# Introduction To Speech Processing

## (2nd lecture)

CSLU, OGI , OHSU

June 2010

# Overview

- Time domain processing =>

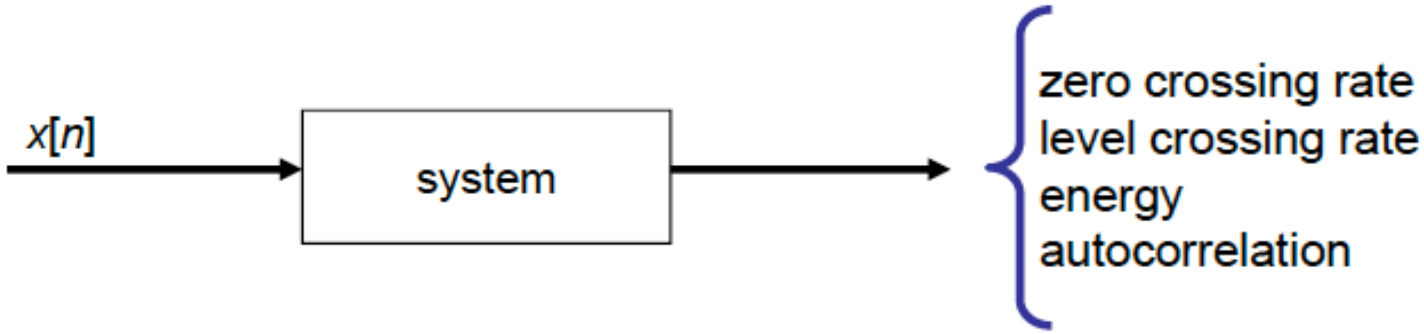
direct operations on the speech waveform

- Frequency domain processing =>

direct operations on a spectral representation of the signal

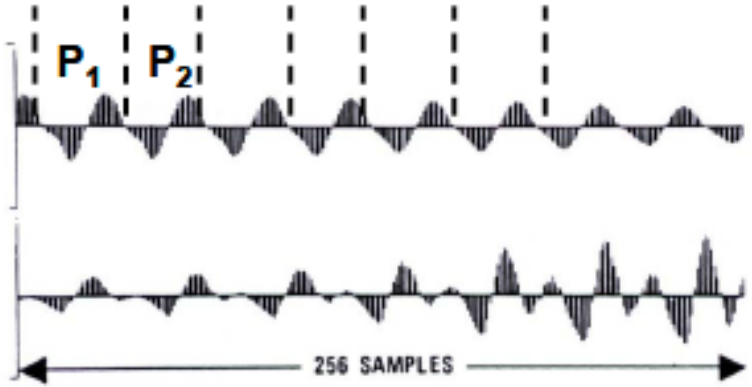
# Time domain processing

- Simple processing
- Enables various types of feature estimation



# Basics in Time domain speech processing

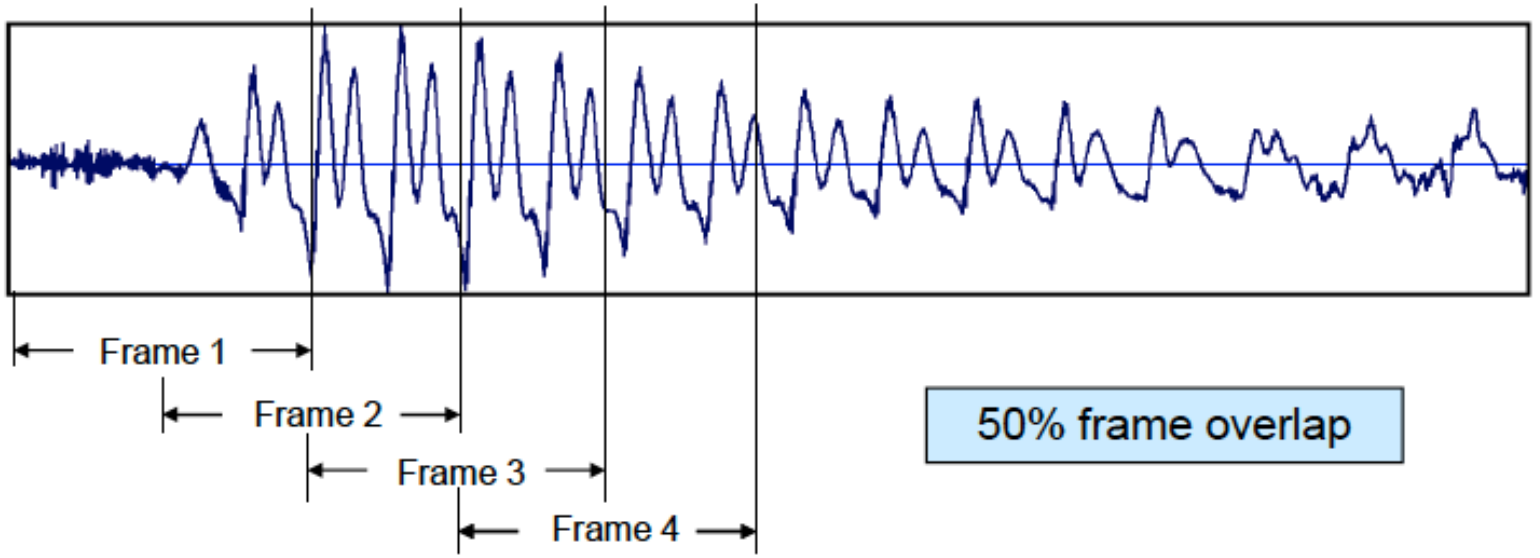
- Properties of speech change with time
- Peak amplitude varies with the sound being produced
- Pitch varies within and across voiced sounds
- Jitter & Shimmer
- Periods of silence where background signals are seen
- The key issue is whether we can create simple time-domain processing methods that enable us to measure/estimate speech representations reliably and accurately



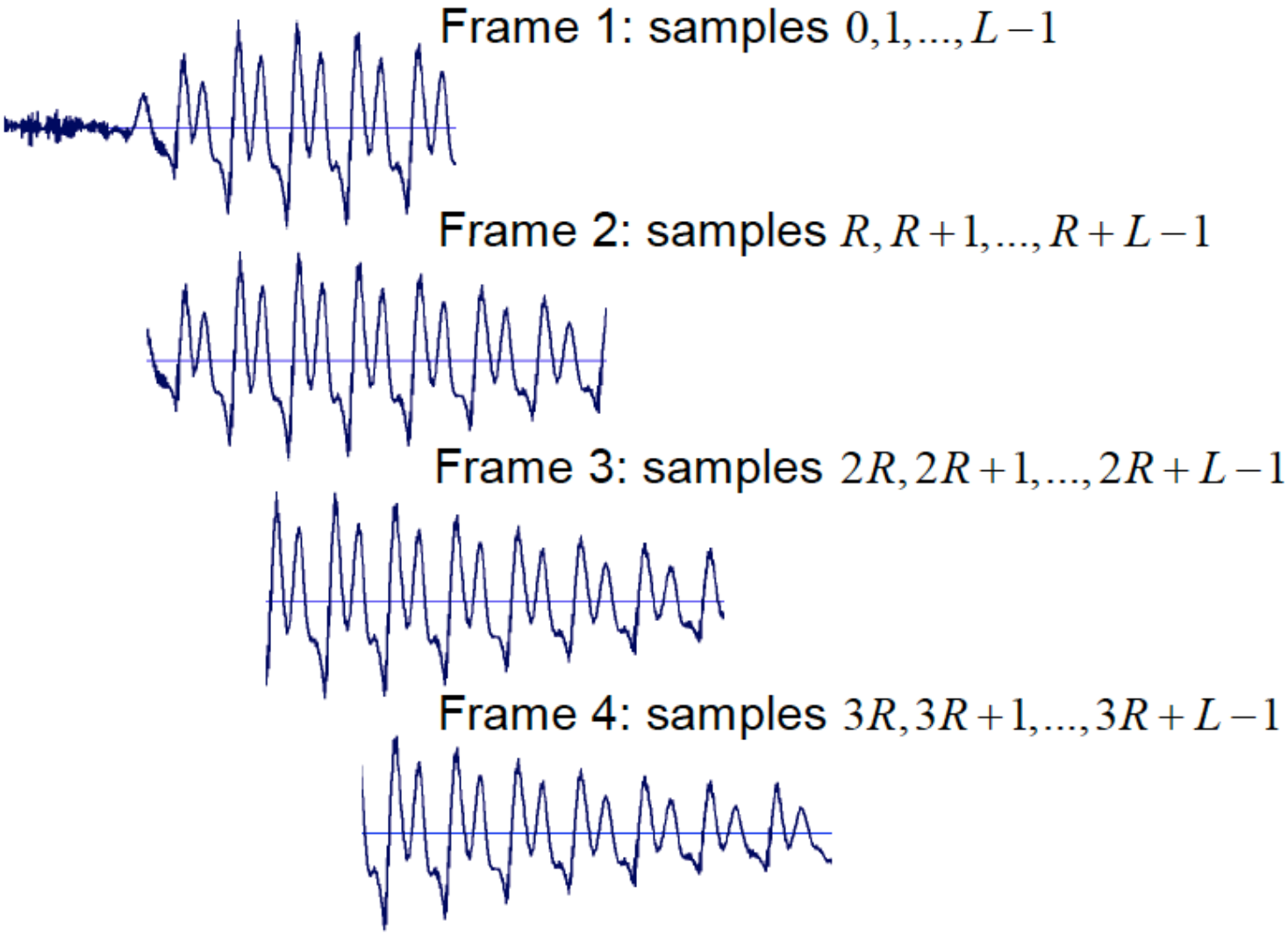
# Fundamental Assumptions

- Because of the slowly varying nature of the speech signal, it is common to process speech in blocks (also called “frames” ) over which
- The properties of the speech waveform can be assumed to remain relatively constant over very short (5-20 msec) intervals
- “short-time” processing methods => Frame-by-Frame Processing
- There is always **uncertainty** in short time measurements and estimates from speech signals

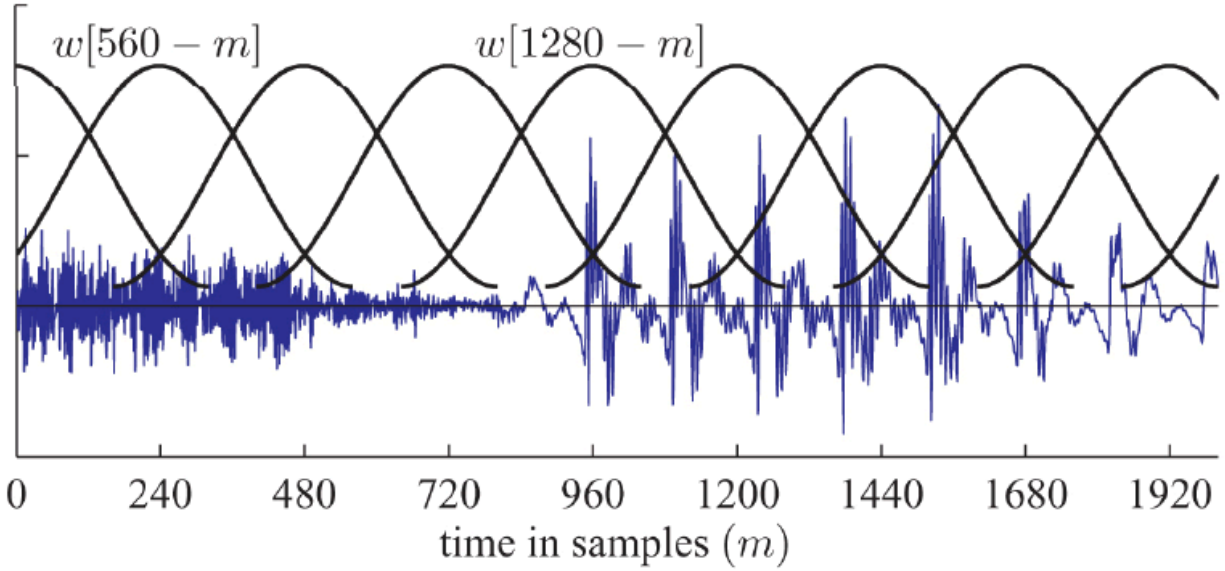
# Frame-by-Frame Processing in Successive Windows




- Speech is processed frame-by-frame in overlapping intervals until entire region of speech is covered by at least one such frame


Frame 1: samples $0,1,...,L-1$
Frame 2: samples $R,R+1,...,R+L-1$
Frame 3: samples $2R,2R+1,...,2R+L-1$
Frame 4: samples $3R,3R+1,...,3R+L-1$

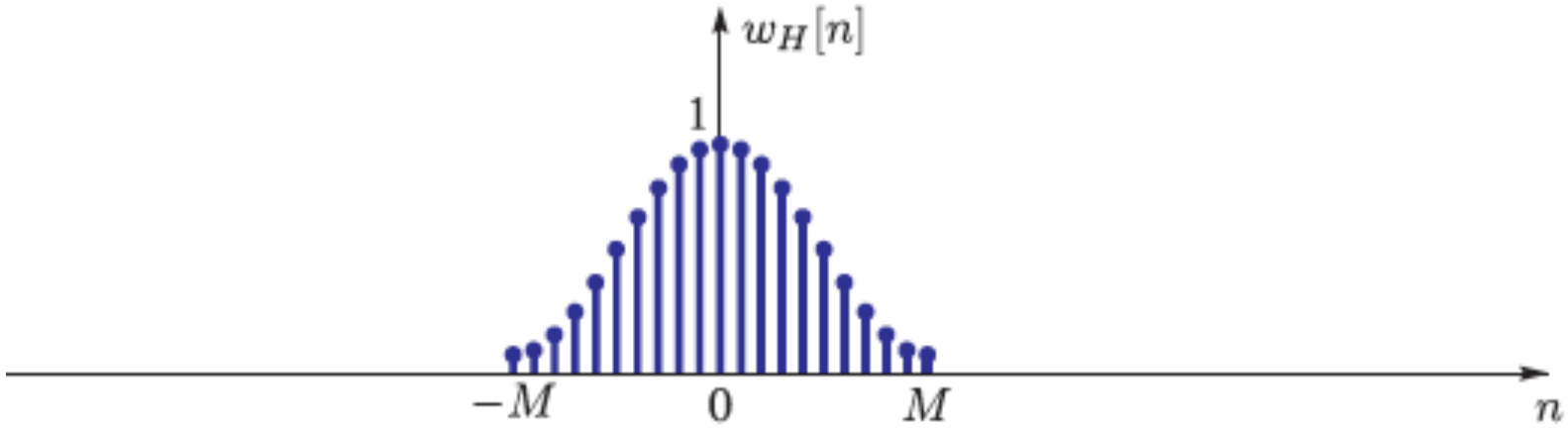
# Frames and Windows



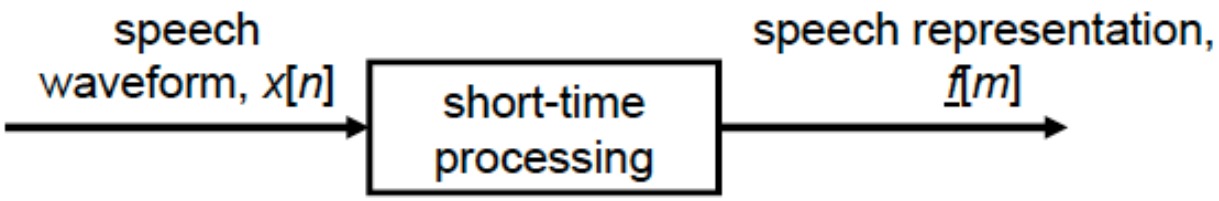

Fs = 16,000 samples/second
Frame rate (overlap percentage) = 10 ms
Window length (Frame length) = 25 ms
=> ( 25ms * 16,000 = 4000 sample/frame)

# Hamming Window

$$w_H[m] = \begin{cases} 0.54 + 0.46\cos(\pi m/M) & -M \leq m \leq M \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$



# Short-Time Processing

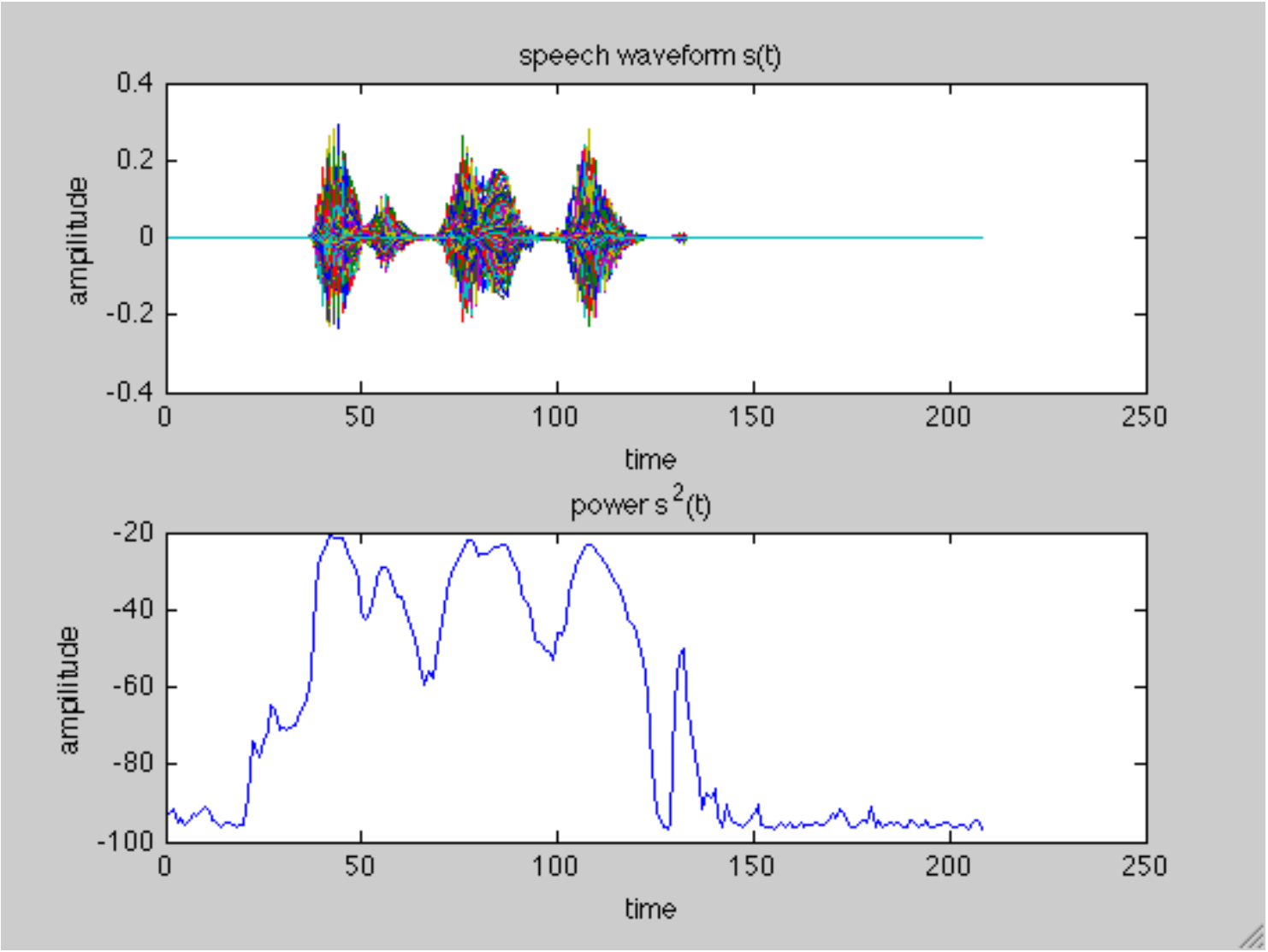



- x[n] = samples of time domain signal
- $\vec{f}[m] = \{f_1[m], f_2[m], \ldots, f_L[m]\}$ frame vectors of signal

# Short-Time Energy and Power

- Simple measure to discriminate voiced/silence (background noise)
- Easy to compute
- Sensitive to background noise energy in case of voiced activity detection
- Short-Time Energy $$E_n = \sum_{n=1}^{N} f[n]^2$$
- Short-Time Power $$P_n = \frac{1}{N} \sum_{n=1}^{N} f[n]^2$$

# Energy based Voiced/silence detection



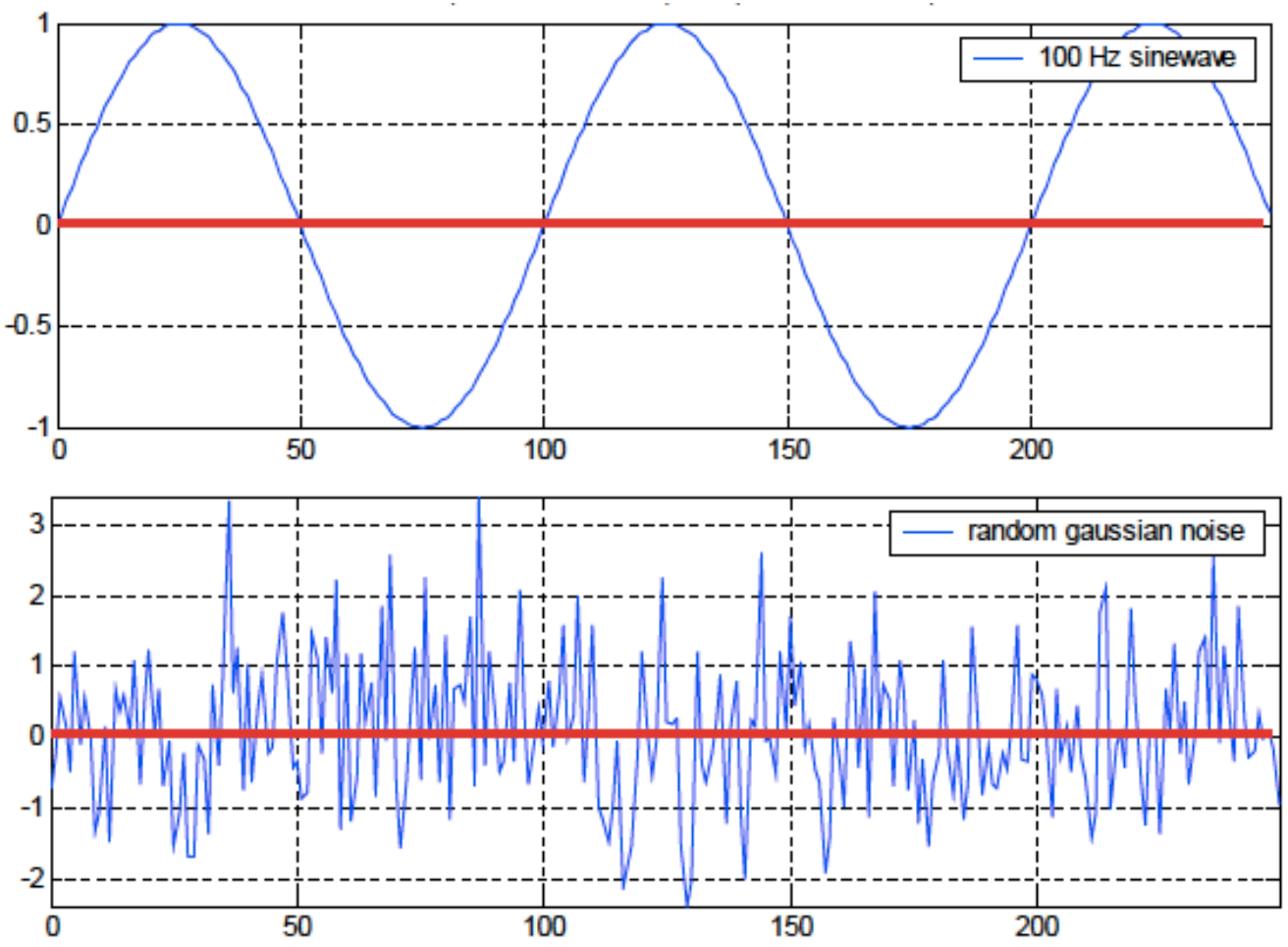
# Zero-Crossing Rate

- ZCR: average of the number of times the speech signal changes sign within the time window.
- Simple to Compute
- Robust against high energy noises in voice activity detection scenario

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{I}\{s_t s_{t-1} < 0\}$$

- **s** is a signal of length T and the indicator function $\mathbb{I}\{A\}$ is 1 if its argument A is true and 0 otherwise.

# Zero-Crossing Rate

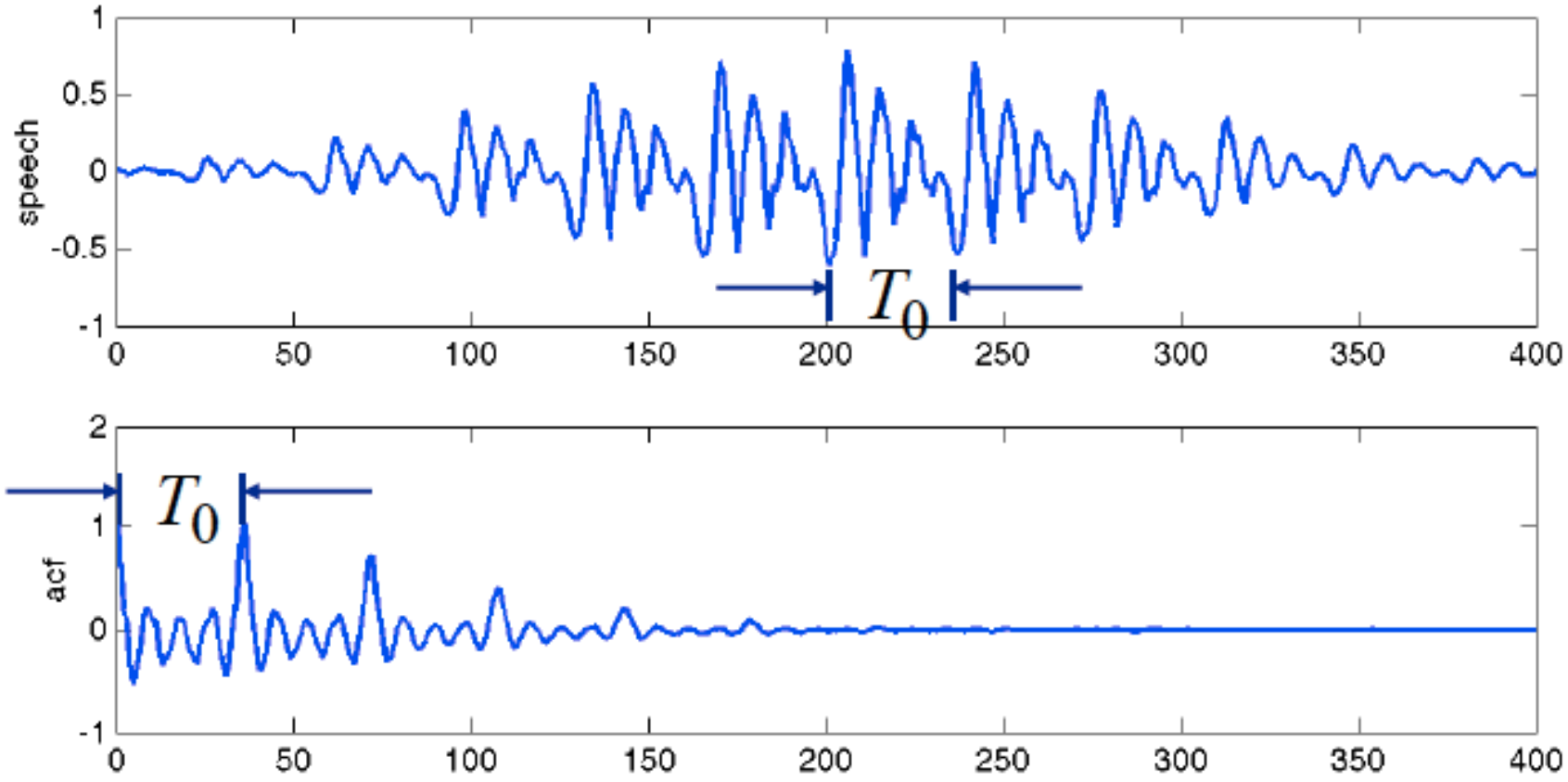


# Short-Time Auto Correlation Function

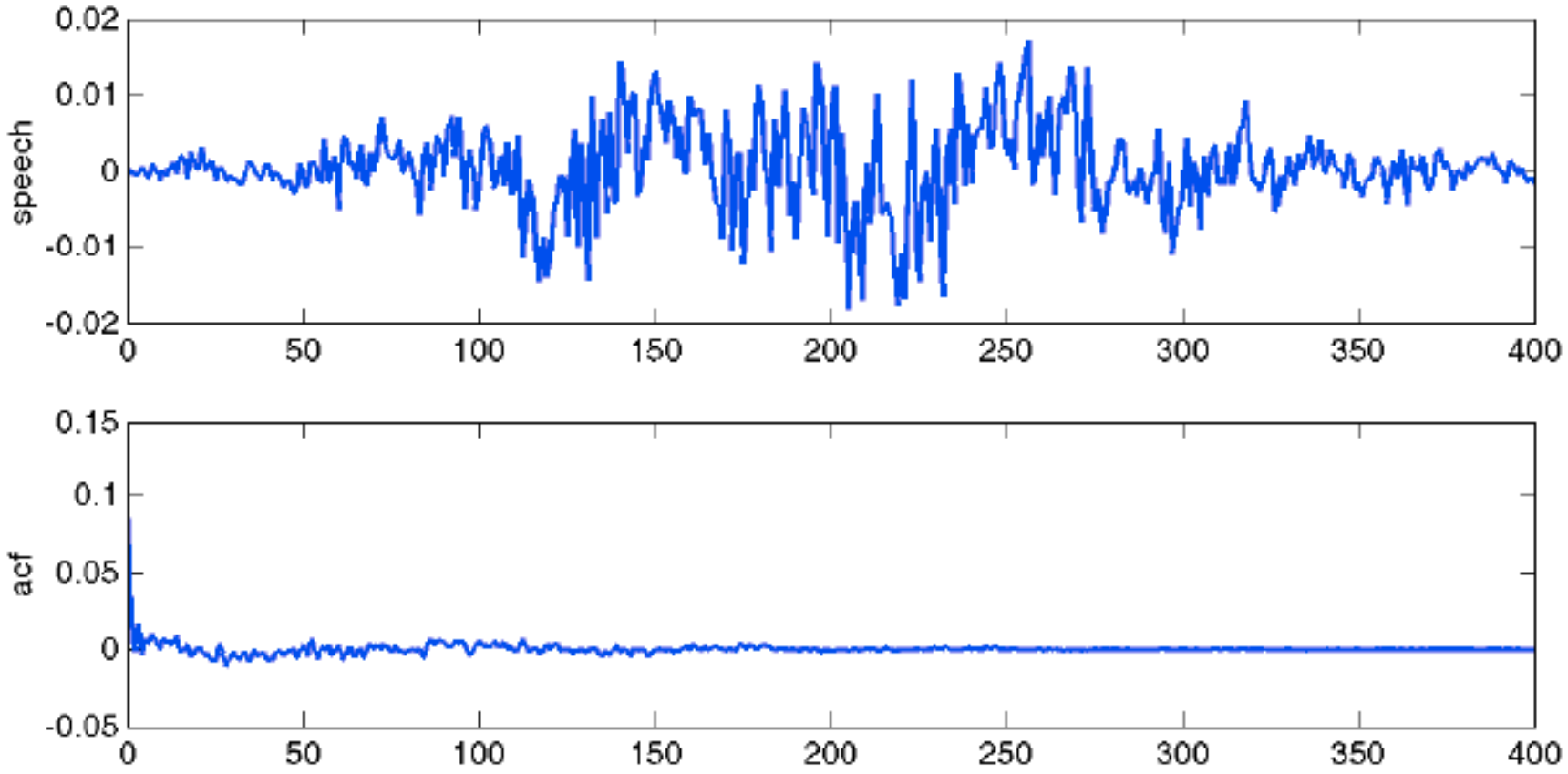
- A measure of similarity
- Autocorrelation function is a good candidate for speech pitch detection algorithms

$$\phi_n[l] = \sum_{n=1}^{N} f[n]f[n+l] \quad , \; 1 = 1, ..., N$$

# Short-Time Auto Correlation Function: Voiced Frame



# Short-Time Auto Correlation Function: Un-voiced Frame



You can find lots of useful MATLAB functions for speech processing here:

http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html