Introduction to Time-Frequency Analysis – Short Term Fourier Transform

Overview

Recall from DSP:

- Discrete time Fourier transform (DTFT)
  - Taking the expression of the Fourier transform $X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$, the DTFT can be derived by numerical integration

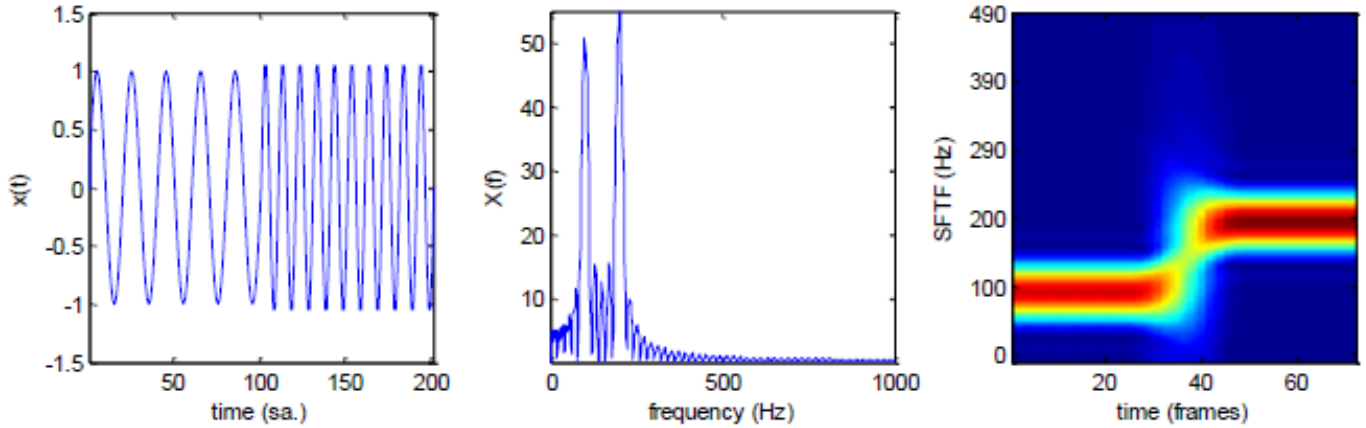
$$X(e^{j\hat{\omega}}) = \sum_{-\infty}^{\infty} x[n]e^{-j\hat{\omega}n}$$

    - where $x[n] = x(nT_S)$ and $\hat{\omega} = 2\pi F/F_S$
- Discrete Fourier transform (DFT)
  - The DFT is obtained by "sampling" the DTFT at N discrete frequencies $\omega_k = 2\pi F_s/N$, which yields the transform

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

Why is another Fourier transform needed?

- The spectral content of speech changes over time (non stationary)
  - As an example, formants change as a function of the spoken phonemes
  - Applying the DFT over a long window does not reveal transitions in spectral content
- To avoid this issue, we apply the DFT over short periods of time
  - For short enough windows, speech can be considered to be stationary
  - Remember, though, that there is a time-frequency tradeoff here



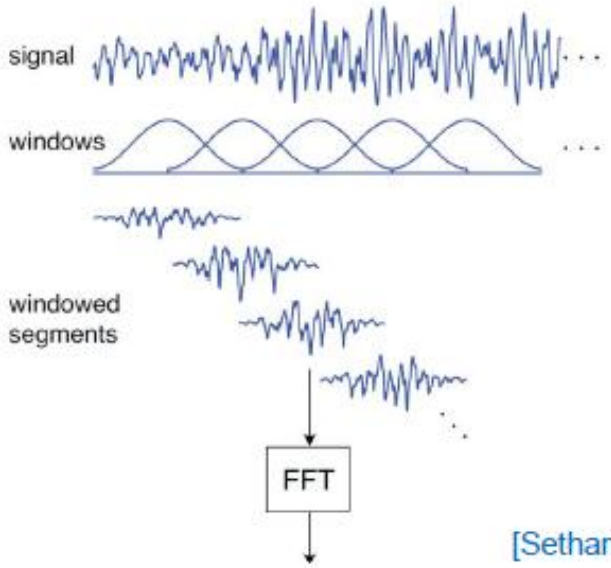
STFT

- Conventional STFT-based spectral analysis techniques build a time-frequency representation of a signal by taking the Fourier Transform (FT) of a windowed section of the signal.

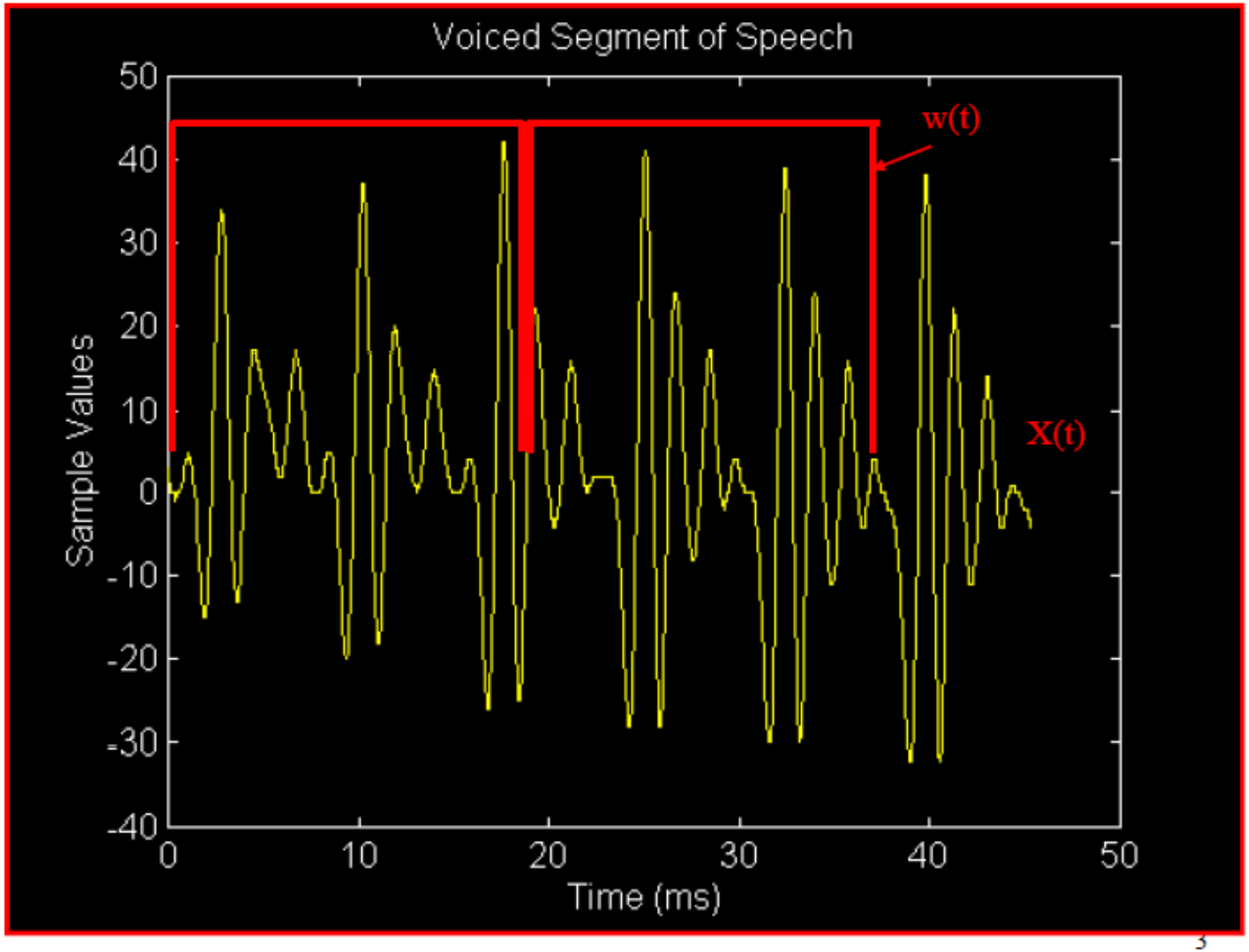
- The window is then moved along the signal in time producing a succession of estimates of the spectral components of the signal. This works well for signals composed of stationary components (e.g. sine waves) and for slowly varying signals.

The short-time Fourier transform

- Define analysis window (e.g., 30ms narrowband, 5 ms wideband)
- Define the amount of overlap between windows (e.g., 30%)
- Define a windowing function (e.g., Hann, Gaussian)
- Generate windowed segments (multiply signal by windowing function)
- Apply the FFT to each windowed segment



[Sethares, 2007]



Window Length

- To achieve effective temporal localisation, a short analysis window is required
- However, the frequency resolution is inversely proportional to the window length. Hence improved time resolution can be achieved only at the expense of poorer frequency resolution
- Thus, choosing a window function short enough to localise the high frequency transients will reduce the ability to distinguish between two adjacent frequency components.
- Conversely, choosing a longer window function will give better frequency resolution, but poorer time resolution

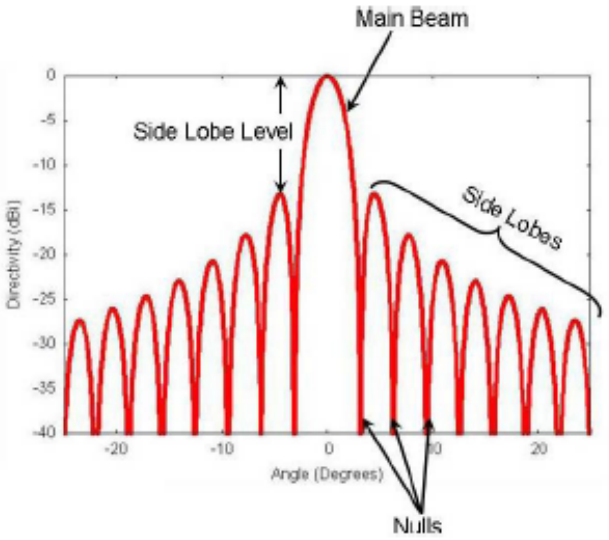
STFT: Fourier analysis view

Windowing function

- To "localize" the speech signal in time, we define a windowing function $w[n,\tau]$, which is generally tapered at its ends to avoid unnatural discontinuities in the speech segment
- Any window affects the spectral estimate computed on it
  - The window is selected to trade off the width of its main lobe and attenuation of its side lobes
- The most common are the Hann and Hamming windows (raised cosines)

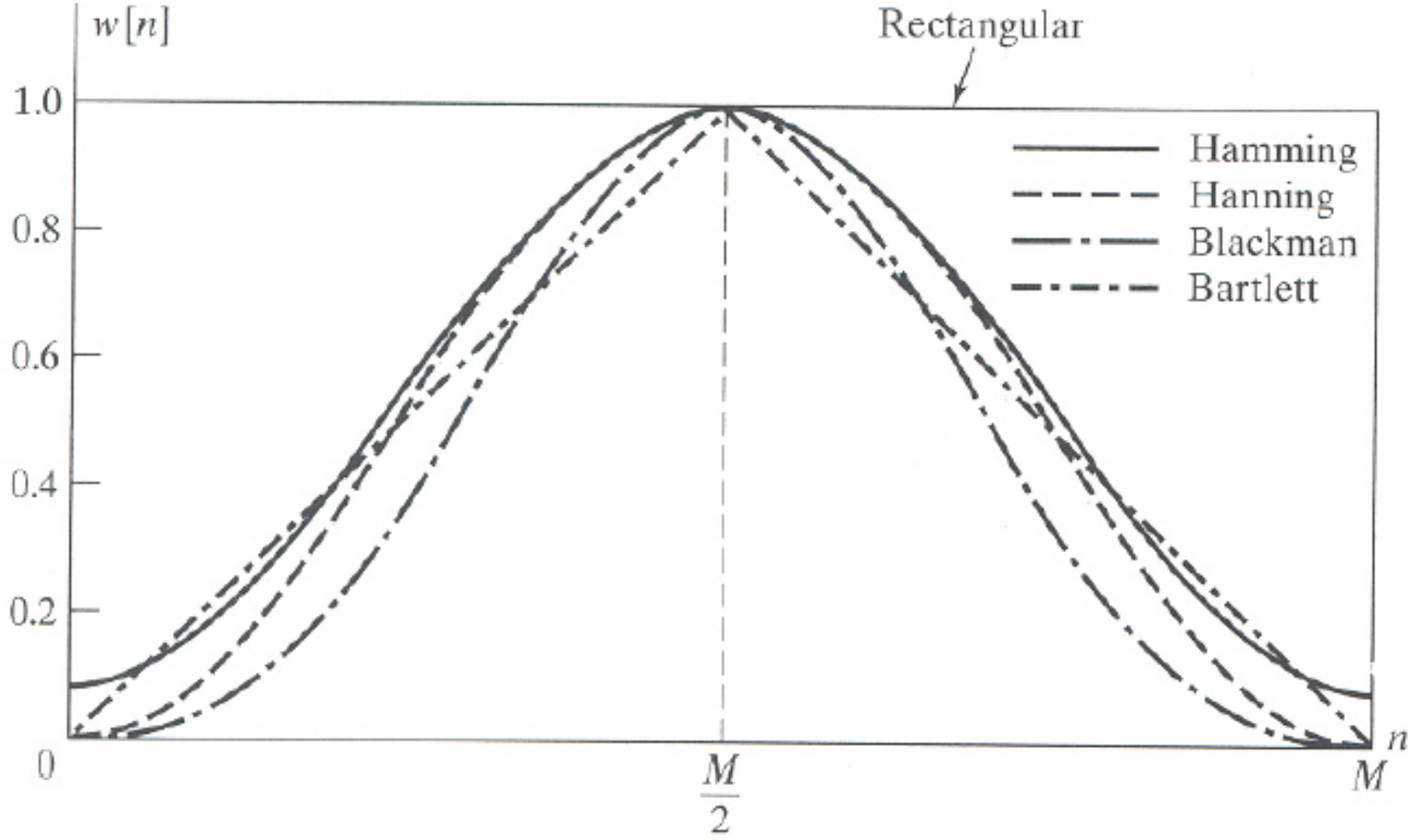
$$w[n,\tau] = 0.54 - 0.4\cos\left[\frac{2\pi(n-\tau)}{N_w - 1}\right]$$

$$w[n,\tau] = 0.5\left(1 - \cos\left(\frac{2\pi(n-\tau)}{N-1}\right)\right)$$



http://en.wikipedia.org/wiki/Window_function

Standard windows – figure



Plotted for convenience. In fact, the window is defined only at integer values of n.

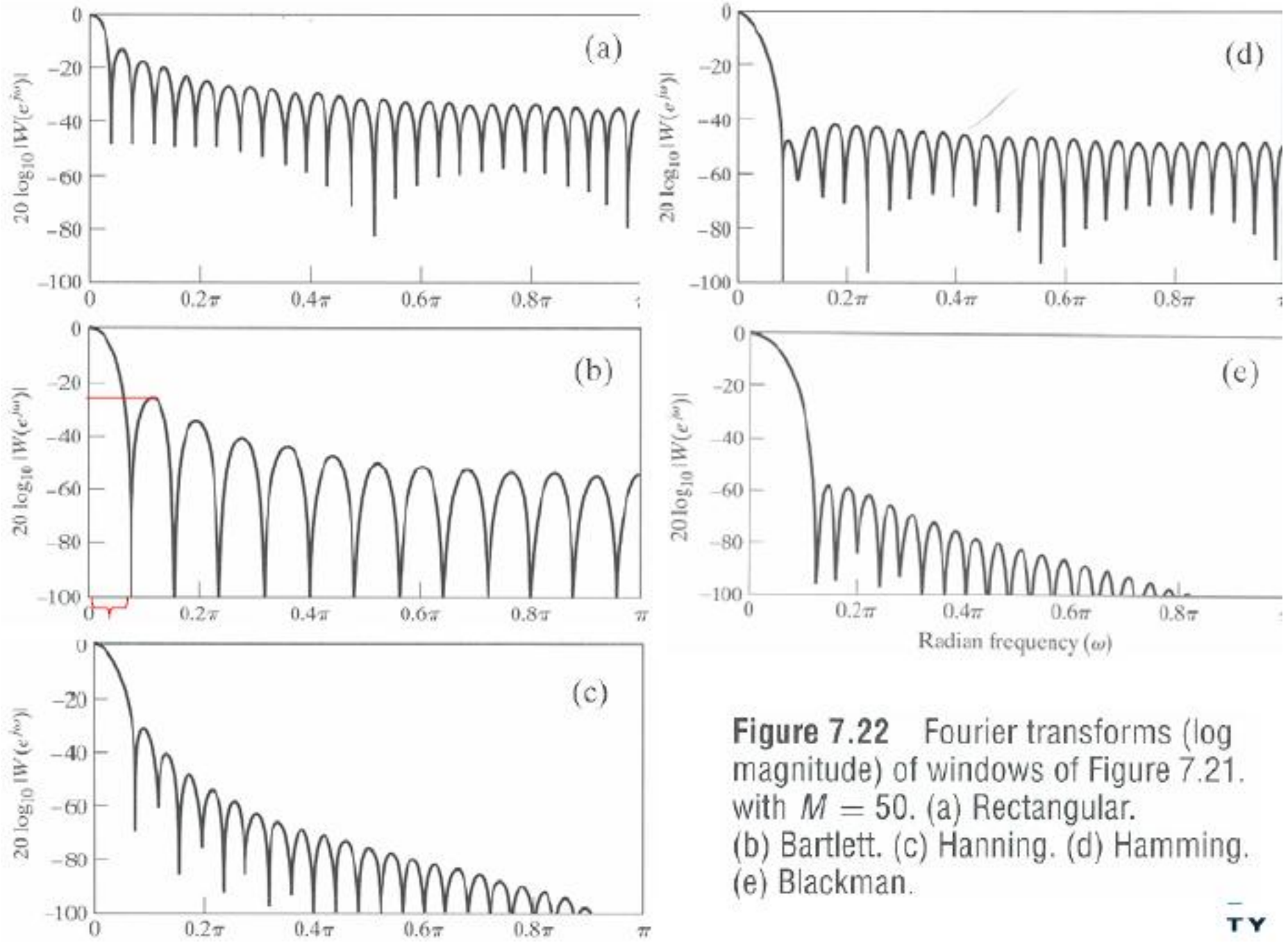


Figure 7.22 Fourier transforms (log magnitude) of windows of Figure 7.21. with $M = 50$. (a) Rectangular. (b) Bartlett. (c) Hanning. (d) Hamming. (e) Blackman.

Standard windows – comparison

Magnitude of side lobes vs width of main lobe

TABLE 7.1 COMPARISON OF COMMONLY USED WINDOWS

Type of Window	Peak Side-Lobe Amplitude (Relative)	Approximate Width of Main Lobe	Peak Approximation Error, $20\log_{10}\delta$ (dB)	Equivalent Kaiser Window, β	Transition Width of Equivalent Kaiser Window
Rectangular	−13	$4\pi/(M+1)$	−21	0	$1.81\pi/M$
Bartlett	−25	$8\pi/M$	−25	1.33	$2.37\pi/M$
Hanning	−31	$8\pi/M$	−44	3.86	$5.01\pi/M$
Hamming	−41	$8\pi/M$	−53	4.86	$6.27\pi/M$
Blackman	−57	$12\pi/M$	−74	7.04	$9.19\pi/M$

Independent of M!

Discrete-time Short-time Fourier transform

- The Fourier transform of the windowed speech waveform is defined as

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega n}$$

  - where the sequence $f_n[m] = x[m]w[n-m]$ is a short-time section of the speech signal $x[m]$ at time n

Discrete STFT

- By analogy with the DTFT/DFT, the discrete STFT is defined as

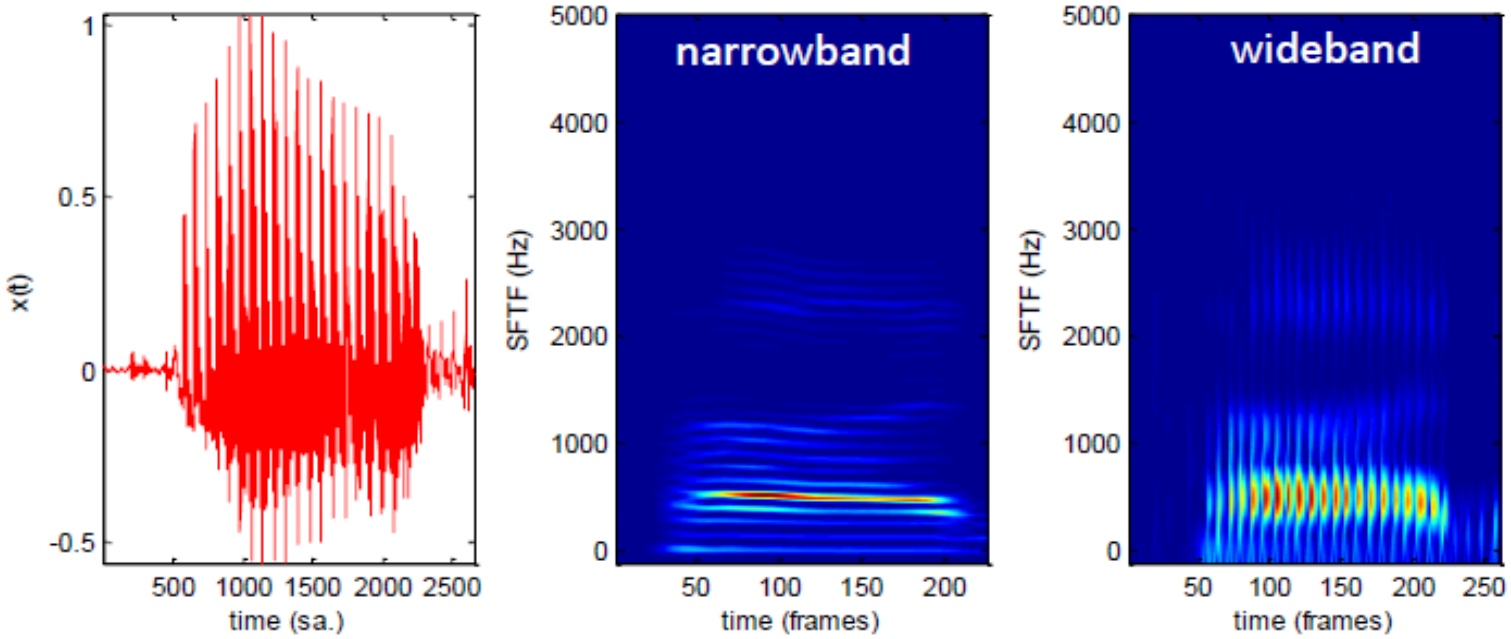
$$X(n,k) = X(n,\omega)\Big|_{\omega=\frac{2\pi}{N}k}$$

- The spectrogram we saw in previous lectures is a graphical display of the magnitude of the discrete STFT, generally in log scale

$$S(n,k) = \log|X(n,k)|^2$$

  - This can be thought of as a 2D plot of the relative energy content in frequency at different time locations

- For a long window $w[n]$, the result is the narrowband spectrogram, which exhibits the harmonic structure in the form of horizontal striations
- For a short window $w[n]$, the result is the wideband spectrogram, which exhibits periodic temporal structure in the form of vertical striations



STFT: filtering view

The STFT can also be interpreted as a filtering operation

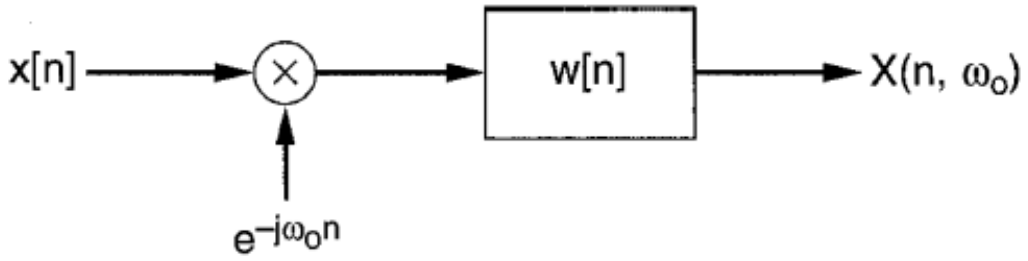
- In this case, the analysis window $w[n]$ plays the role of the filter impulse response
- To illustrate this view, we fix the value of ω at ω_0, and rewrite

$$X(n,\omega_0) = \sum_{m=-\infty}^{\infty} \left(x[m]e^{-j\omega_0 m}\right) w[n-m]$$

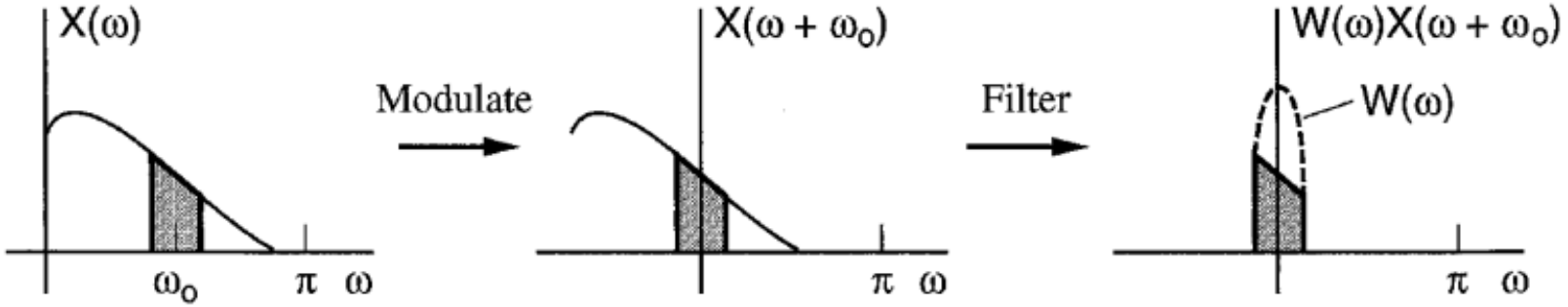
  - which can be interpreted as the convolution of the signal $\left(x[n]e^{-j\omega_0 n}\right)$ with the sequence $w[n]$:

$$X(n,\omega_0) = \left(x[n]e^{-j\omega_0 n}\right) * w[n]$$

  - and the product $x[n]e^{-j\omega_0 n}$ can be interpreted as the modulation of $x[n]$ up to frequency ω_0 (i.e., per the frequency shift property of the FT)



(a)



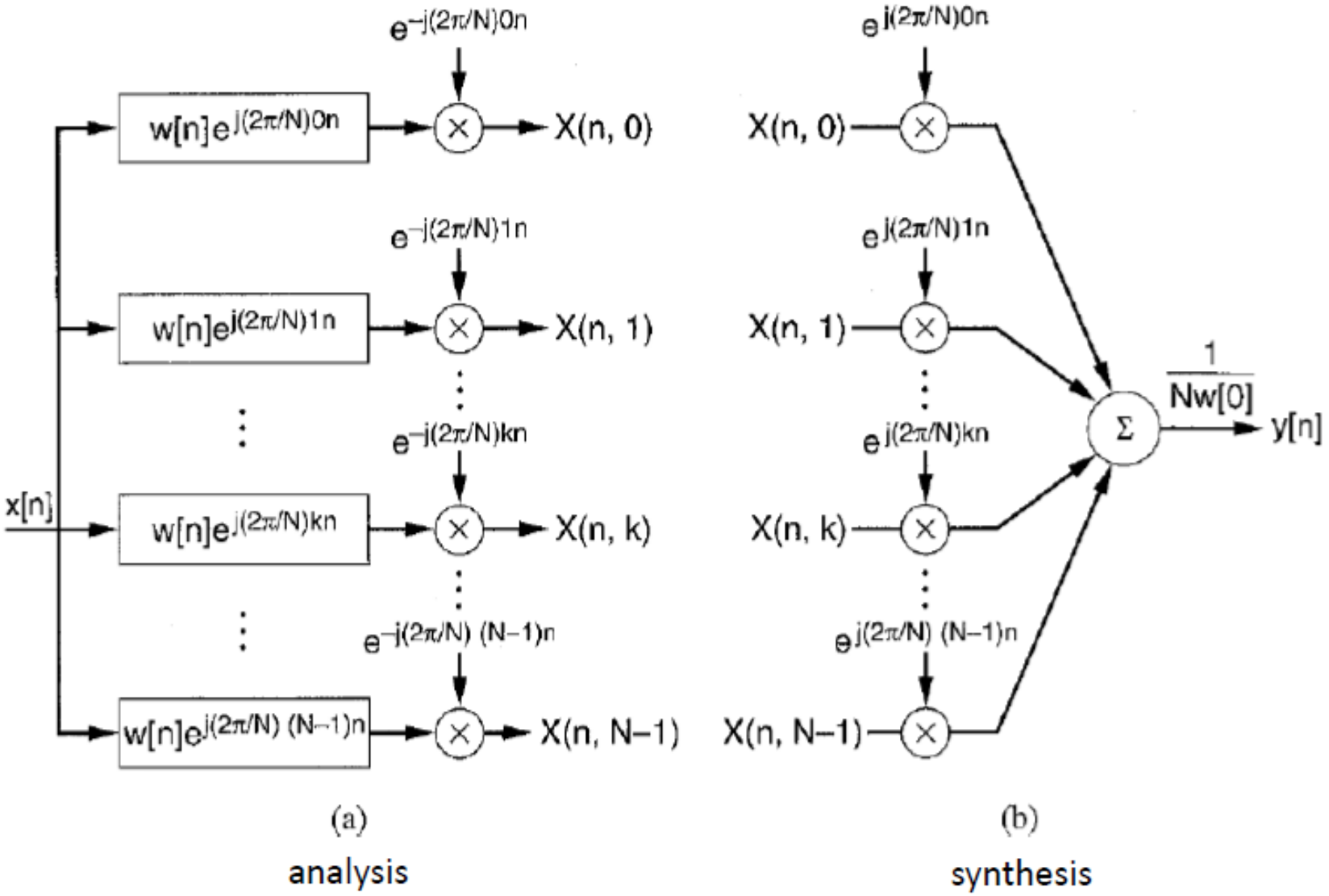
(b)

[Quatieri, 2002]

- This later rearrangement allows us to interpret the discrete STFT as the output of a filter bank

$$X(n,k) = e^{-j\frac{2\pi}{N}kn}\left(x[n] * w[n]e^{-j\frac{2\pi}{N}kn}\right)$$

  - Note that each filter is acting as a bandpass filter centered around its selected frequency
- Thus, the discrete STFT can be viewed as a collection of sequences, each corresponding to the frequency components of $x[n]$ falling within a particular frequency band
  - This filtering view is shown in the next slide, both from the analysis side and from the synthesis (reconstruction) side



[Quatieri, 2002]

Time-Frequency Resolution of the STFT

- Because both Δt and Δf are controlled by the same window length, it is not possible to decrease one without increasing the other. In fact, the time-bandwidth product Dt.Df is lower bounded (Rioul and Vetterli, 1991).
- Time Bandwidth Product = $\Delta t \Delta f \geq (1/4\pi)$
- Once a window has been chosen for the STFT, then both the time resolution and the frequency resolution are fixed. Thus, the window can be chosen to give good frequency resolution or good time resolution, but not both.

Calculating the short-time spectrum

- The short-time spectrum can be computed in various ways. However, they all involve:
  - Low-pass filtering
  - Analogue-to-Digital (A/D) conversion - convert analogue signal into a digital signal
  - Windowing - select a short section of speech centred at time *t,* and smooth its edges
  - Frequency Analysis - estimate distribution of power (amplitude) with respect to frequency over a time interval centred at *t*

Windowing

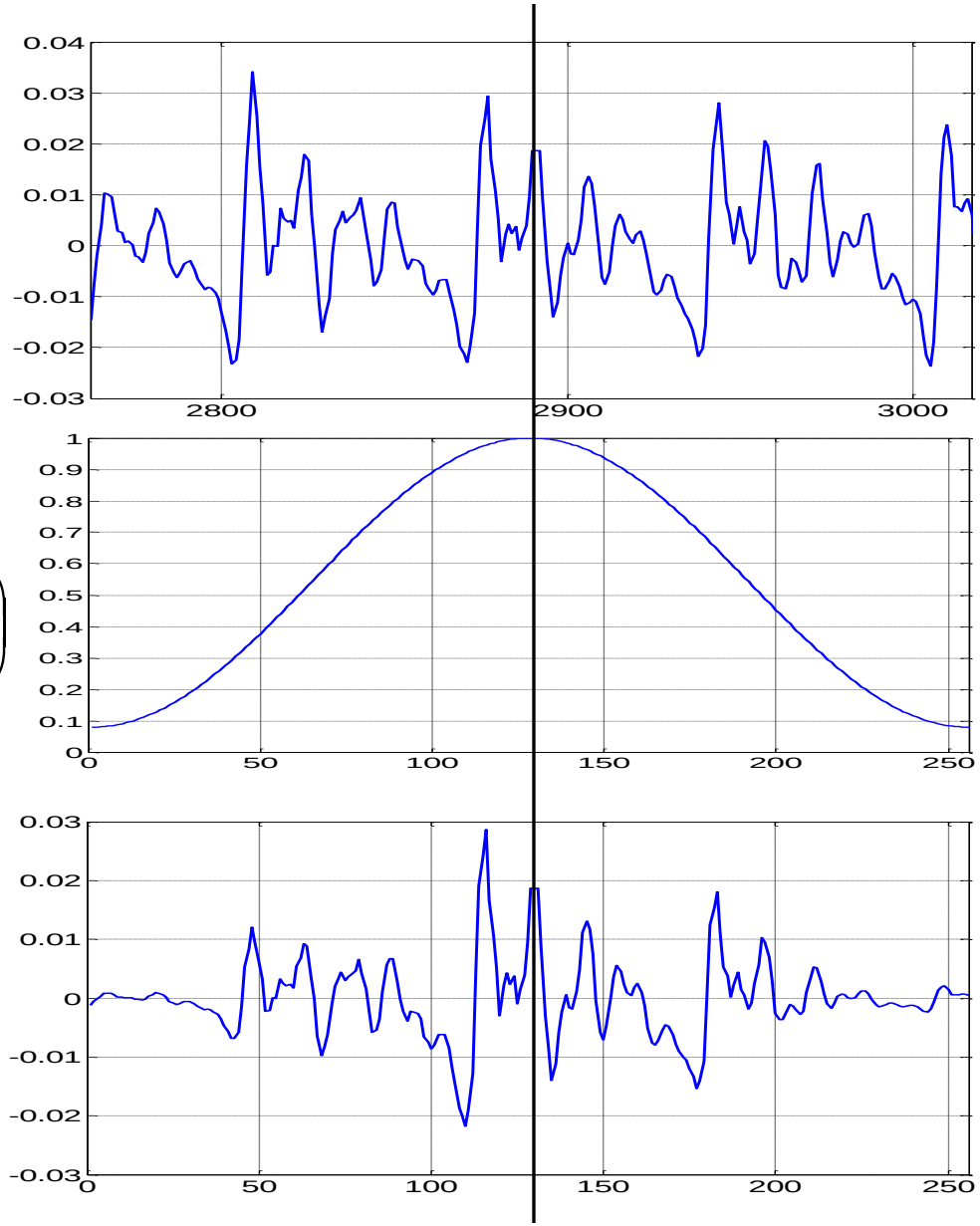
Original section of signal $s(n)$

Hamming window

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)$$

Windowed signal

$s'(n) = w(n)s(n)$



Frequency Analysis

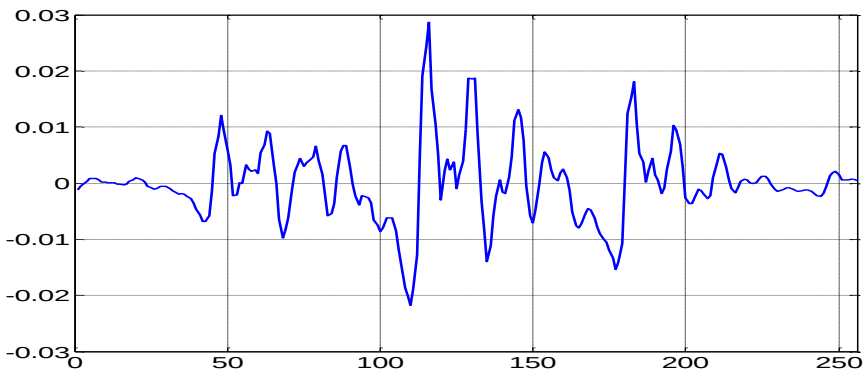
- **Discrete Fourier Transform** (DFT) applied to windowed digital waveform *{s(n):n=1,…,N}*.
- Assuming *N* sample window, this results in an *N/2* point **complex** spectrum *{S(f): f=1,..,N/2}*.
- Take **modulus** - *N/2* point **magnitude spectrum** *{P(f)=|S(f)|: f=1,…,N/2}*. (phase ignored)
- Take **logarithm** to compress dynamic range, - **log-magnitude spectrum** *{LP(f)=log|S(f)|:f=1,…,N/2}*.
- The log-magnitude spectrum, computed over a short window centred at time *t*, is referred to as the **short-time (Fourier) spectrum at time *t***.

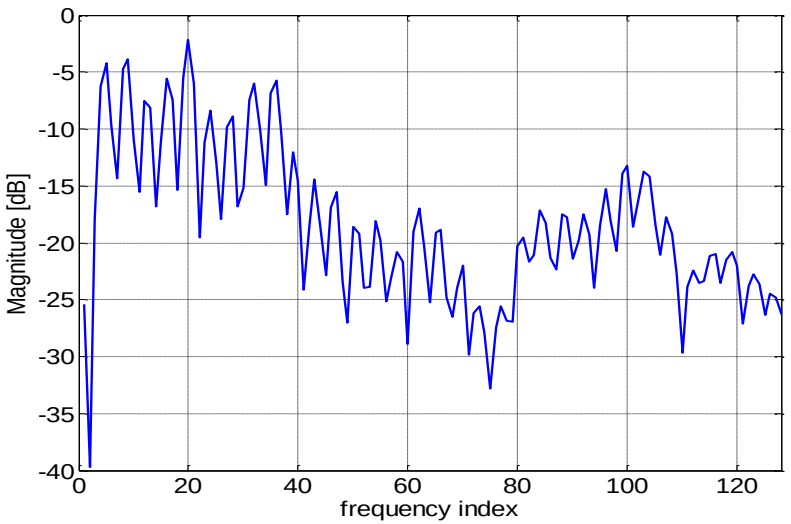
Short-time Fourier Spectrum – Example

- Fs - sampling frequency;
  - number of samples in the frame-window

N

- The frequency index ‘f’ corresponds to the frequency $f \cdot \frac{Fs}{N}$





Why short-time spectrum?

- The **short-time spectrum at time *t***
  - From the perspective of the human speech production, it tells us about the shape of the vocal tract at time *t*
  - From the perspective of human speech perception, we know that a similar analysis is performed in the cochlea in the initial stages of human speech perception

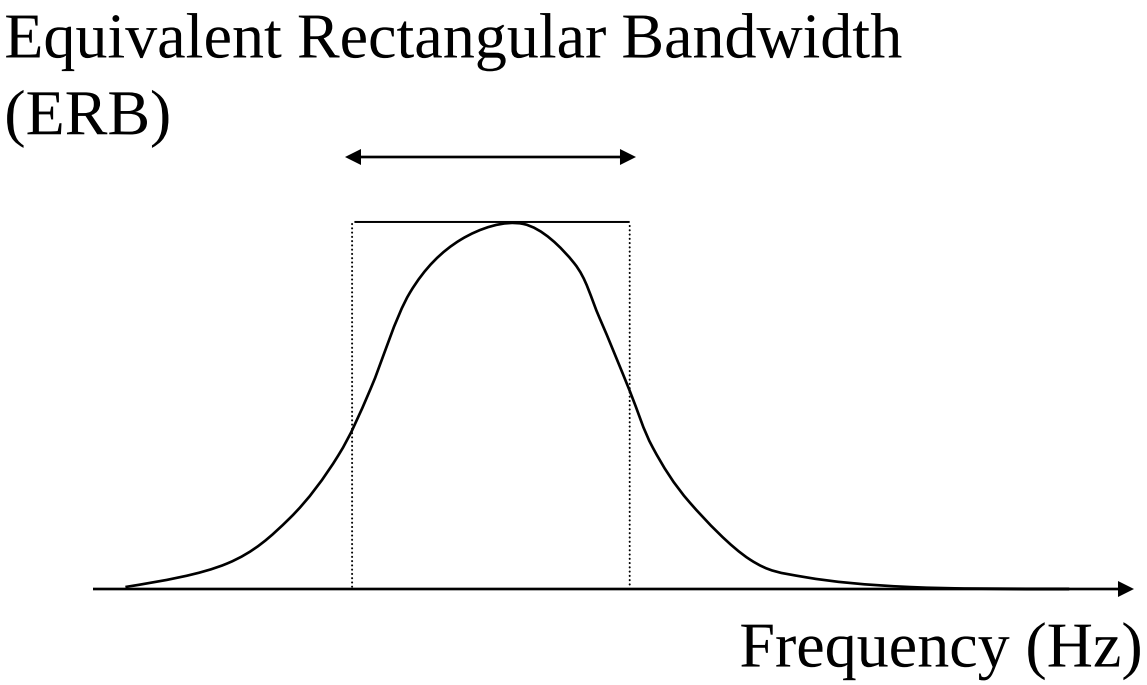
Frequency Analysis

- The use of DFT is **not** the only way to compute a short-time power spectrum
- Other approaches include:
  - **Filter-bank** analysis (based on a set of **band-pass** filters)

  and
  - **Linear Predictive Coding** (LPC-Spectrum)

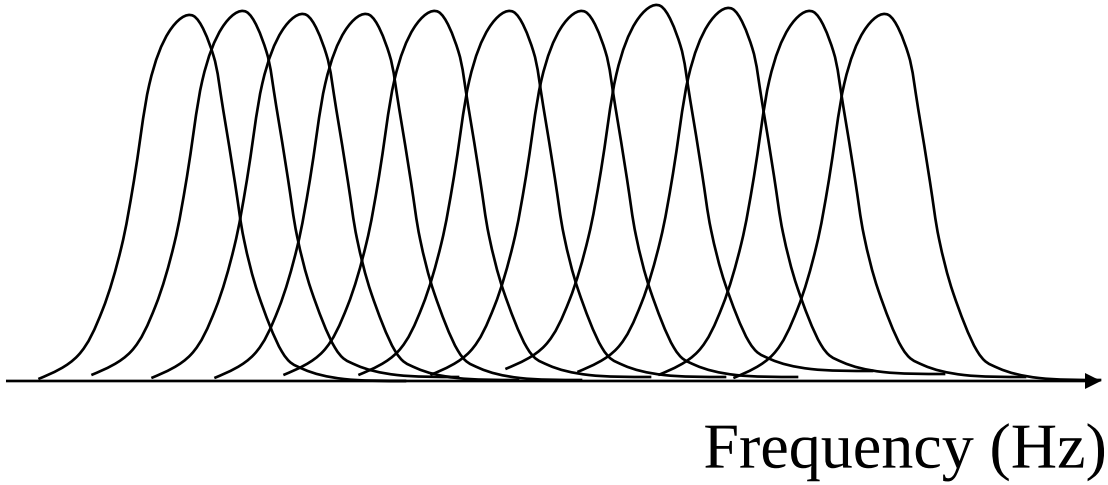
Band-pass filter

- Filters encountered in speech processing typically result from, or simulate, physiological processes



Filter-bank

- Spectrum can be estimated as a set of outputs from a bank of band-pass filters
- $y_1,\ldots,y_K$ where y_k is the output of the k^{th} filter



Time - Frequency Resolution

- Back to the DFT…
- If the window is **long** then
  - number of points *N* in frequency analysis is large

    $\Rightarrow$ the number of points in the spectrum is large,

    $\Rightarrow$ **fine** frequency resolution, **poor** temporal resolution
  - **long** window $\Rightarrow$ **narrow-band** frequency analysis - **narrow-band spectra**.

Time - Frequency Resolution

- If the window is **short** then
  - **poor frequency resolution**, but **fine temporal resolution**

    short window $\Rightarrow$ **broad-band** frequency analysis
    - **broad-band spectra**.
- In summary:

  short window $\Rightarrow$ **broad-band** frequency analysis

  long window $\Rightarrow$ **narrow-band** frequency analysis

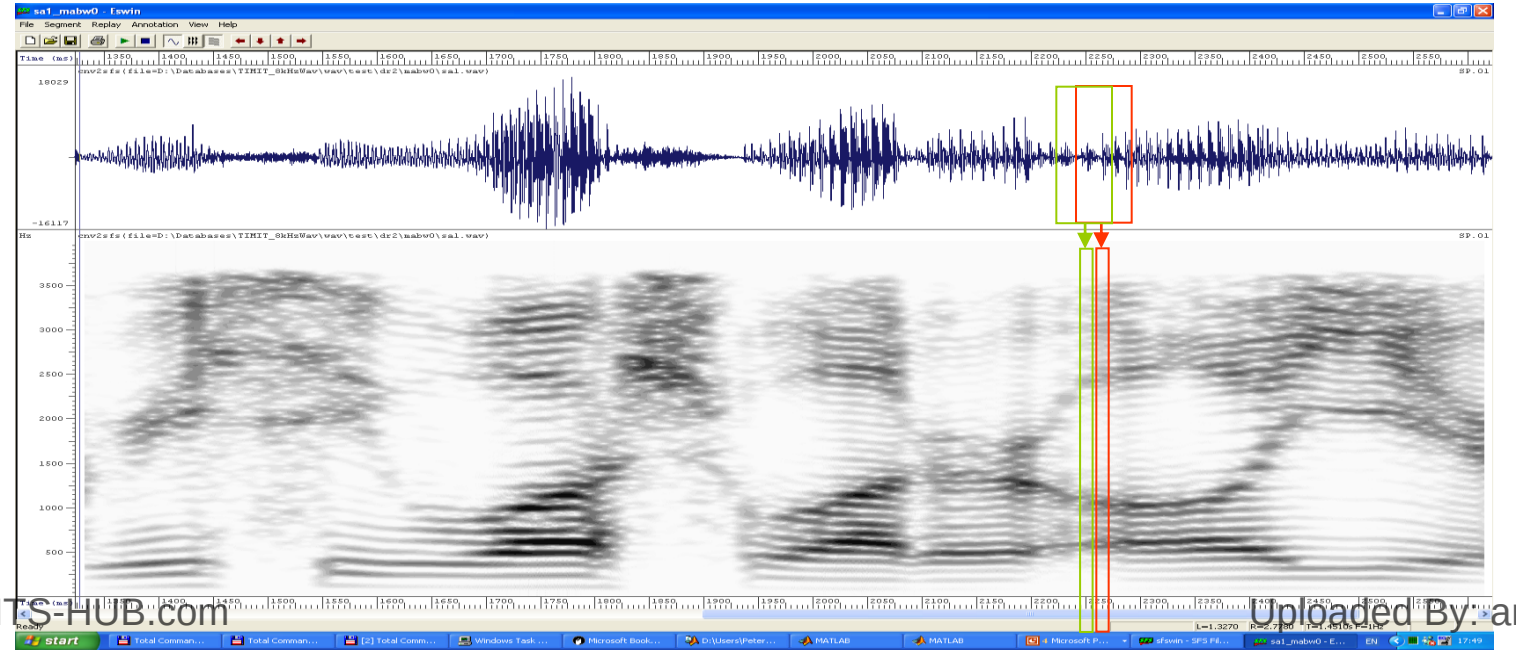
Bandwidth for implicit DFT filters

- The value of the spectrum at a particular frequency *f* can be thought of as the output of a band-pass filter, with bandwidth dependent on window type and window size (in seconds)

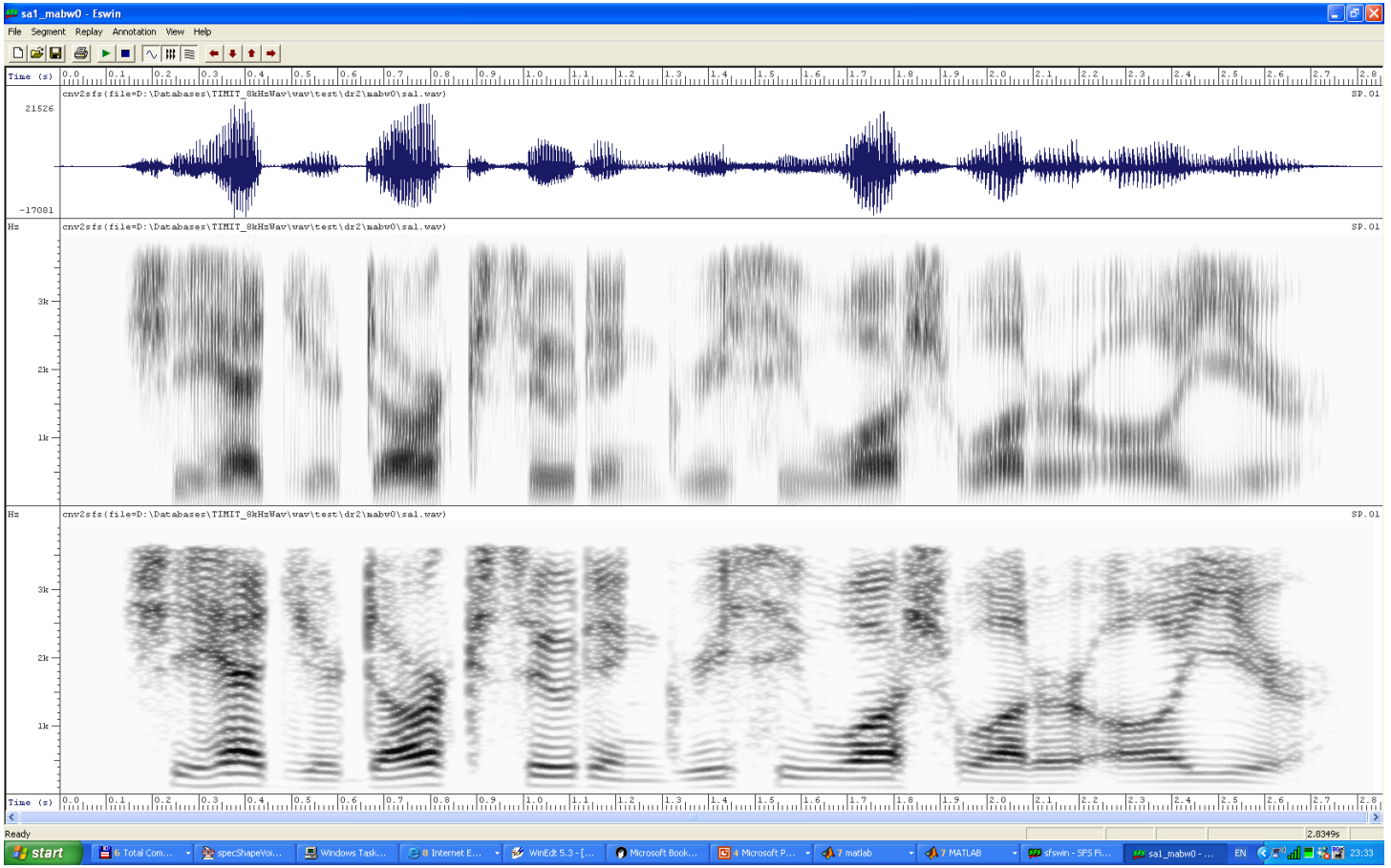
- If the sample rate is *Fs* samples per second and the window length is *N* samples, then for a Hamming window, **the implicit filter bandwidth is *4·Fs/N***

Spectrogram

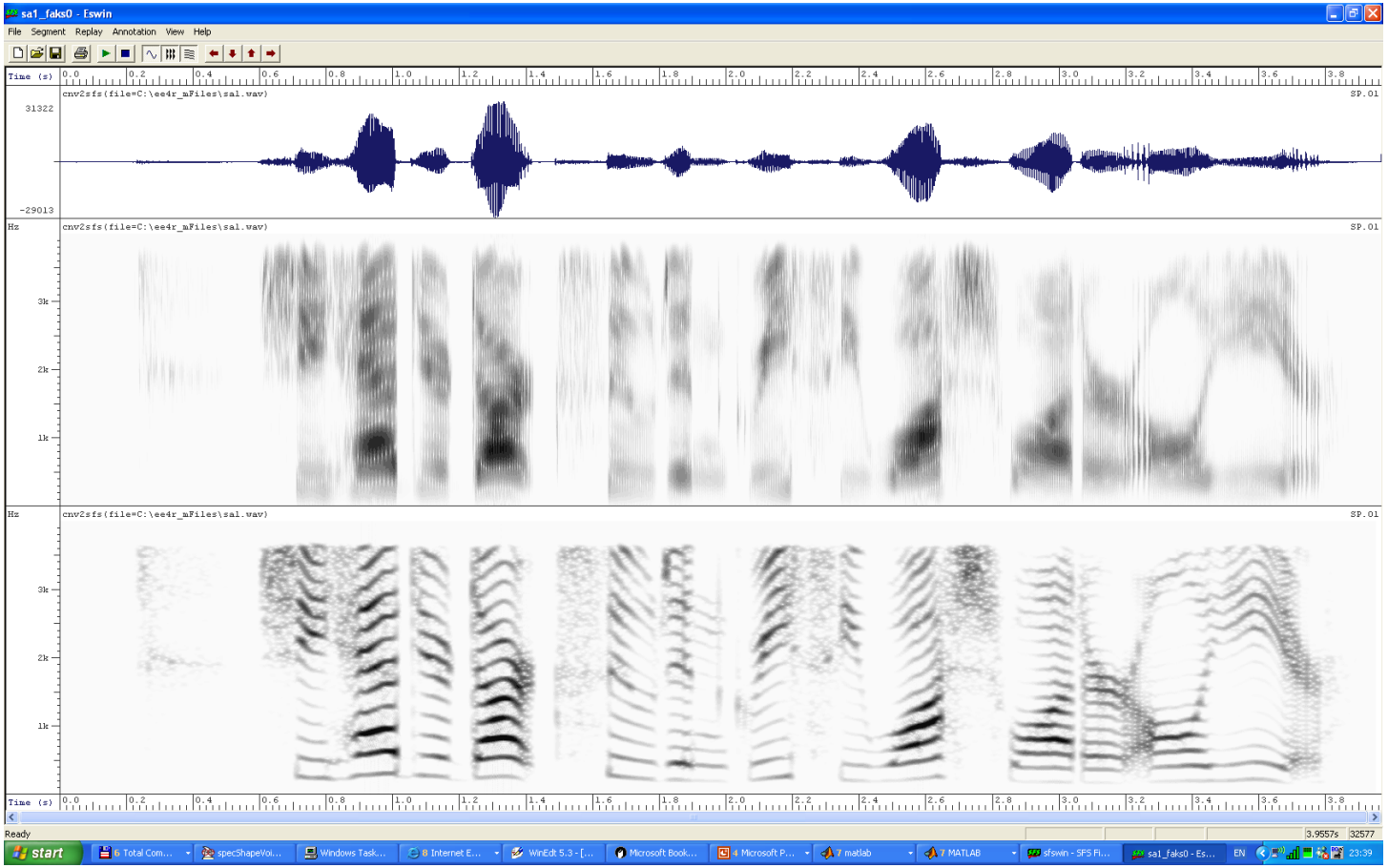
- Spectrogram (time-frequency representation). Calculation:
  - Move the analysis-window in time (i.e. the signal split into frames)
  - Calculate the short-time spectrum for each frame of the signal
  - Collect into vertical ‘slices’
- A vertical ‘slice’ through the spectrogram represents distribution of power with respect to frequency over a time interval centred at *t*



Wide- and Narrow-band Spectrograms



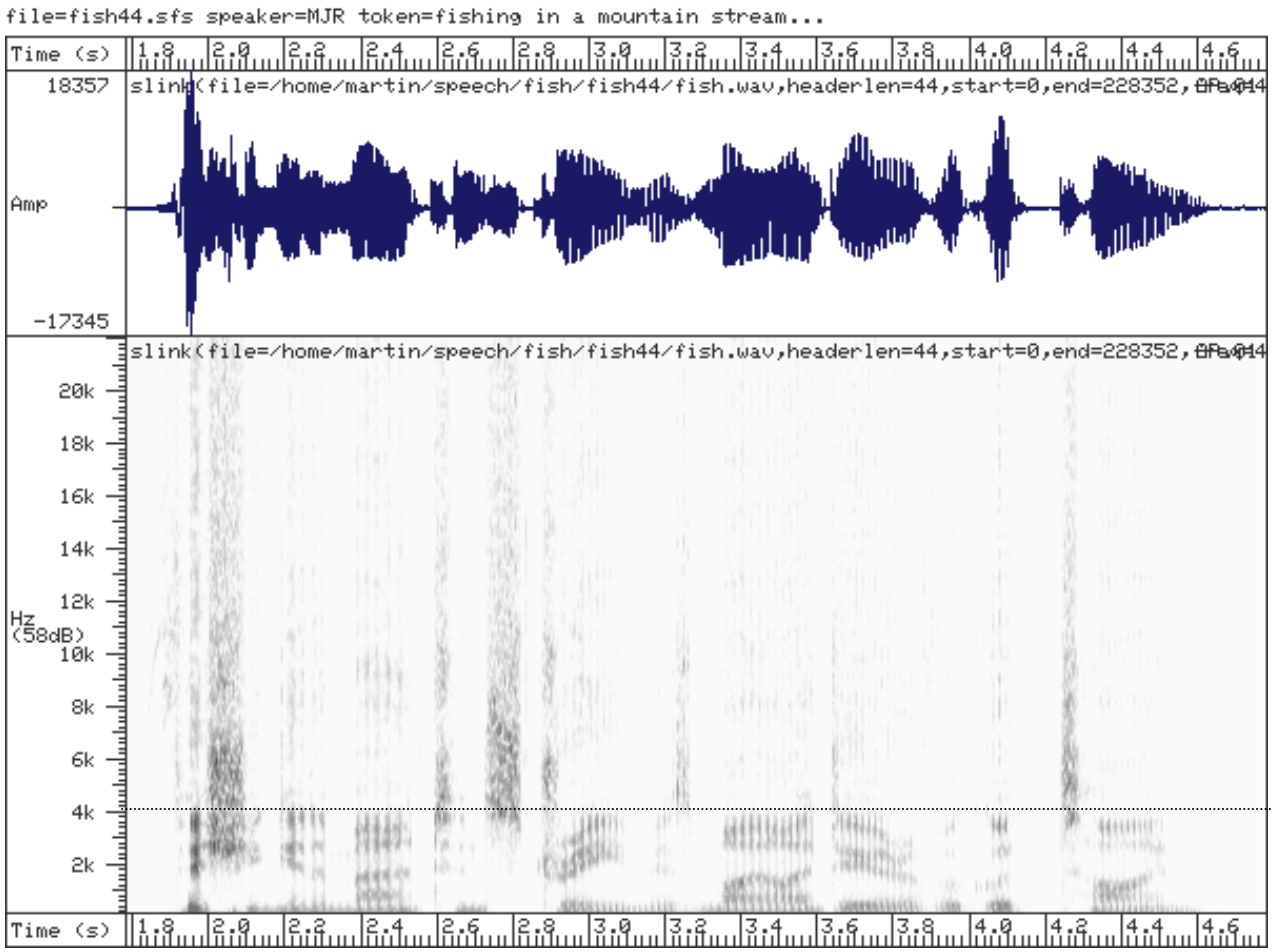
Wide- and Narrow-band Spectrograms



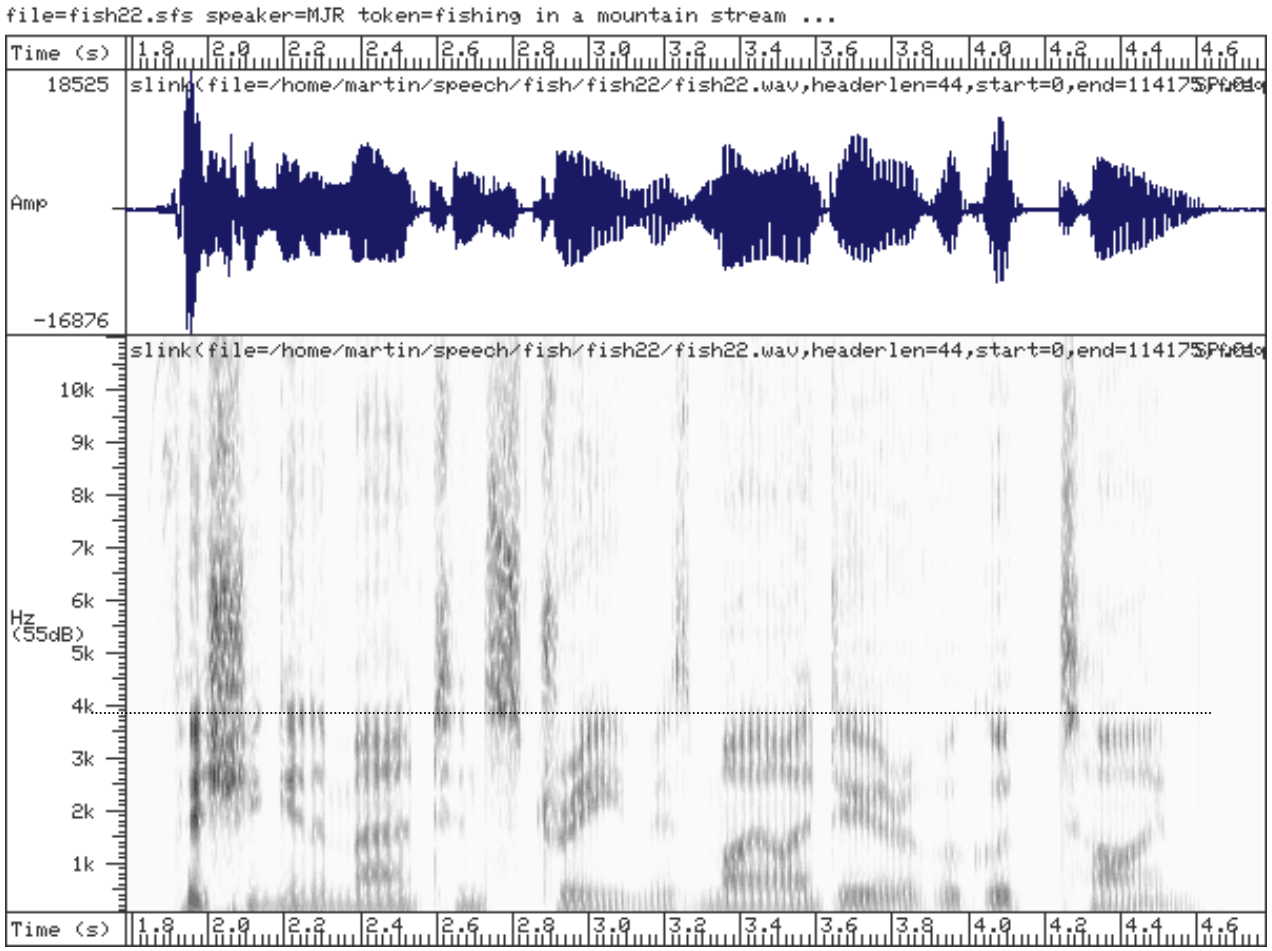
Bandwidth of speech signals

- CD quality speech sampled at 44kHz, giving 22kHz bandwidth
- In the case of speech, almost all of the relevant information lies below 10kHz, so a sample rate of 20kHz gives good quality
- Restricting the bandwidth to 3.75kHz results in intelligible speech, but quality is degraded
- Intelligibility compromised at bandwidths below 3.75kHz

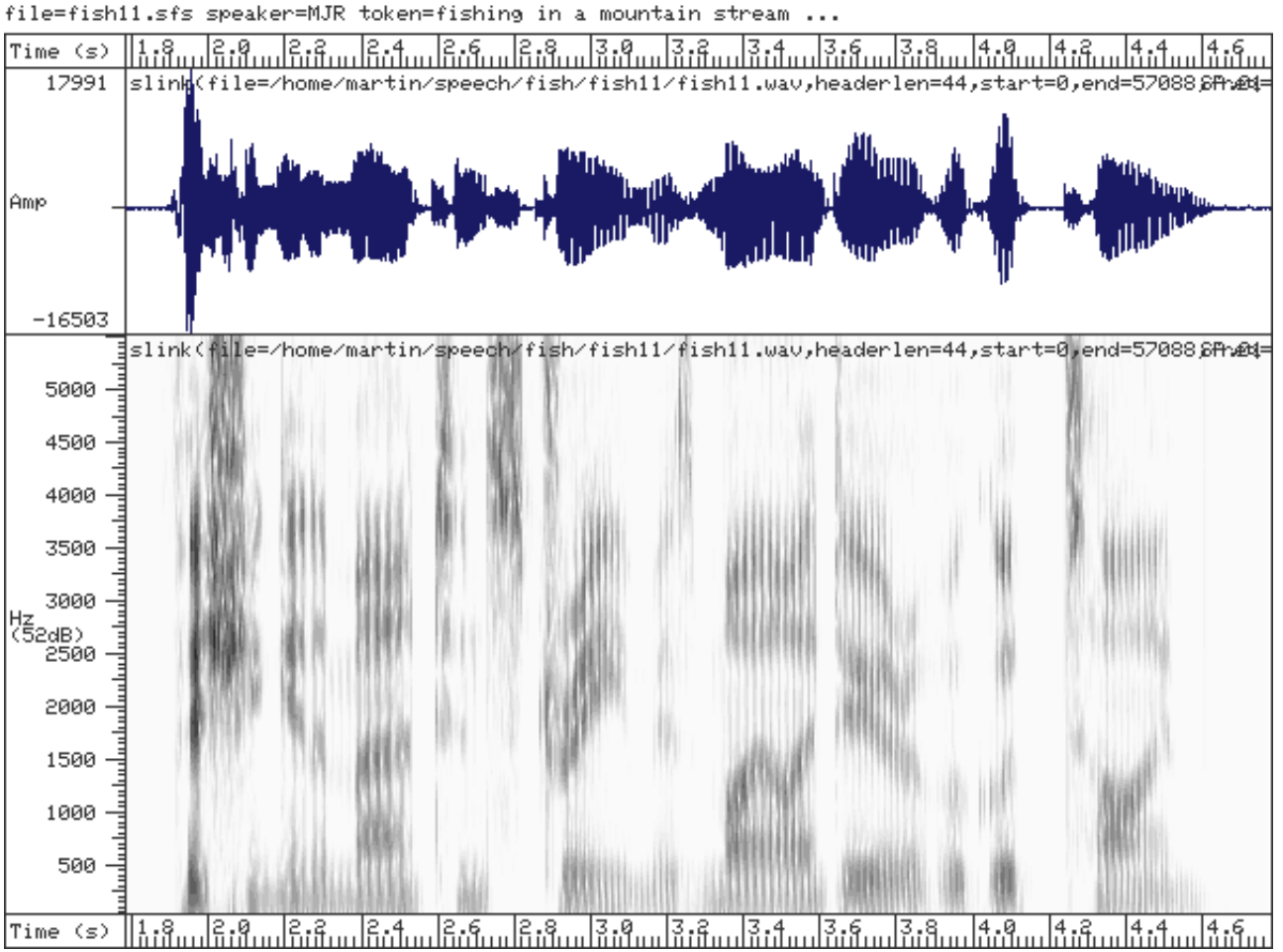
Speech (22kHz bandwidth)



Speech (11kHz bandwidth)



Speech (5.6kHz bandwidth)



Speech (2.8kHz bandwidth)