Digital Audio and Speech Processing

(Sayısal Ses ve Konuşma İşleme)

Prof. Dr. Nizamettin AYDIN

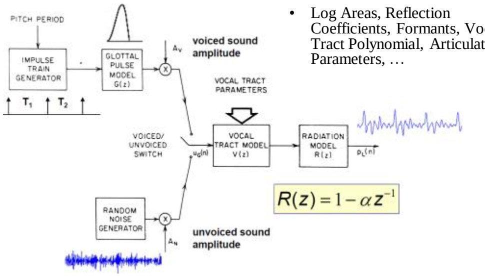
naydin@yildiz.edu.tr

nizamettinaydin@gmail.com

http://www3.yildiz.edu.tr/~naydin

Frequency Domain Methods in Speech Processing

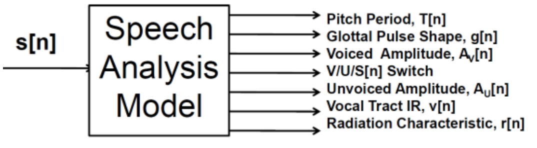
General Synthesis Model



- Log Areas, Reflection Coefficients, Formants, Vocal Tract Polynomial, Articulatory Parameters, …

- Pitch Detection, Voiced/Unvoiced/Silence Detection, Gain Estimation, Vocal Tract Parameter Estimation, Glottal Pulse Shape, Radiation Model

General Analysis Model



- All analysis parameters are time-varying at rates commensurate with information in the parameters;
- We need algorithms for estimating the analysis parameters and their variations over time

Overview of Lecture

- Define time-varying Fourier transform (STFT) analysis method
- Define synthesis method from time-varying FT (filter-bank summation, overlap addition)
- Show how time-varying FT can be viewed in terms of a bank of filters model
- Computation methods based on using FFT
- Application to
  - vocoders, spectrum displays, format estimation, pitch period estimation

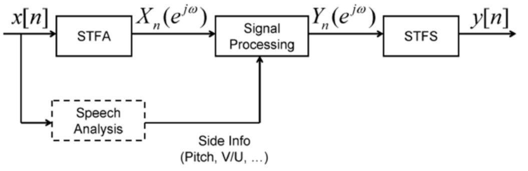
Short-Time Fourier Transform (STFT) Analysis

- Represent signal by sum of sinusoids or complex exponentials as it leads to convenient solutions to problems such as
  - formant estimation,
  - pitch period estimation,
  - analysis-by-synthesis methods
- Such Fourier representations provide
  - convenient means to determine response to a sum of sinusoids for linear systems
  - clear evidence of signal properties that are obscured in the original signal

Why STFT for Speech Signals

- Steady state sounds, like vowels, are produced by periodic excitation of a linear system
  - speech spectrum is the product of the excitation spectrum and the vocal tract frequency response
- Speech is a time-varying signal
- Need more sophisticated analysis to reflect time varying properties
  - Changes occur at syllabic rates
    - ~10 times/sec
  - Over fixed time intervals of 10-30 msec, properties of most speech signals are relatively constant

Frequency Domain Processing



- Coding:
  - Transform, subband, homomorphic, channel vocoders
- Restoration/Enhancement/Modification:
  - Noise and reverberation removal, helium restoration, time-scale modifications (speed-up and slow-down of speech)

Frequency and the DTFT

- Sinusoids

$$x(n) = \cos(\omega_0 n) = \frac{e^{j\omega n} + e^{-j\omega n}}{2}$$

  - where ω_0 is the frequency (in radians) of the sinusoid
- The Discrete-Time Fourier Transform (DTFT)

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

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega = \text{DTFT}^{-1}\{X(e^{j\omega})\}$$

  - where ω is the frequency variable of $X(e^{j\omega})$

DTFT and DFT of Speech

- The DTFT and the DFT for the infinite duration signal could be calculated (the DTFT) and approximated (the DFT) by the following:

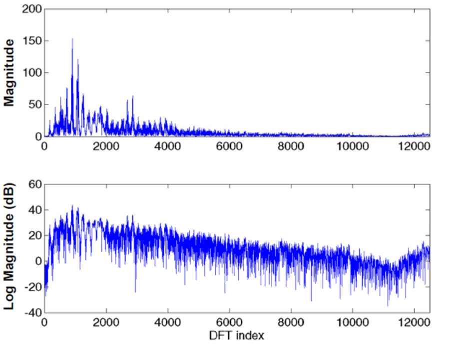
$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \qquad \text{(DTFT)}$$

$$X(k) = \sum_{m=0}^{L-1} x(m)w(m)e^{-j\left(\frac{2\pi}{L}\right)km}, \qquad k = 0,1,\cdots L-1$$

$$= X(e^{j\omega})\Big|_{\omega=\frac{2\pi k}{L}} \qquad \text{(DFT)}$$

- Using a value of L=25000 we get the plot given in the next slide

25000-Point DFT of Speech



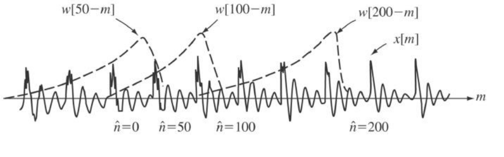
Short-Time Fourier Transform

- Speech is not a stationary signal,
  - i.e., it has properties that change with time
- 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 that changes periodically as the speech properties change over time

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 $X_{\hat{n}}(e^{j\hat{\omega}})$

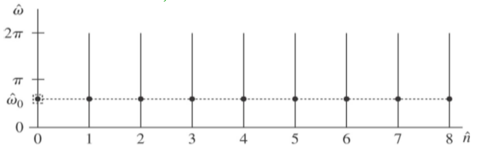


Short Time Fourier Transform

- STFT is a function of two variables,
  - the time index, $\hat{n}$, which is discrete,
  - the frequency variable, $\hat{\omega}$, which is continuous

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

  - where $\hat{n}$ is fixed, $\hat{\omega}$ is variable



Fourier Transform Interpretation

- Consider $X_{\hat{n}}(e^{j\hat{\omega}})$ as the normal Fourier transform of the sequence $x(m)w(\hat{n}-m), -\infty < m < \infty$ for fixed $\hat{n}$.
- The window $w(\hat{n}-m)$ slides along the sequence $x(m)$ and defines a new STFT for every value of $\hat{n}$.
- Conditions for the existence of the STFT:
  - The sequence $x(m)w(\hat{n}-m)$ must be absolutely summable for all values of $\hat{n}$
    - since $|x(\hat{n})| \leq L$ (32767 for 16-bit sampling)
    - since $|w(\hat{n})| \leq 1$ (normalized window levels)
    - since window duration is usually finite
  - $x(m)w(\hat{n}-m)$ is absolutely summable for all $\hat{n}$

Signal Recovery from STFT

- Since for a given value of $\hat{n}$, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform, we can recover the input sequence exactly.
- Since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the windowed sequnce $x(m)w(\hat{n}-m)$, then

$$x(m)w(\hat{n}-m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- Assuming the window satisfies the property that $w(0) \neq 0$ (a trivial requirement), then by evaluating the inverse Fourier transform when $m = \hat{n}$, we obtain

$$x(\hat{n}) = \frac{1}{2\pi w(0)}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Signal Recovery from STFT

$$x(\hat{n}) = \frac{1}{2\pi w(0)}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

- with the requirement that $w(0) \neq 0$, the sequence $x(\hat{n})$ can be recovered exactly from $X_{\hat{n}}(e^{j\hat{\omega}})$, if $X_{\hat{n}}(e^{j\hat{\omega}})$ is known for all values of $\hat{\omega}$ over one complete period.
  - sample-by-sample recovery process
  - Since $X_{\hat{n}}(e^{j\hat{\omega}})$ must be known for every value of $\hat{n}$ and for all $\hat{\omega}$ then
- Can also recover sequence $x(m)w(\hat{n}-m)$ but cannot guarantee that $x(m)$ can be recovered since $w(\hat{n}-m)$ can equal 0

Alternative Forms of STFT

- Alternative forms of $X_{\hat{n}}(e^{j\hat{\omega}})$:
  - Real and Imaginary parts

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \text{Re}[X_{\hat{n}}(e^{j\hat{\omega}})] + j\text{Im}[X_{\hat{n}}(e^{j\hat{\omega}})]$$
$$= a_{\hat{n}}(\hat{\omega}) - jb_{\hat{n}}(\hat{\omega})$$
$$a_{\hat{n}}(\hat{\omega}) = \text{Re}[X_{\hat{n}}(e^{j\hat{\omega}})]$$
$$b_{\hat{n}}(\hat{\omega}) = -\text{Im}[X_{\hat{n}}(e^{j\hat{\omega}})]$$

    - When $x(m)$ and $w(\hat{n}-m)$ are both real can show that $a_{\hat{n}}(\hat{\omega})$ is symmetric in $\hat{\omega}$, and $b_{\hat{n}}(\hat{\omega})$ is anti-symmetric in $\hat{\omega}$
  - Magnitude and Phase

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

    - Can relate $|X_{\hat{n}}(e^{j\hat{\omega}})|$ and $\theta_{\hat{n}}(\hat{\omega})$ to $a_{\hat{n}}(\hat{\omega})$ and $b_{\hat{n}}(\hat{\omega})$

Role of Window in STFT

- The window $w(\hat{n}-m)$ does the following:
  - Chooses portion of $x(m)$ to be analyzed
  - Window shape determines the nature of $X_{\hat{n}}(e^{j\hat{\omega}})$
- Since $X_{\hat{n}}(e^{j\hat{\omega}})$ (for fixed $\hat{n}$) is the normal FT of $x(m)w(\hat{n}-m)$, then if we consider the normal FT's of both $x(m)$ and $w(m)$ individually, we get

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

Role of Window in STFT

- Then for fixed $\hat{n}$, the normal Fourier transform of the product $x(m)w(\hat{n}-m)$ is the convolution of the transforms of $x(m)$ and $w(m)$
- For fixed $\hat{n}$, the FT of $w(\hat{n}-m)$ is $W(e^{-j\hat{\omega}})e^{-j\hat{\omega}\hat{n}}$, thus

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})e^{-j\theta\hat{n}}X(e^{j(\hat{\omega}-\theta)})d\theta$$

- And replacing θ by $-\theta$ gives

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta\hat{n}}X(e^{j(\hat{\omega}+\theta)})d\theta$$

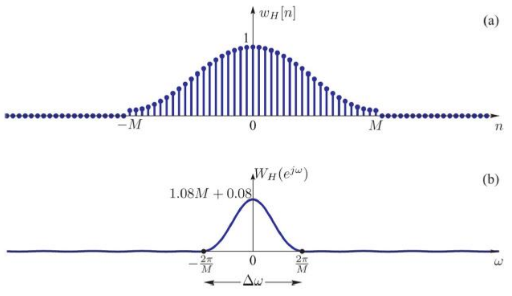
19

Windows in STFT

- For $X_{\hat{n}}(e^{j\hat{\omega}})$ to represent the short-time spectral properties of $x(\hat{n})$ inside the window
  - $W(e^{j\theta})$ should be much narrower in frequency than significant spectral regions of $X(e^{j\theta})$
    - i.e. Almost an impulse in frequency
- Consider Rectangular and Hamming windows, where width of the main spectral lobe is inversely proportional to window length, and side lobe levels are essentially independent of window length
  - 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

20

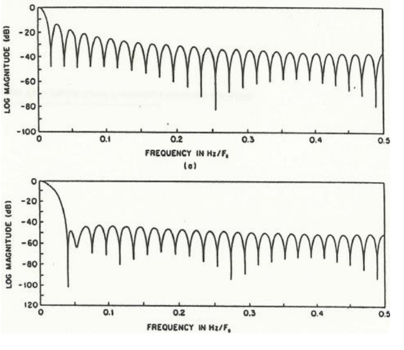
Windows



L=2*M*+*1*-point Hamming window and its corresponding DTFT

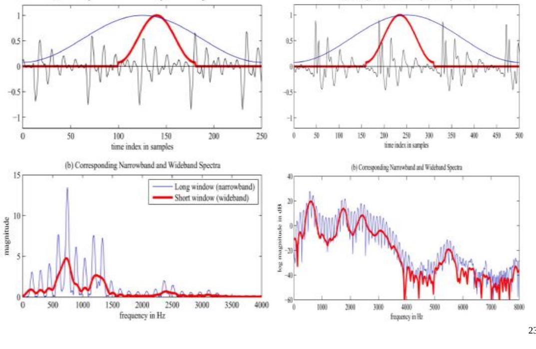
21

Frequency Responses of Windows



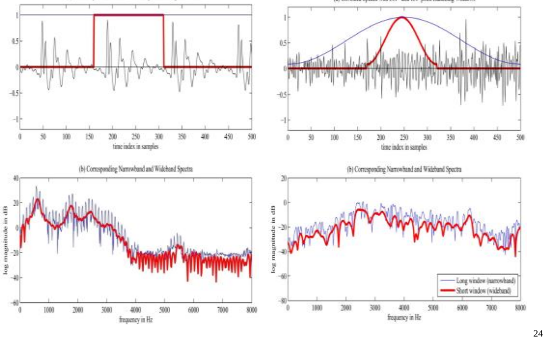
22

Effect of Window Length



23

Effect of Window Length



24

Summary of FT view of STFT

- Interpret $X_{\hat{n}}(e^{j\hat{\omega}})$ as the normal Fourier transform of the sequence $x(m)w(\hat{n}-m), -\infty < m < \infty$
- Properties of this Fourier transform depend on the window
  - Frequency resolution of $X_{\hat{n}}(e^{j\hat{\omega}})$ varies inversly with the length of the window
    - Long windows for high frequency resolution
  - $x(n)$ should be relatively stationary (non-time-varying) during duration of window for most stable spectrum
    - Short windows for high temporal resolution
- As usual in speech processing, there needs to be a compromise between good temporal resolution (short windows) and good frequency resolution (long windows)

25