Short-time Fourier Transform

Tạ Đăng Khoa
Phạm Cao Bằng
Lê Đình Duy

Content

- Overview
  - Recap from previous lectures
  - Why is another Fourier transform needed?
  - The short-time Fourier transform in a nutshell
- Analysis: Fourier-transform view
- Analysis: Filtering view
- Short-time synthesis
- STFT magnitude

Overview

Recap from previous lectures

- Discrete time Fourier transform (DTFT)

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

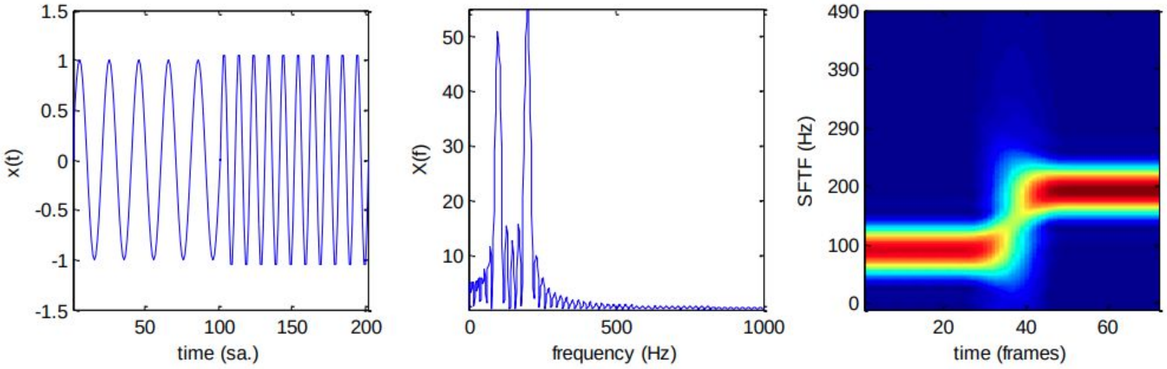
- Discrete Fourier transform (DFT)
  + The DFT is obtained by "sampling" the DTFT at N discrete frequencies
  + N > L

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-j2\pi kn/N}, \qquad k = 0, 1, 2, ..., N-1$$

Overview

Why is another Fourier transform needed?

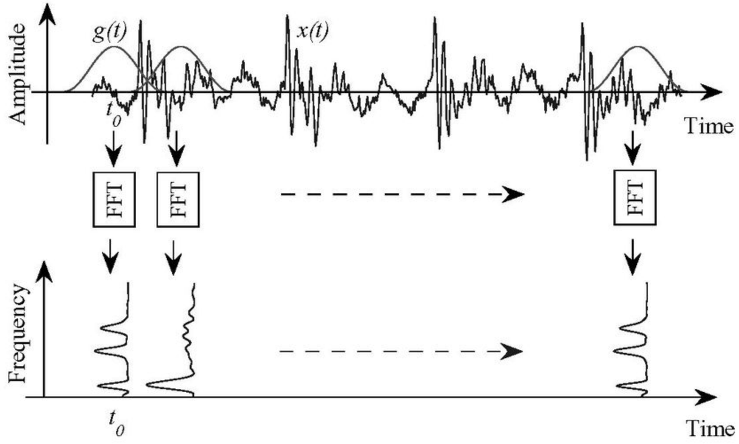
- The spectral content of speech changes over time (non stationary)
  + As an example, formants change as a function of the spoken phonemes
  + Applying the DFT over a long window does not reveal transitions in spectral content
- To avoid this issue, we apply the DFT over short periods of time
  + For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



Overview

The short-time Fourier transform in a nutshell

- 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



Analysis: Fourier-transform view
Windowing function

- 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
- For example, in the speech signal, we define a windowing function $w[n]$, which is generally tapered at its ends to avoid unnatural discontinuities in the speech segment
  - The most common are the Hanning and Hamming windows (raised cosines)

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

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

Analysis: Fourier-transform view

Windowing function

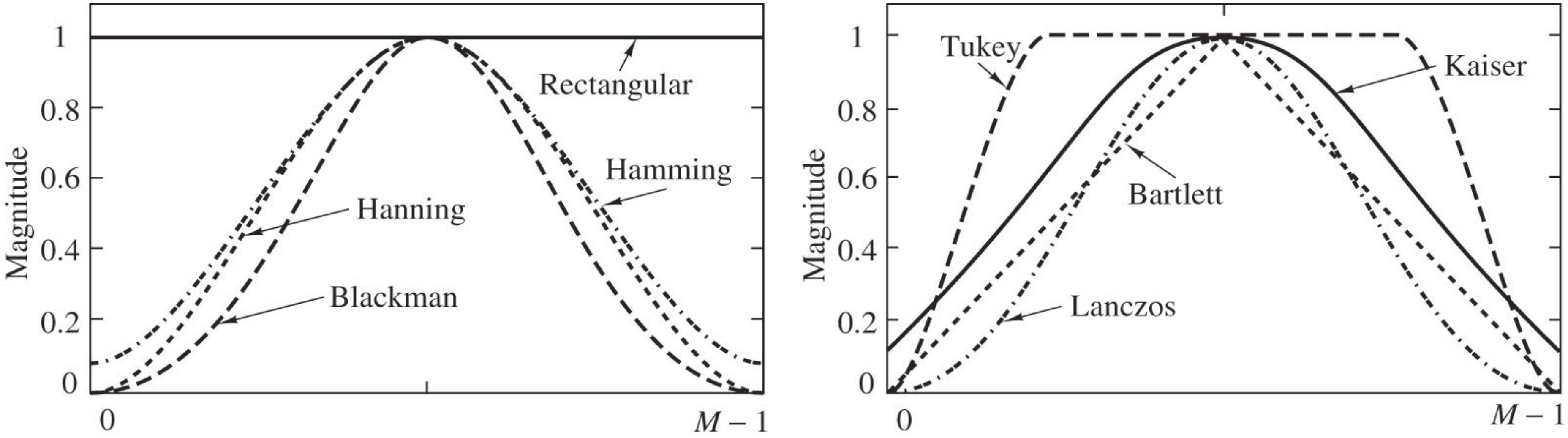


Figure 10.2.3 Shapes of several window functions.

Analysis: Fourier-transform view

Windowing function

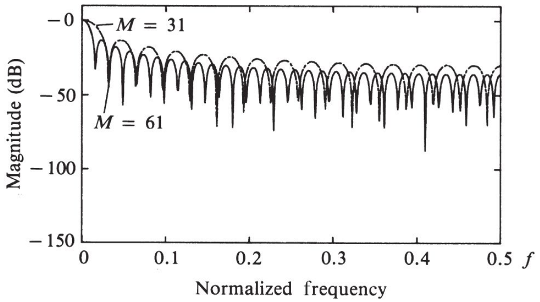


Figure 10.2.2 Frequency response for rectangular window of lengths (a) $M = 31$, (b) $M = 61$.

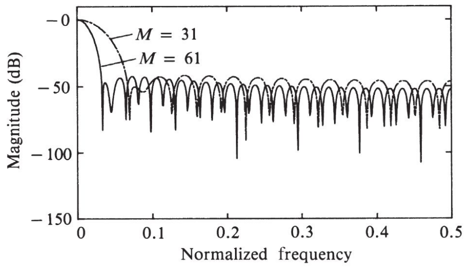


Figure 10.2.5 Frequency responses for Hamming window for (a) $M = 31$ and (b) $M = 61$.

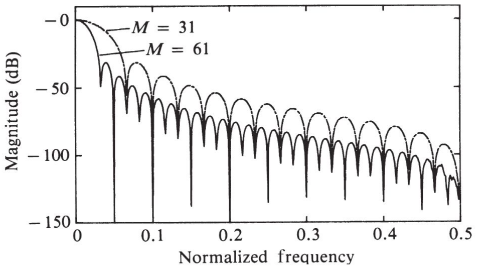


Figure 10.2.4 Frequency responses of Hanning window for (a) $M = 31$ and (b) $M = 61$.

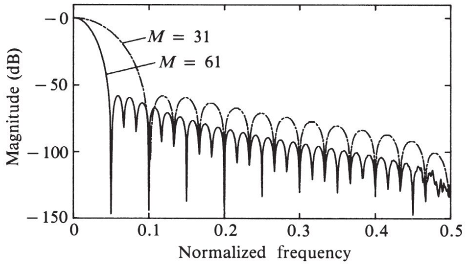


Figure 10.2.6 Frequency responses for Blackman window for (a) $M = 31$ and (b) $M = 61$.

Analysis: Fourier-transform view

Discrete-time Short-time Fourier transform

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

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

  + $f_m(n) = x[n]w[n-m]$ is a short-time section of the speech signal $x[n]$ at time m

Analysis: Fourier-transform view

Discrete STFT

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

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

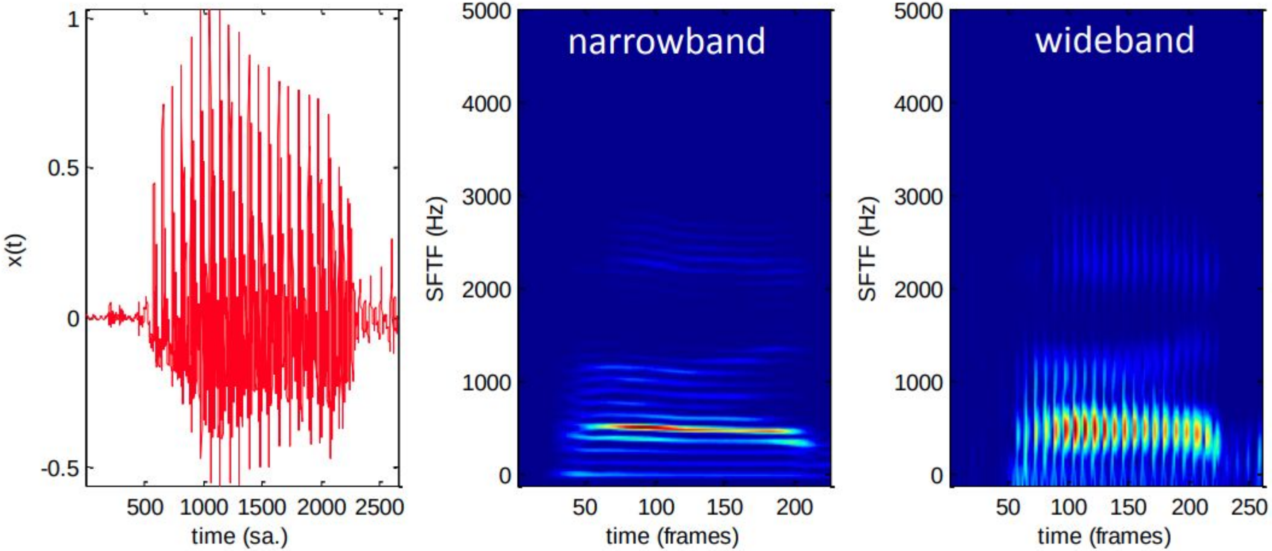
- The spectrogram we saw in previous lectures is a graphical display of the magnitude of the discrete STFT, generally in log scale

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

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

Analysis: Fourier-transform view

Narrowband vs Wideband



Analysis: Filtering view

STFT 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 ω0, and rewrite

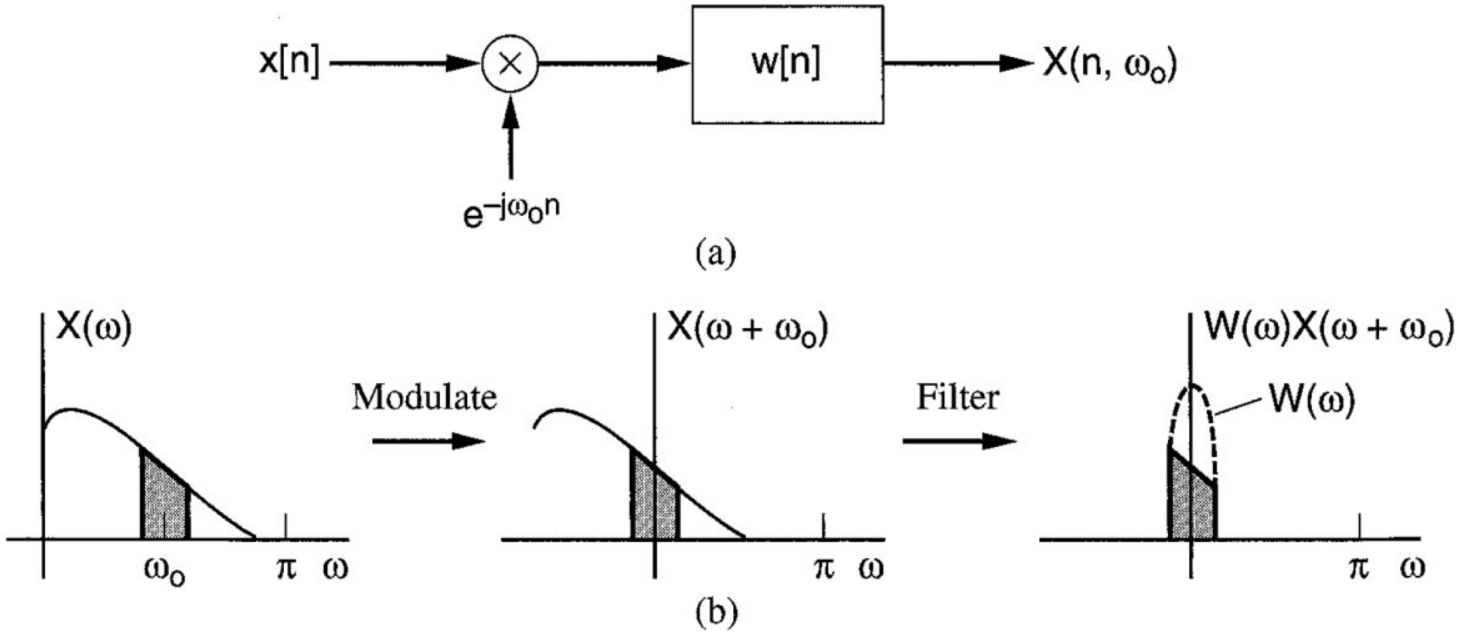
$$X(n, \omega_0) = \sum_{m=-\infty}^{\infty} (x[m]e^{-j\omega_0 m})w[m-n]$$

  + Which can be interpreted as convolution

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

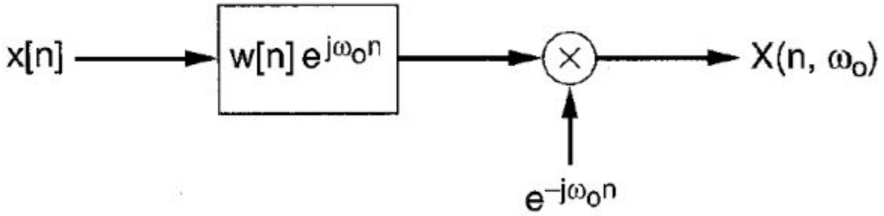
Analysis: Filtering view

STFT as a filtering operation

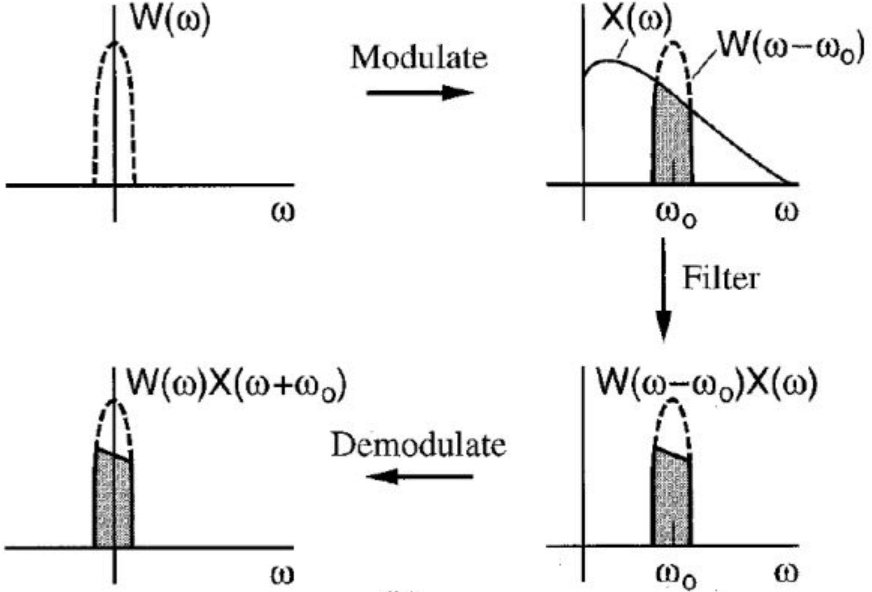


Analysis: Filtering view

STFT as a filtering operation



(a)



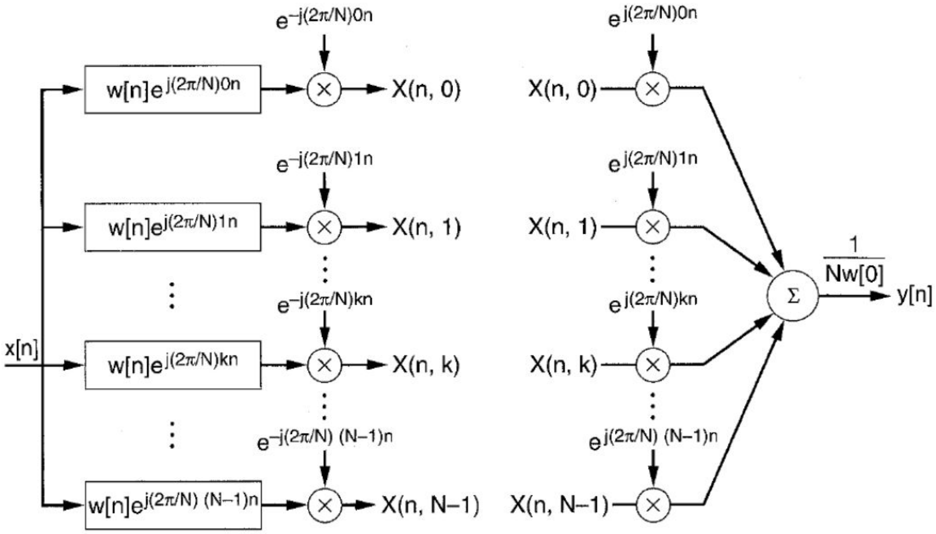
(b)

Analysis: Filtering view

- Interpret the discrete STFT as the output of a filter bank

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

$$\omega = \frac{2\pi}{N}k$$



(a)
analysis

(b)
synthesis

Short-time synthesis

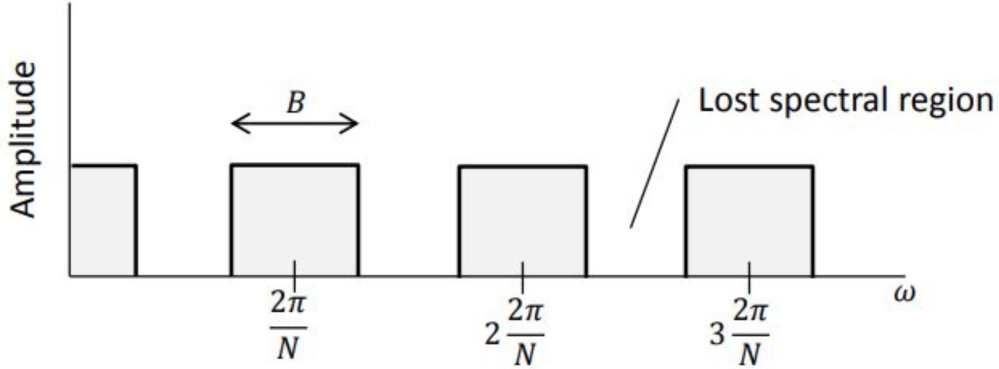
- Recall that $$X(n,\omega) = \sum_{m=-\infty}^{\infty} f_n[m] e^{-j\omega m}$$ $$with\ f_n[m] = x[m]w[m-n]$$

- We obtain $$f_n[n] = x[n]w[0]$$

- We can estimate x[n] as $$x[n] = \frac{1}{2\pi w[0]} \int_{-\pi}^{\pi} X(n,\omega) e^{j\omega n} d\omega$$

Short-time synthesis

- With the discrete STFT:
  + Consider the case where w[n] is band-limited with bandwidth B
  + If 2π/N is greater than B, some of the frequency components in $x\ n$ do not pass through any of the filters of the STFT
  + So the discrete STFT may become non invertible



STFT magnitude

- The spectrogram (STFT magnitude) is widely used in speech
  + For one, evidence suggests that the human ear extracts information strictly from a spectrogram representation of the speech signal
  + Likewise, trained researchers can visually “read” spectrograms, which further indicates that the spectrogram retains most of the information in the speech signal (at least at the phonetic level)
  + Hence, one may question whether the original signal x[n] can be recovered from $|X(n, \omega)|$, that is, by ignoring phase information

STFT magnitude

Inversion of the STFTM

Thanks for Listening