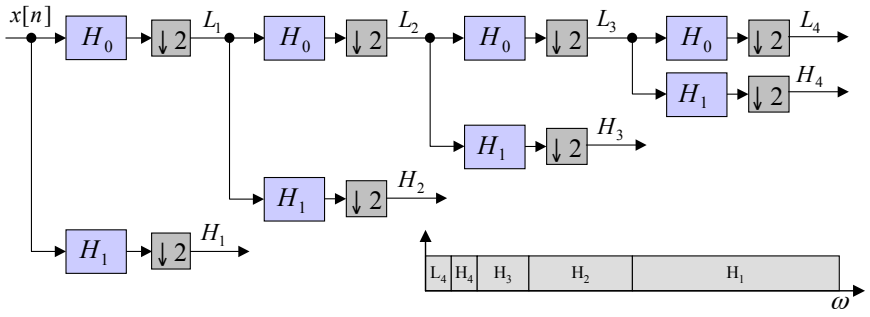

Figure 3: *DWT as an iterated filter bank, with its logarithmic passband structure.*

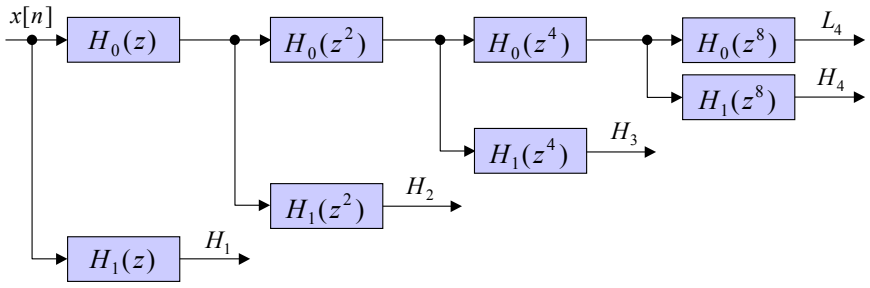

Figure 4: *DWT without any decimation.*

eliminate the decimators from the DWT structure in Figure 3, we obtain the analysis filter bank depicted in Figure 4. It has exactly the same passband structure as that shown in Figure 3. To obtain this representation, we make use of the identity

$$\boxed{\downarrow N} \longrightarrow \boxed{H\left(z\right)} \quad \equiv \quad \boxed{H\left(z^N\right)} \longrightarrow \boxed{\downarrow N}$$

A filter with Z-transform $H\left(z^N\right)$ has impulse response equal to $h\left[n\right]$ at the locations nN and 0 in between. It follows that the impulse response corresponding to $H\left(z^N\right)$ is N times longer than the impulse response corresponding to $H\left(z\right)$. We see, therefore, that low frequency bands have small bandwidths and filters with very long impulse responses, while the high frequency bands have larger bandwidths and filters with much shorter impulse responses.

Filter banks of this form are known as wavelet transforms so long as the basic filters, $H_0(z)$ and $H_1(z)$ exhibit certain regularity properties, which are required to ensure that the products,

$$H_0(z) H_0\left(z^2\right) \cdots H_0\left(z^{2^{K-1}}\right) H_0\left(z^{2^K}\right)$$

and

$$H_0(z) H_0\left(z^2\right) \cdots H_0\left(z^{2^{K-1}}\right) H_1\left(z^{2^K}\right)$$

represent smooth impulse responses, with smooth compact frequency responses, for arbitrarily large K.