Lecture 03: [Rabiner] Chapter 7. Frequency-Domain Representations

DEEE725 음성신호처리실습

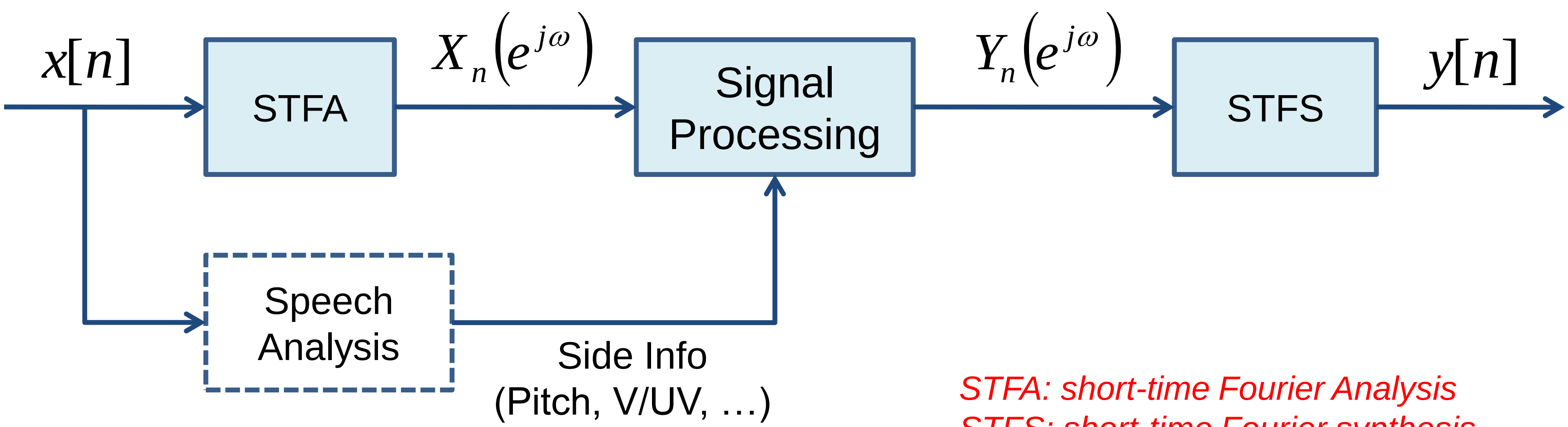
Speech Signal Processing Lab

Instructor: 장길진

Original slides from Lawrence Rabiner

Frequency Domain Processing

- Restoration/Enhancement/Modification:
  - noise and reverberation removal
  - High-pass / Low-pass / Bandpass filering
- Feature extraction:
  - Filterbank energies / Cepstral coefficients



STFA: short-time Fourier Analysis
STFS: short-time Fourier synthesis

Short-Time Fourier Transform

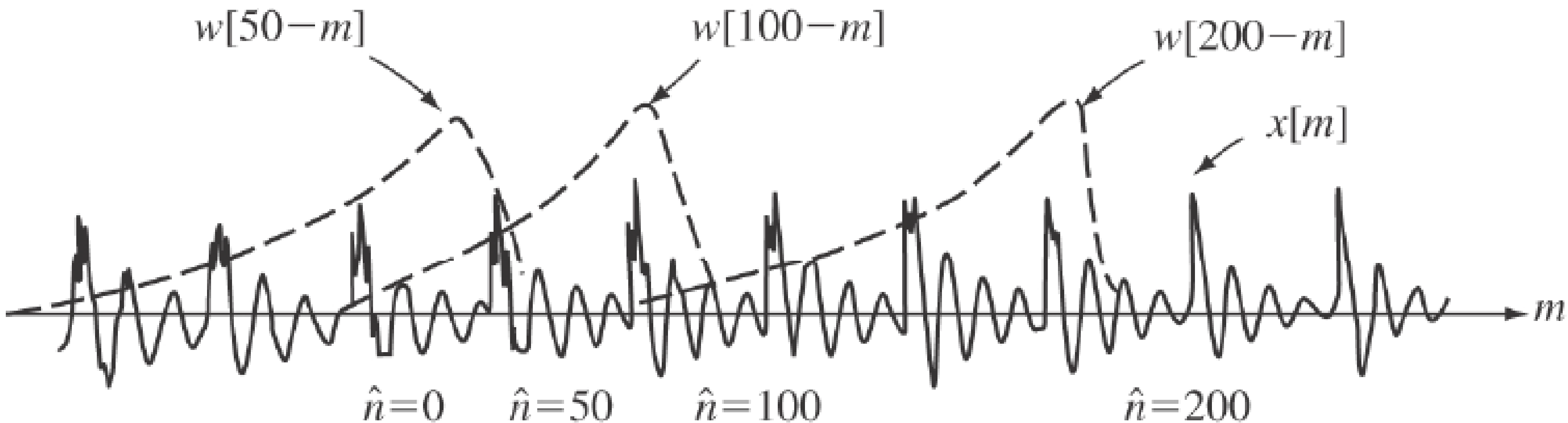
- Speech is not **pure stationary**, i.e., its properties **change with time** (*time-varying*)
  - changes occur at syllabic rates (~10 times/sec)
  - over fixed time intervals of 10-30 milliseconds, properties of speech signals are relatively constant
- Thus a **single representation** based on all the samples of a speech utterance, for the most part, has no meaning
- Instead, we define a **time-dependent Fourier transform (TDFT or STFT)** of speech

Definition of STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

both $\hat{n}$ and $\hat{\omega}$ are variables

- $w(\hat{n}-m)$ is a real window which determines the portion of $x(\hat{n})$ that is used in the computation of $X_{\hat{n}}(e^{j\hat{\omega}})$



STFT Interpretation

- STFT is a function of two variables, the **discrete** time index *n*; **continuous** variable *ω*

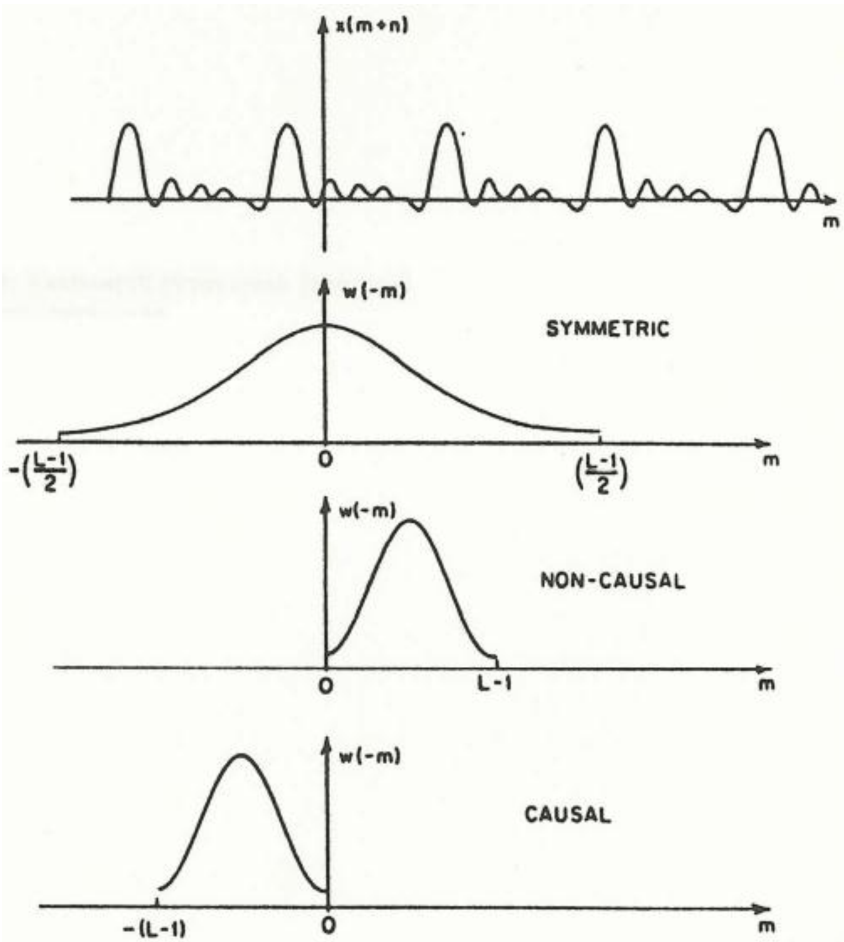
$$X_{\hat{n}}(\hat{\omega}) = \sum_{m=-\infty}^{\infty} x(m)\, w(\hat{n}-m) e^{-j\hat{\omega}m} = DTFT\left(x(m)w(\hat{n}-m)\right)$$

- Alternative form of STFT – origin from windows

$$X_{\hat{n}}(\hat{\omega}) = \sum_{m=-\infty}^{\infty} x(m)\, w(\hat{n}-m) e^{-j\hat{\omega}m}$$

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

Time Origin for STFT



- By using different windows
  - Which one is the best?
  - Causal filter only depends on past, but has delay
  - Matter of where to put the reference point

Alternative Forms of STFT

- Real and imaginary parts
  - when $x(m)$ and $w(n\text{-}m)$ are both real (usually the case), $a_n(\omega)$ is symmetric in ω, and $b_n(\omega)$ is anti-symmetric in ω
- magnitude and phase representation

$$X_{\hat{n}}(\hat{\omega}) = \mathrm{Re}\left[X_{\hat{n}}(\hat{\omega})\right] + j\,\mathrm{Im}\left[X_{\hat{n}}(\hat{\omega})\right] = a_{\hat{n}}(\hat{\omega}) - jb_{\hat{n}}(\hat{\omega})$$

$$a_{\hat{n}}(\hat{\omega}) = \mathrm{Re}\left[X_{\hat{n}}(\hat{\omega})\right]$$

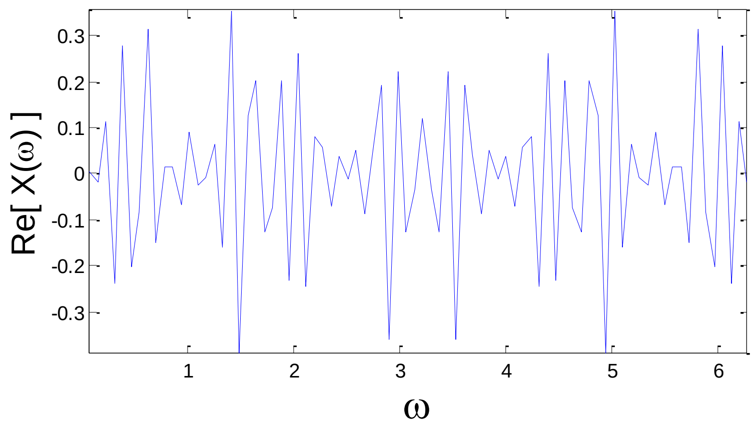
$$b_{\hat{n}}(\hat{\omega}) = -\mathrm{Im}\left[X_{\hat{n}}(\hat{\omega})\right]$$

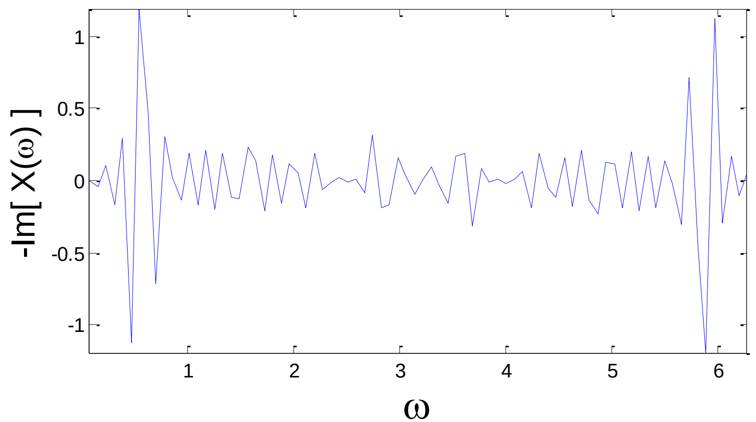
$$X_{\hat{n}}(\hat{\omega}) = \left|X_{\hat{n}}(\hat{\omega})\right| \cdot e^{j\theta_{\hat{n}}(\hat{\omega})}$$

$$\theta_{\hat{n}}(\hat{\omega}) = -j\log\frac{X_{\hat{n}}(\hat{\omega})}{\left|X_{\hat{n}}(\hat{\omega})\right|} = \tan^{-1}\frac{-b_{\hat{n}}(\hat{\omega})}{a_{\hat{n}}(\hat{\omega})}$$

Real Fourier Transform

- when $x(m)$ and $w(n\text{-}m)$ are both real
  - $a_n(\omega)$ is symmetric in ω with respect to π
  - $b_n(\omega)$ is anti-symmetric in ω with respect to π





Role of Window in STFT

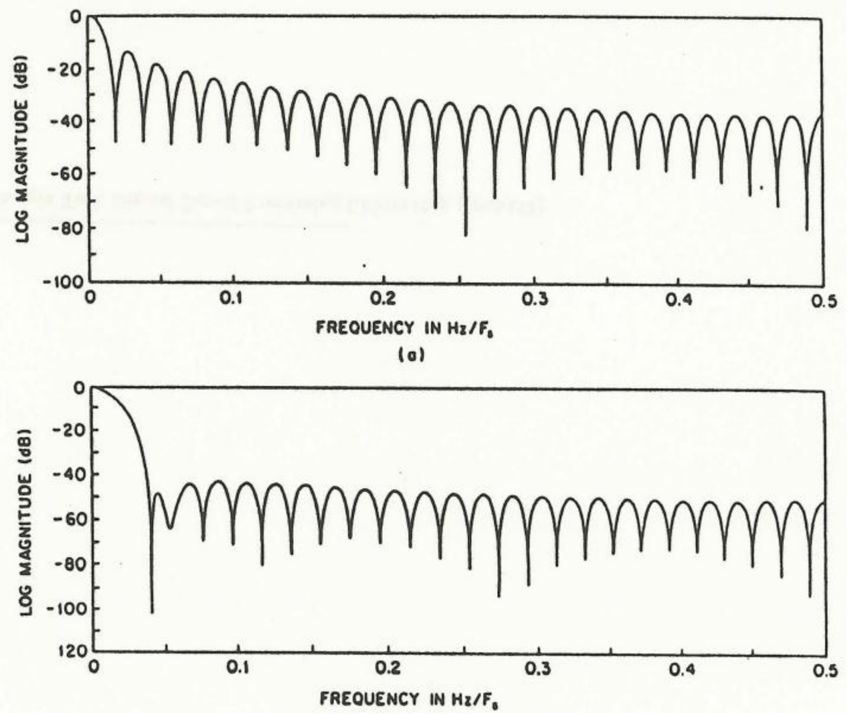
- The window $w(n-m)$ does the following:
  1) chooses portion of $x(m)$ to be analyzed
  2) window shape determines the nature of $X_n(\omega)$

- $X_n(\omega)$ is the convolution of $X(\omega)$ – true spectrum – with the Fourier transform of the shifted window sequence $W(-\omega)\, e^{-j\omega n}$
  – $X_n(\omega)$ is the smoothed version of the short-time spectral properties of $x(n)$

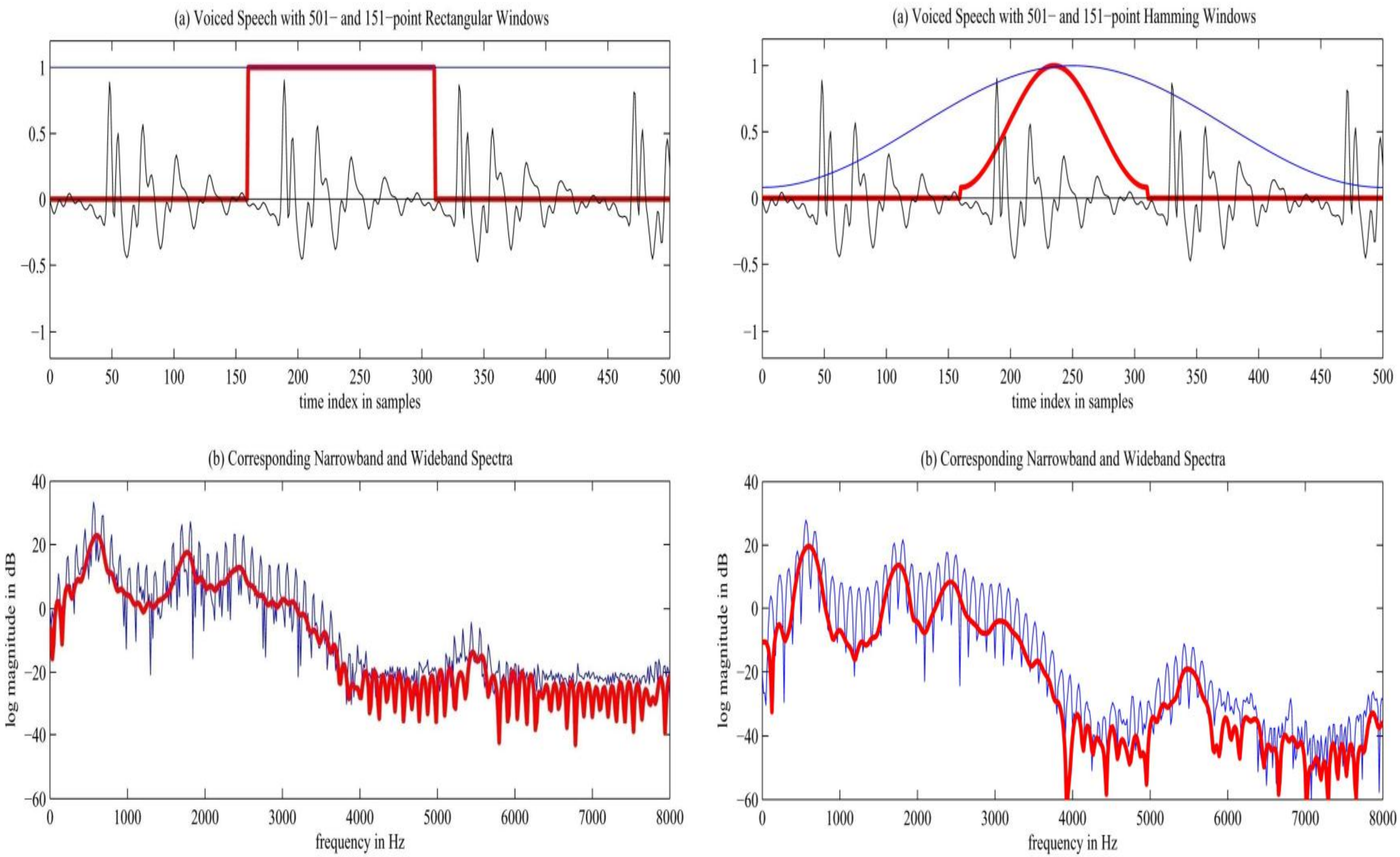
Windows in STFT

- ***Rectangular Window***: flat window of length L samples; first zero in frequency response occurs at F_S/L, with sidelobe levels of -14 dB or lower
- ***Hamming Window***: raised cosine window of length L samples; first zero in frequency response occurs at $2F_S/L$, with sidelobe levels of -40 dB or lower

Frequency responses



Rectangular and Hamming Windows



Discrete STFT

- Terminology
  - Frame (window): the analysis unit
  - Frame size (window size): the size of a single frame; either in time or number of samples (***Nf***)
  - Shift length: how much to slide, 1/*Fs* second (1 sample) to frame size (***Ns***)
  - FT size: number of FT sampling in frequency (***Nft*** ≥ ***Nf***)
  - Frequency index: discretized frequency number (***k***)

Discrete STFT

- Define STDFT
  - Reduce to N_f points
  - Sample ω by N_{ft} times in $[0\ 2\pi)$
  - Substitute ω with $\omega(k)$
  - Consider the frame not from a long signal but just a fixed length sequence
  - Slide the frame by the shift size

$$X_{\hat{n}}(\hat{\omega}) = \sum_{m=0}^{N_f-1} x(n-m)w(m)e^{-j\hat{\omega}(\hat{n}-m)}$$

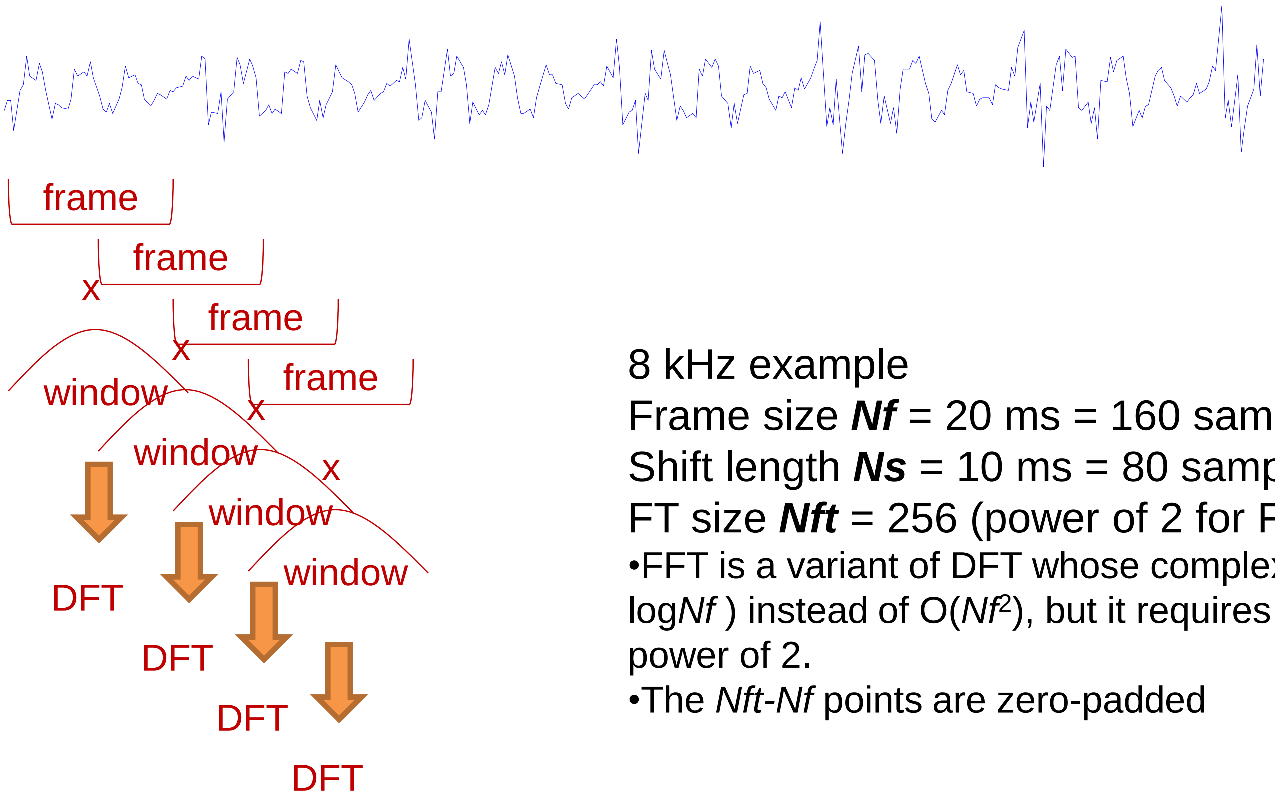
$$\hat{\omega}(k) = \frac{2\pi}{N_{ft}}(k-1), \quad k = 1, \ldots, N_{ft}$$

$$X_{\hat{n}}(k) = \sum_{m=0}^{N_f-1} x(\hat{n}-m)w(m)e^{-j\frac{2\pi}{N_{ft}}(k-1)(\hat{n}-m)}$$

$$\hat{n} \leftarrow 0, \quad m \leftarrow -m$$

$$X(k) = \sum_{m=0}^{N_f-1} x(m)w(-m)e^{j\frac{2\pi}{N_{ft}}(k-1)(m)}$$

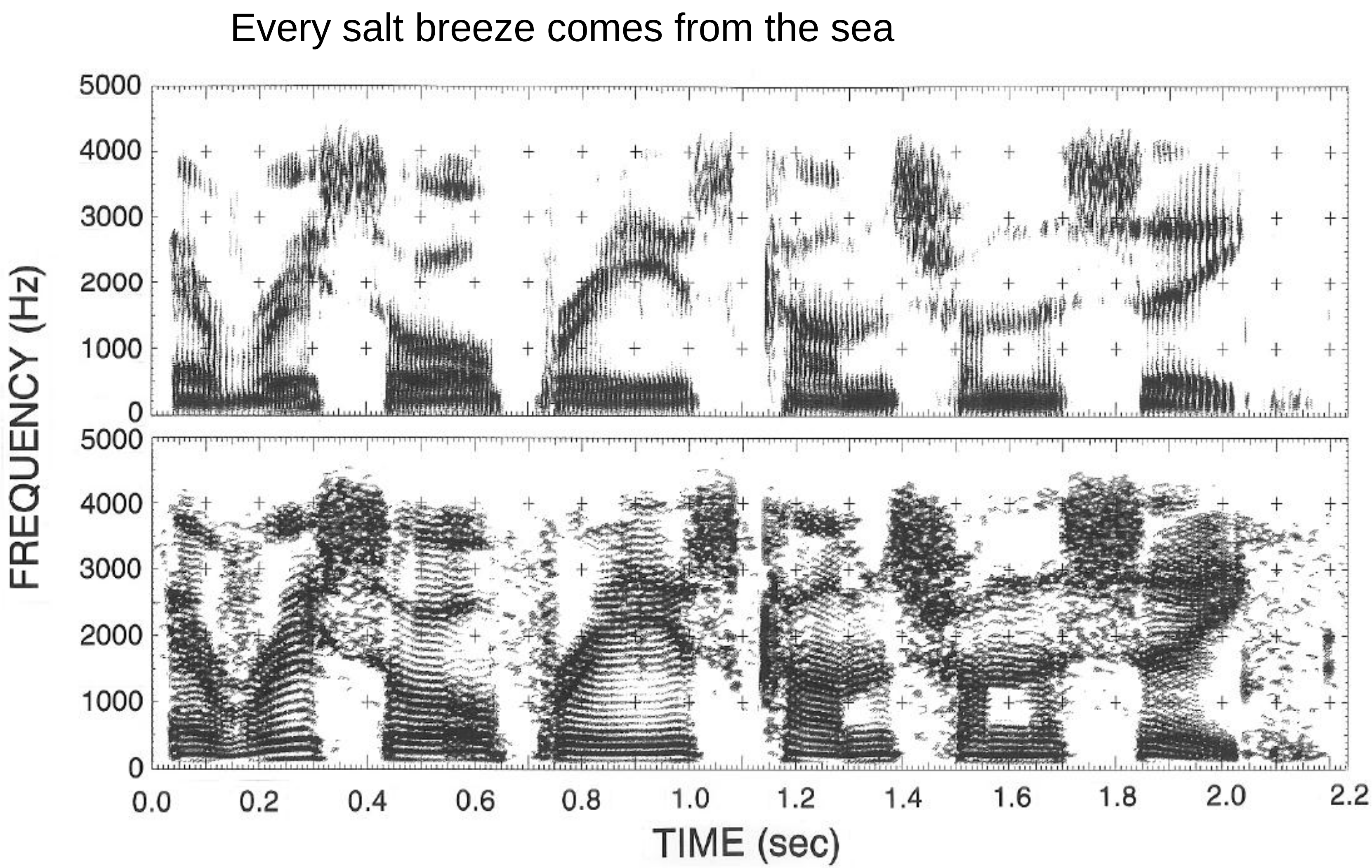
Illustration of STDFT



8 kHz example
Frame size ***Nf*** = 20 ms = 160 samples
Shift length ***Ns*** = 10 ms = 80 samples
FT size ***Nft*** = 256 (power of 2 for FFT*)

•FFT is a variant of DFT whose complexity is $O(Nf \log Nf)$ instead of $O(Nf^2)$, but it requires *Nf* to be power of 2.
•The *Nft*-*Nf* points are zero-padded

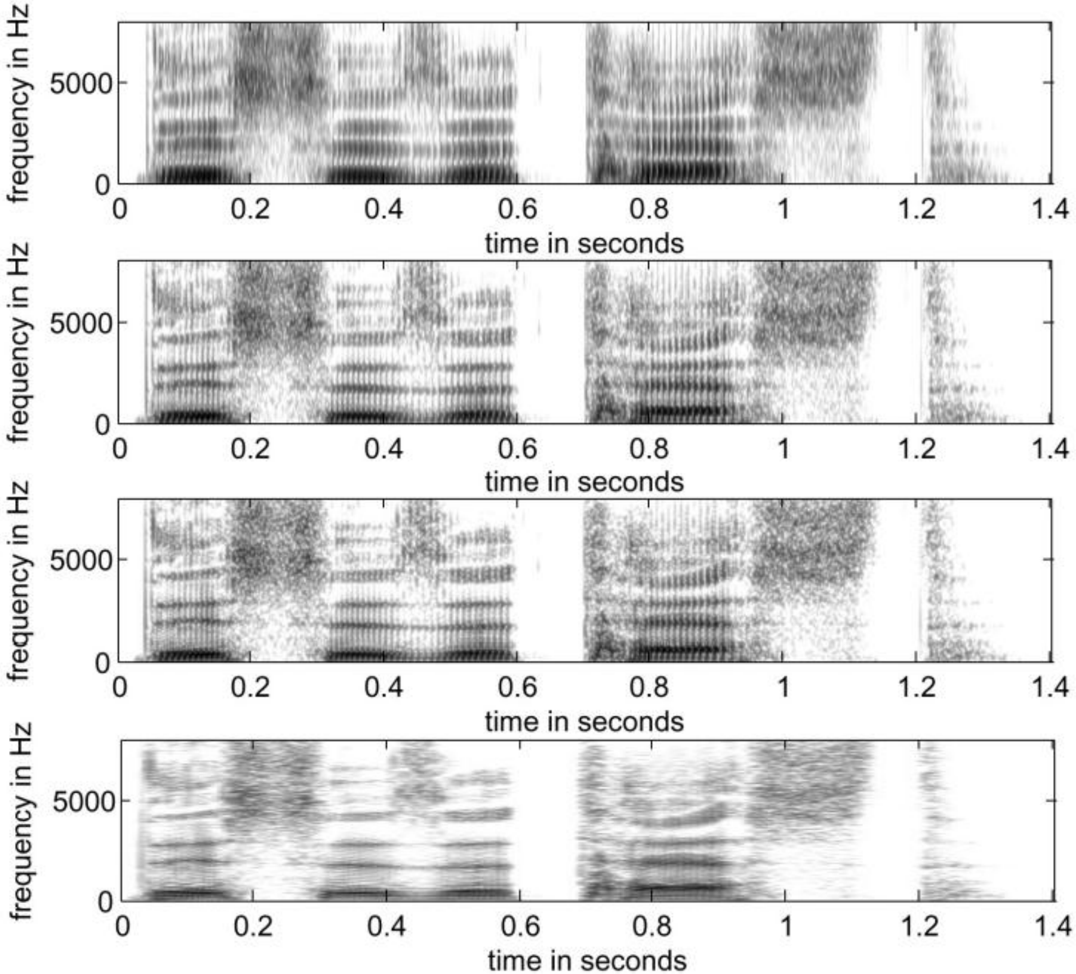
Spectrogram Display



Digital Speech Spectrograms

- Speech Parameters ("This is a test"):
  - sampling rate: 16 kHz
  - speech duration: 1.406 seconds
  - speaker: male
- Wideband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 6 ms (96 samples)
  - analysis window shift: 0.625 ms (10 samples)
  - FFT size: 512
- Narrowband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 60 ms (960 samples)
  - analysis window shift: 6 ms (96 samples)
  - FFT size: 1024

Digital Speech Spectrograms



Top Panel:
3 msec (48 samples) window

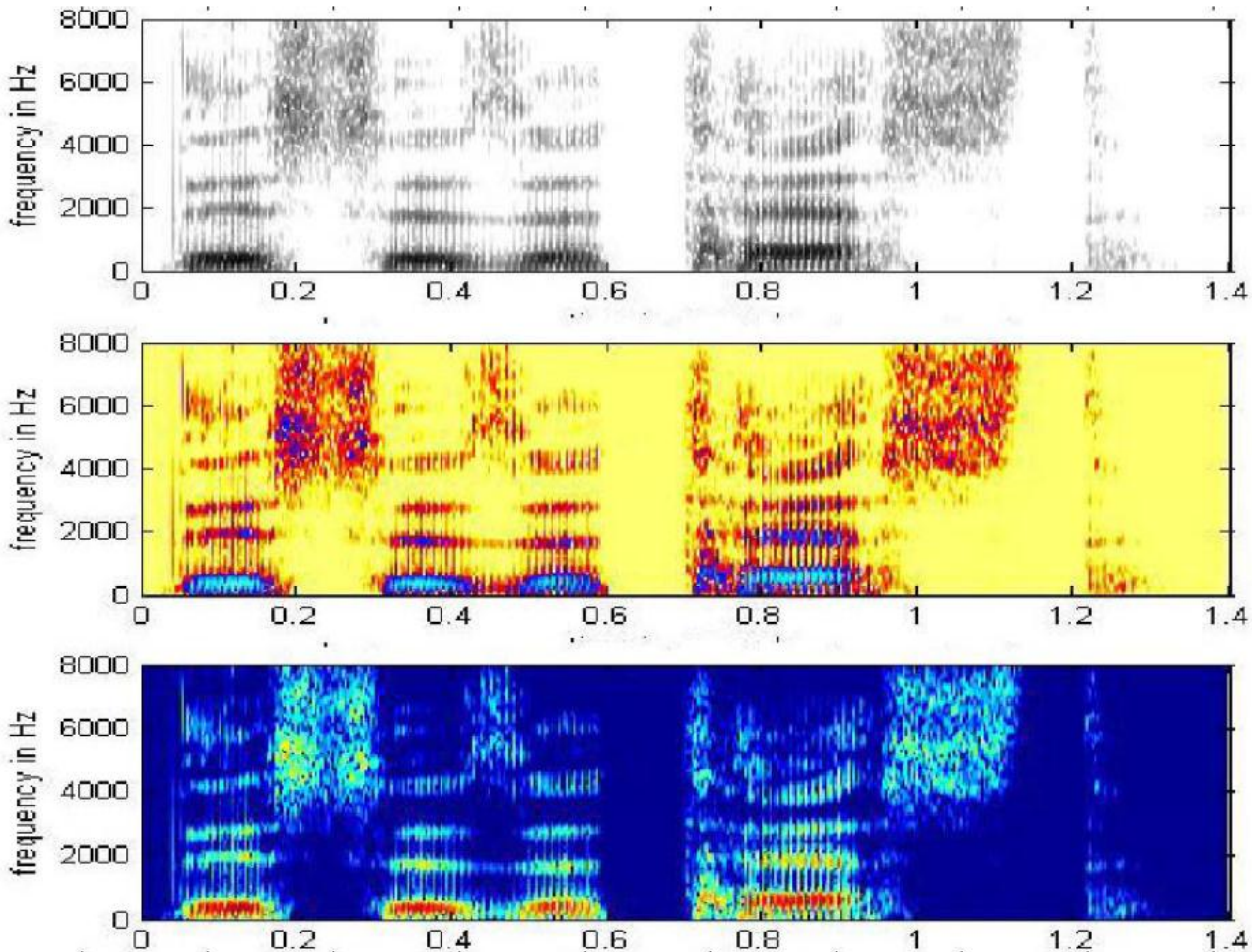
Second Panel:
6 msec (96 samples) window

Third Panel:
9 msec (144 sample) window

Fourth Panel:
30 msec (480 sample) window

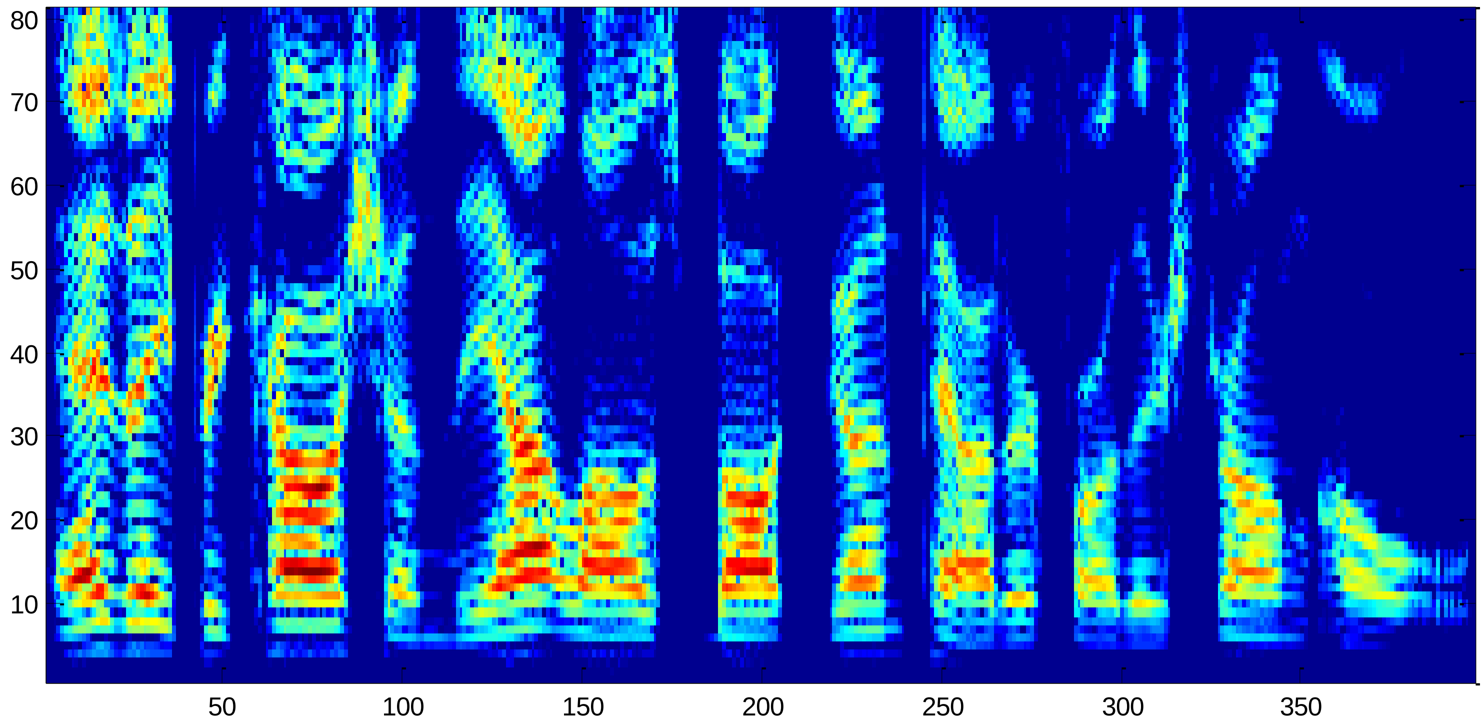
61

Color Display



MATLAB Exercises

```
% file drawSpectrogram.m
```



ELEC747 Speech Signal Processing

Gil-Jin Jang

END OF CHAPTER 7. FREQUENCY-DOMAIN REPRESENTATIONS