Lecture 06:
[Rabiner Chapter 6] Time-Domain Methods for Speech Processing part 1. short-time energies and zero-crossing rates (ZCR)

DEEE725 음성신호처리실습

Instructor: 장길진

Original slides from Lawrence Rabiner

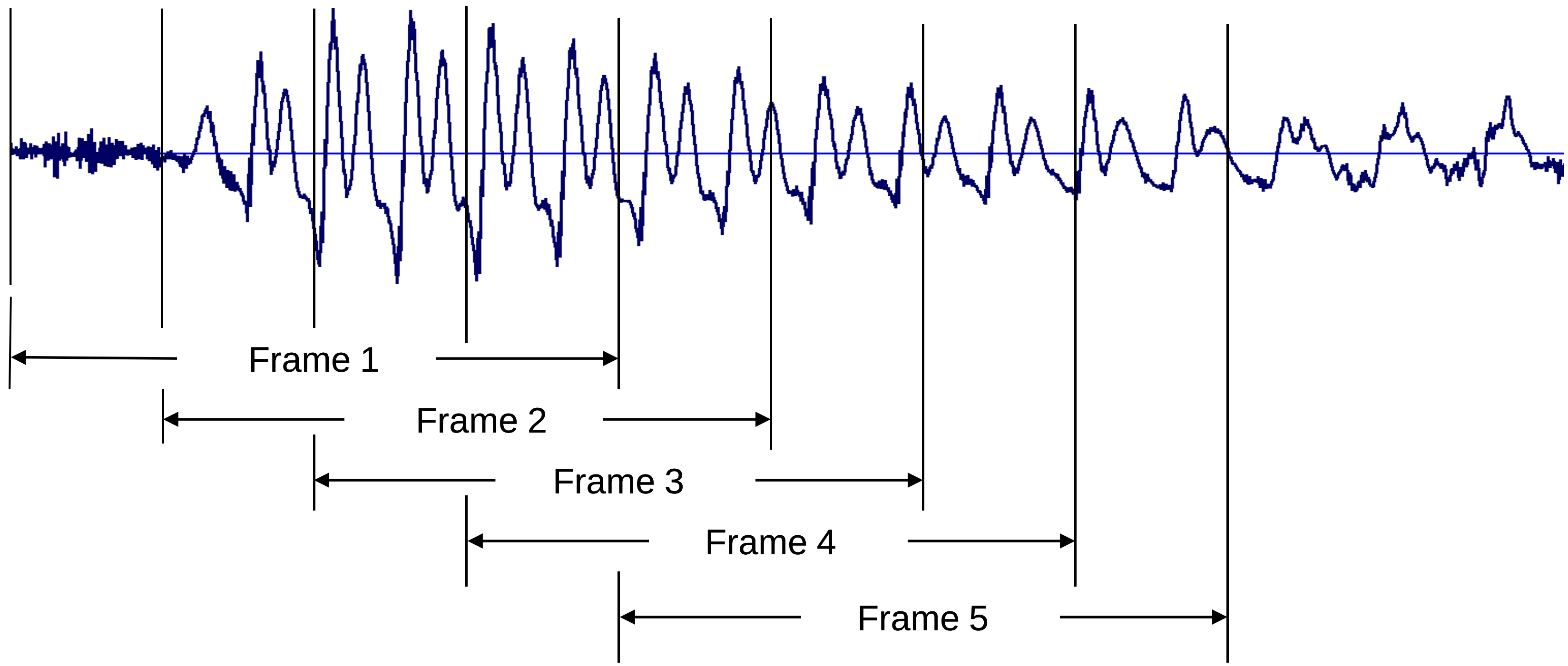
Fundamental Assumptions

- properties of the speech signal change relatively slowly with time (5-10 sounds per second)
  - over very short (5-20 msec) intervals => ***uncertainty*** due to small amount of data, varying pitch, varying amplitude
  - over medium length (20-100 msec) intervals => ***uncertainty*** due to changes in sound quality, transitions between sounds, rapid transients in speech
  - over long (100-500 msec) intervals => ***uncertainty*** due to large amount of sound changes
- there is ***always uncertainty*** in short time measurements and estimates from speech signals

Compromise Solution

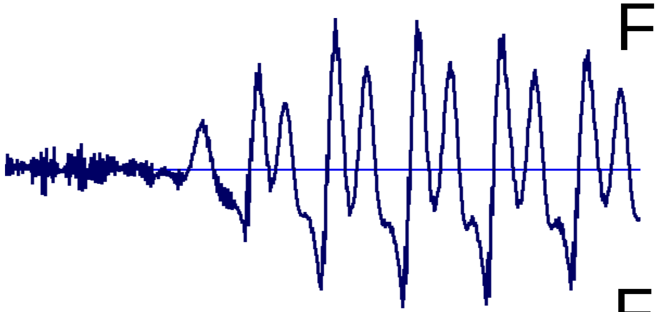
- “short-time” processing methods => short segments of the speech signal are “isolated” and “processed” as if they were short segments from a “sustained” sound with fixed (non-time-varying) properties
  - this short-time processing is periodically repeated for the duration of the waveform
  - these short analysis segments, or “analysis frames” often ***overlap*** one another
  - the results of short-time processing can be a single number (e.g., an estimate of the pitch period within the frame), or a set of numbers (an estimate of the formant frequencies for the analysis frame)
  - the end result of the processing is a new, time-varying sequence that serves as a new representation of the speech signal

Frame-by-Frame Processing in Successive Windows

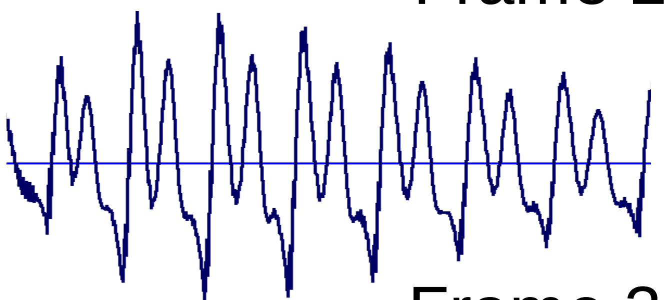


75% frame overlap => frame length=L, frame shift=R=L/4
Frame1={x[0],x[1],…,x[L-1]}
Frame2={x[R],x[R+1],…,x[R+L-1]}
Frame3={x[2R],x[2R+1],…,x[2R+L-1]}
…

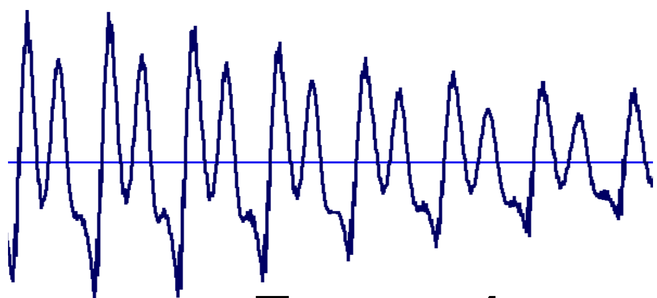
Frame 1: samples $0,1,\ldots,L-1$



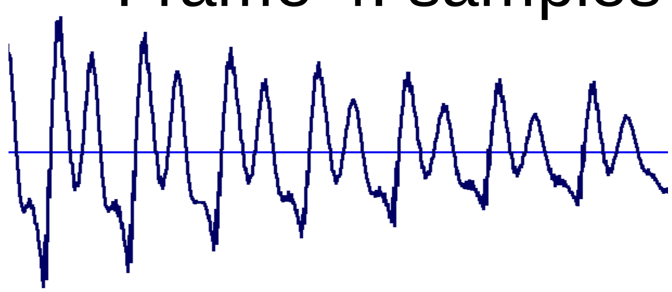
Frame 2: samples $R, R+1,\ldots,R+L-1$



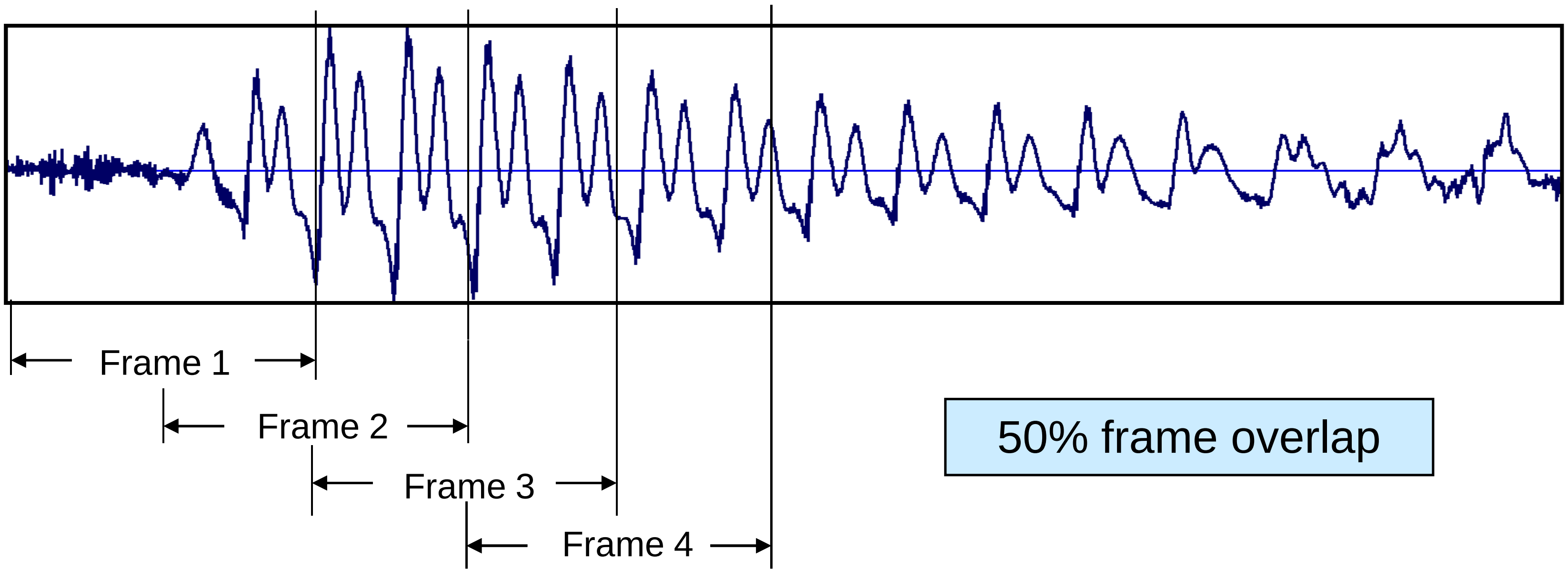
Frame 3: samples $2R, 2R+1,\ldots,2R+L-1$



Frame 4: samples $3R, 3R+1,\ldots,3R+L-1$



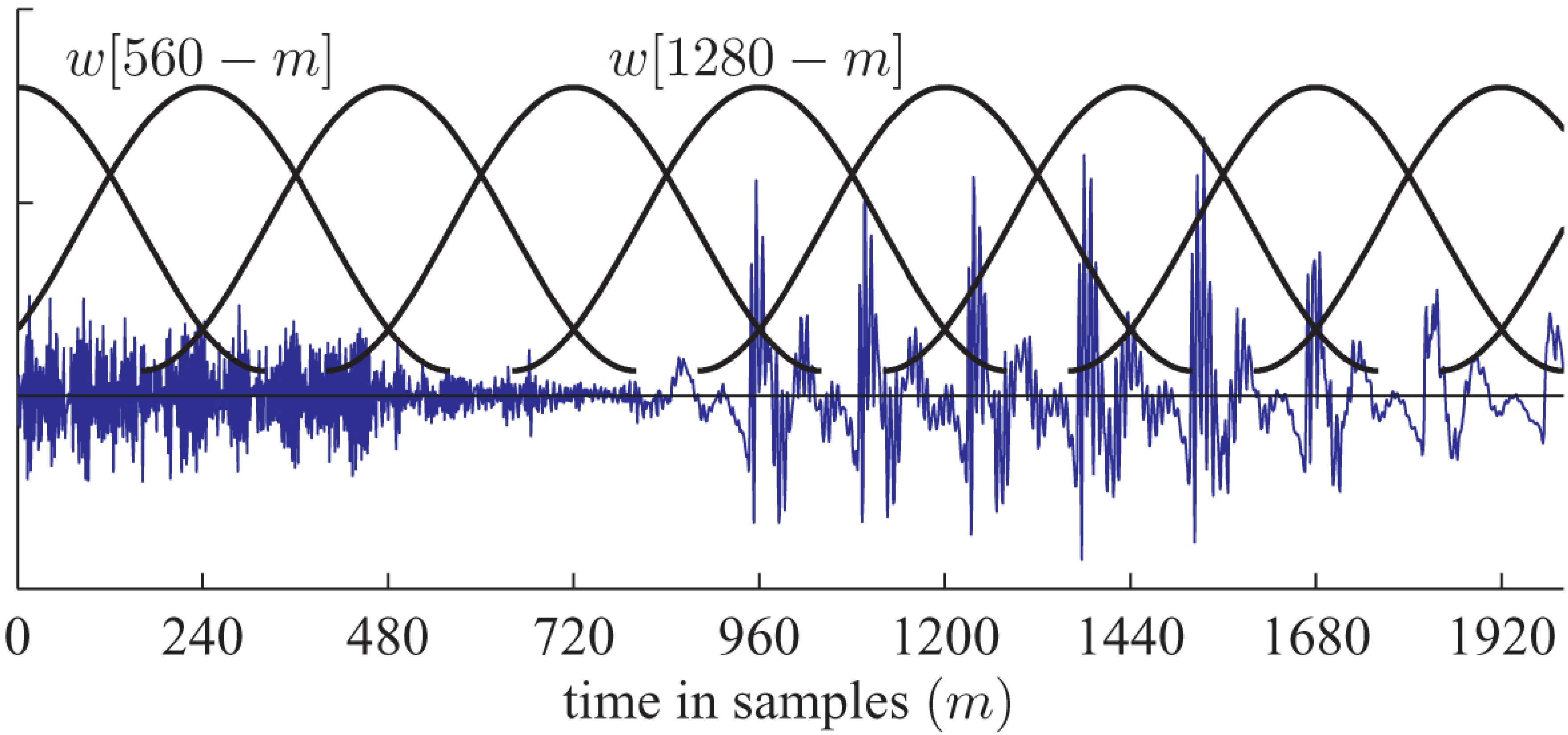
Frame-by-Frame Processing in Successive Windows



50% frame overlap

- Speech is processed frame-by-frame in overlapping intervals until entire region of speech is covered by at least one such frame
- Results of analysis of individual frames used to derive model parameters in some manner
- Representation goes from time sample $x[n], n = \cdots, 0, 1, 2, \cdots$ to parameter vector $\mathbf{f}[m], m = 0, 1, 2, \cdots$ where n is the time index and m is the frame index.

Frames and Windows



$F_S = 16,000$ samples/second

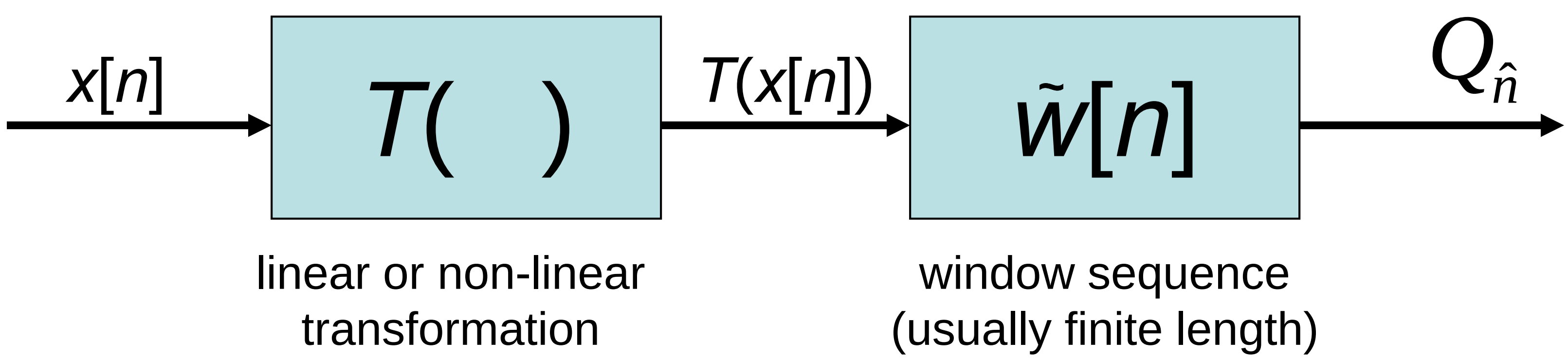
$L = 641$ samples (equivalent to 40 msec frame (window) length)

$R = 240$ samples (equivalent to 15 msec frame (window) shift)

Frame rate of 66.7 frames/second

Generic Short-Time Processing

$$Q_{\hat{n}} = \left(\sum_{m=-\infty}^{\infty} T(x[m])\,\tilde{w}[n-m] \right) \Bigg|_{n=\hat{n}}$$



- $Q_{\hat{n}}$ is a sequence of ***local weighted average values*** of the sequence $T(x[n])$ at time $n = \hat{n}$

Short-Time Energy

$$E = \sum_{m=-\infty}^{\infty} x^2[m]$$

-- this is the long term definition of signal energy

-- there is little or no utility of this definition for time-varying signals

$$E_{\hat{n}} = \sum_{m=\hat{n}-N+1}^{\hat{n}} x^2[m] = x^2[\hat{n}-N+1] + \ldots + x^2[\hat{n}]$$

-- short-time energy in vicinity of time $\hat{n}$

$$T(x) = x^2$$

$$\begin{aligned} \tilde{w}[n] &= 1 \qquad 0 \le n \le N-1 \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Computation of Short-Time Energy

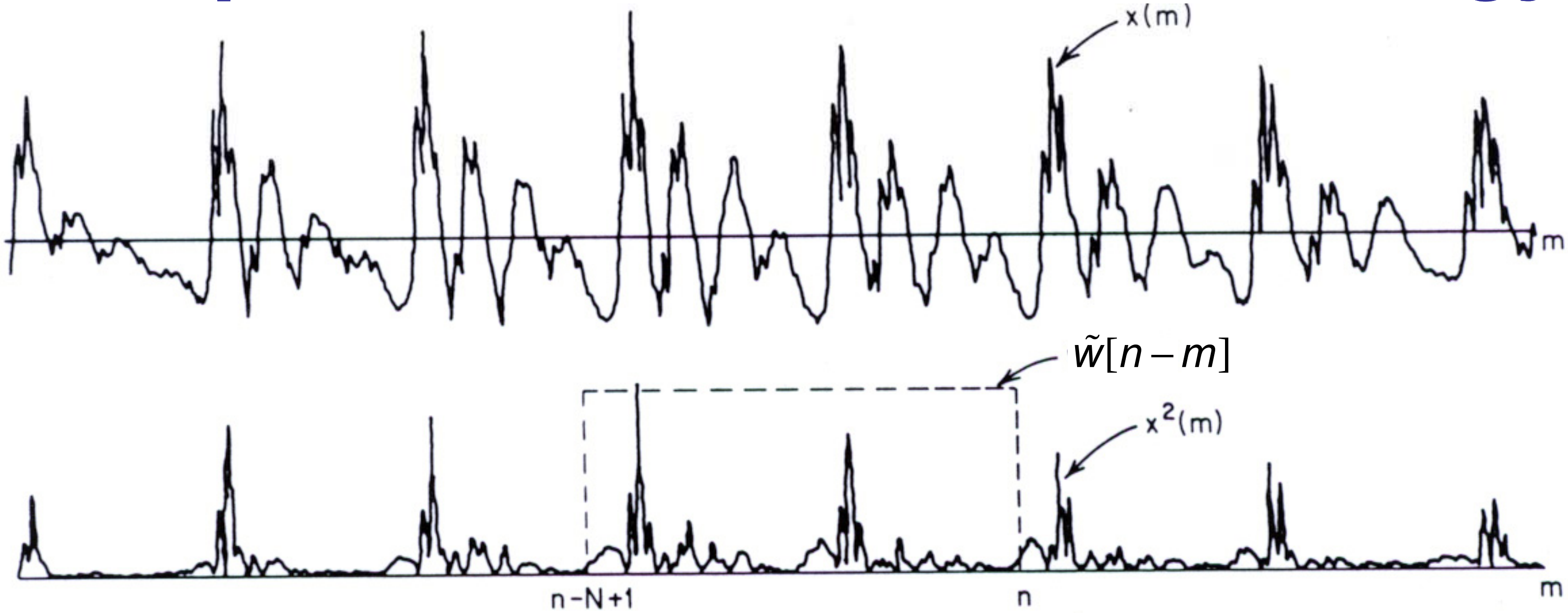


Fig. 4.2 Illustration of the computation of short-time energy.

- **window jumps/slides across sequence of squared values**, selecting interval for processing
- what happens to $E_{\hat{n}}$ as sequence jumps by 2,4,8,...,L samples ($E_{\hat{n}}$ is a lowpass function—so it can be decimated without lost of information; why is $E_{\hat{n}}$ lowpass?)
- effects of decimation depend on L; if L is small, then $E_{\hat{n}}$ is a lot more variable than if L is large (window bandwidth changes with L!)

Short-Time Energy

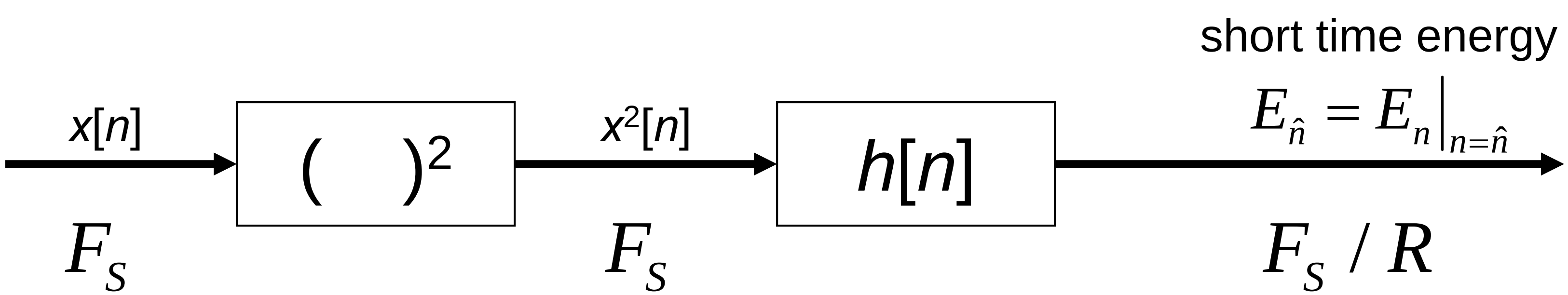
- serves to **<u>differentiate voiced and unvoiced sounds</u>** in speech from silence (background signal)
- natural definition of energy of weighted signal is:

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} \left[x[m]\tilde{w}[\hat{n}-m] \right]^2 \text{ (sum or squares of portion of signal)}$$

-- concentrates measurement at sample $\hat{n}$, using weighting $\tilde{w}[\hat{n} - m]$

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} x^2[m]\, \tilde{w}^2[\hat{n}-m] = \sum_{m=-\infty}^{\infty} x^2[m]\, h[\hat{n}-m]$$

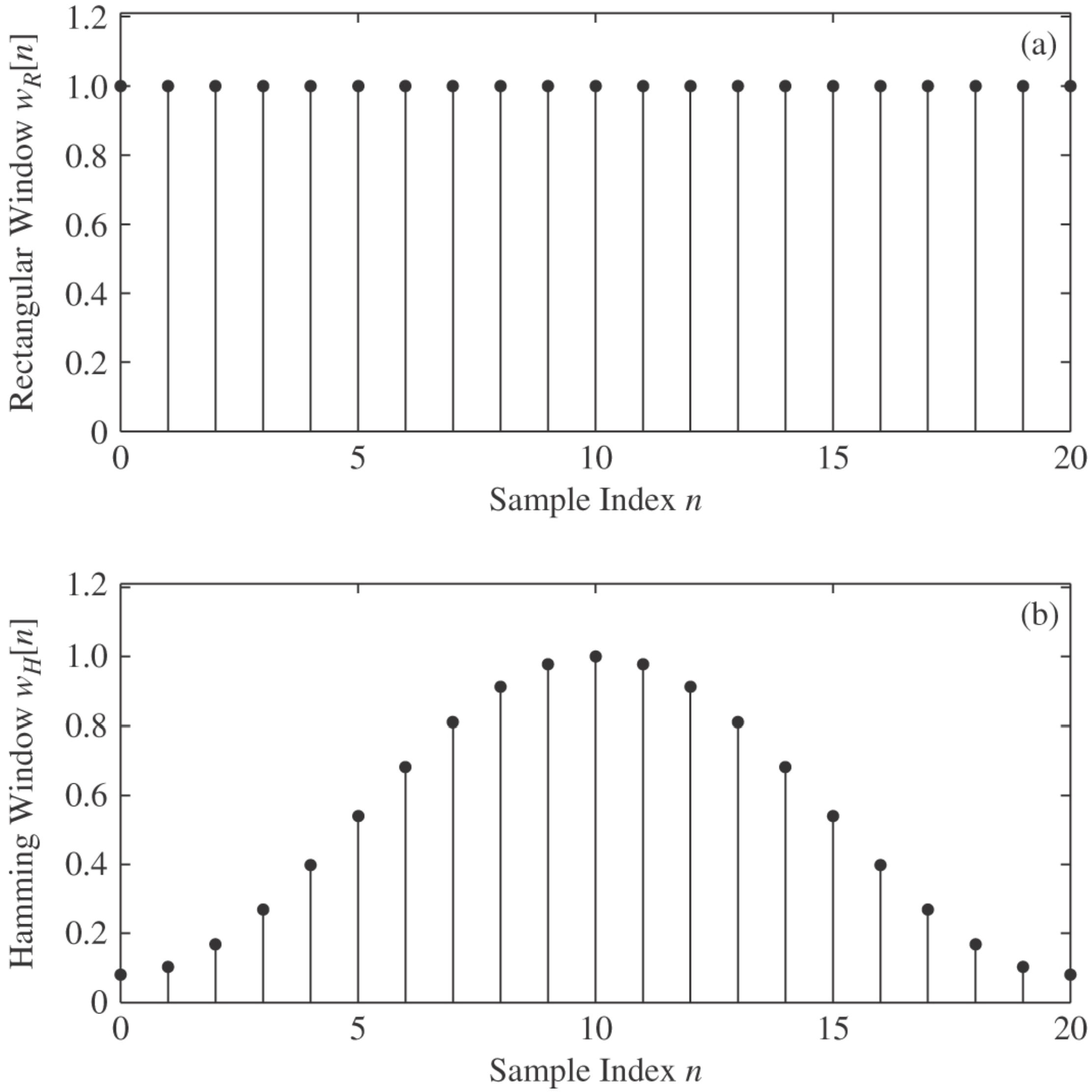
$$h[n] = \tilde{w}^2[n]$$



Windows

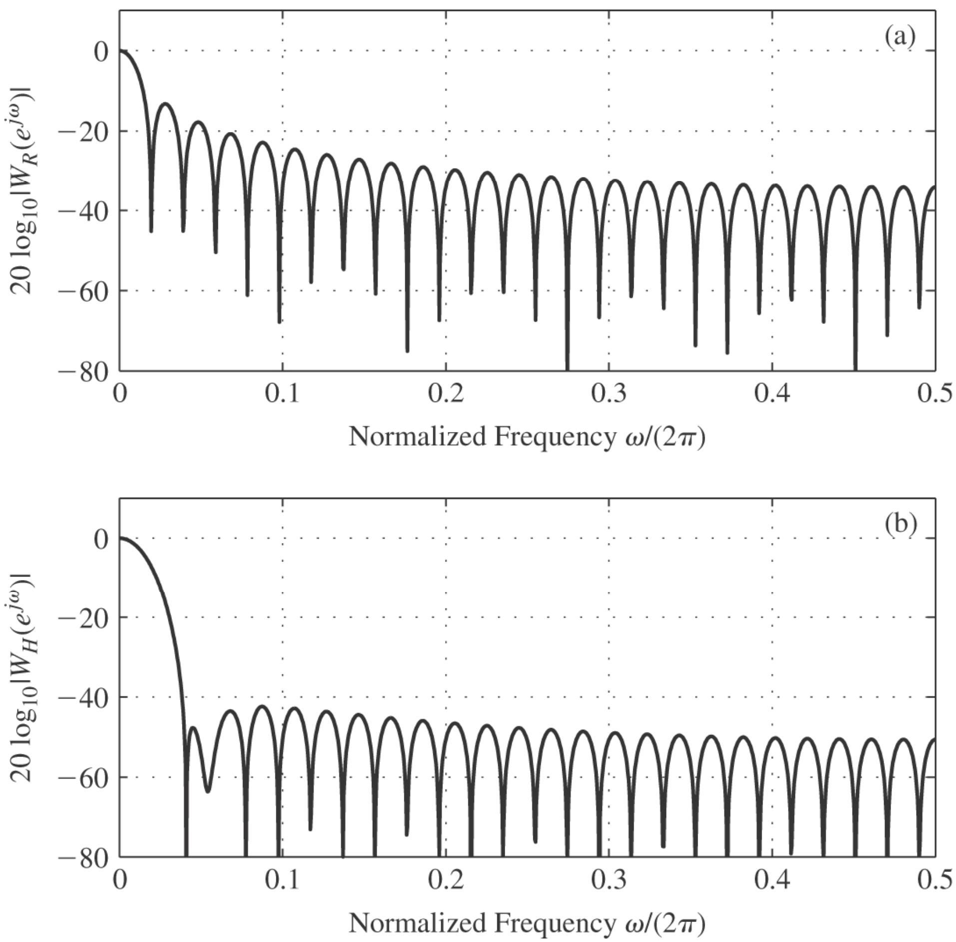
- consider two windows, $\tilde{w}[n]$
  - rectangular window:
    - $h[n]=1,\ 0 \le n \le L-1$ and 0 otherwise
  - Hamming window (raised cosine window):
    - $h[n]=0.54-0.46\cos(2\pi n/(L-1)),\ 0 \le n \le L-1$ and 0 otherwise
  - rectangular window gives ***equal weight*** to all L samples in the window ($n,\ldots,n-L+1$)
  - Hamming window gives ***most weight*** to middle samples and ***tapers off*** strongly at the beginning and the end of the window

Rectangular and Hamming Windows



$L = 21$ samples

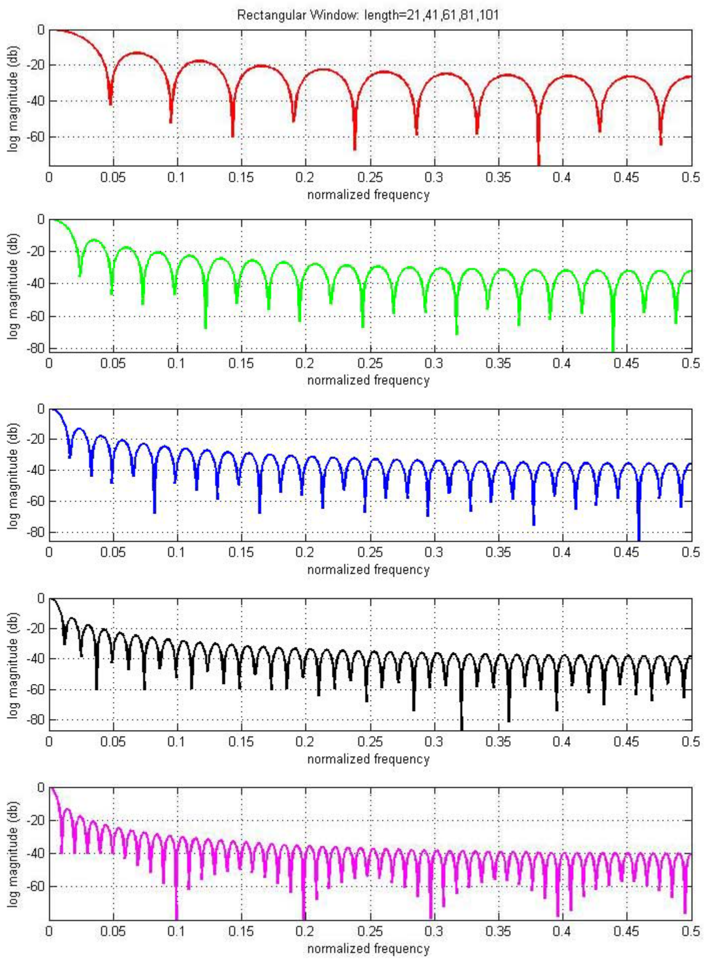
RW and HW Frequency Responses

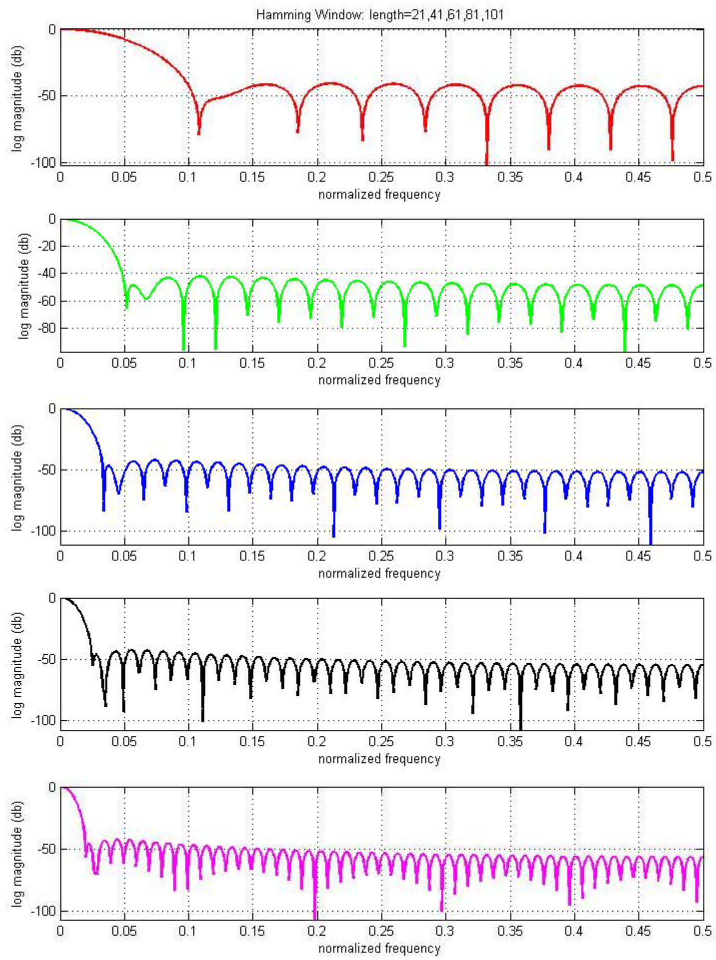


- log magnitude response of RW and HW
- ***bandwidth*** of HW is approximately twice the bandwidth of RW
- ***attenuation*** of more than 40 dB for HW outside passband, versus 14 dB for RW
- stopband attenuation is essentially **independent** of L, the window duration => increasing L simply decreases window bandwidth
- L needs to be larger than a pitch period (or severe fluctuations will occur in E_n), but smaller than a sound duration (or E_n will not adequately reflect the changes in the speech signal)

There is no perfect value of L, since a pitch period can be as short as 20 samples (500 Hz at a 10 kHz sampling rate) for a high pitch child or female, and up to 250 samples (40 Hz pitch at a 10 kHz sampling rate) for a low pitch male; a compromise value of L on the order of 100-200 samples for a 10 kHz sampling rate is often used in practice

Window Frequency Responses

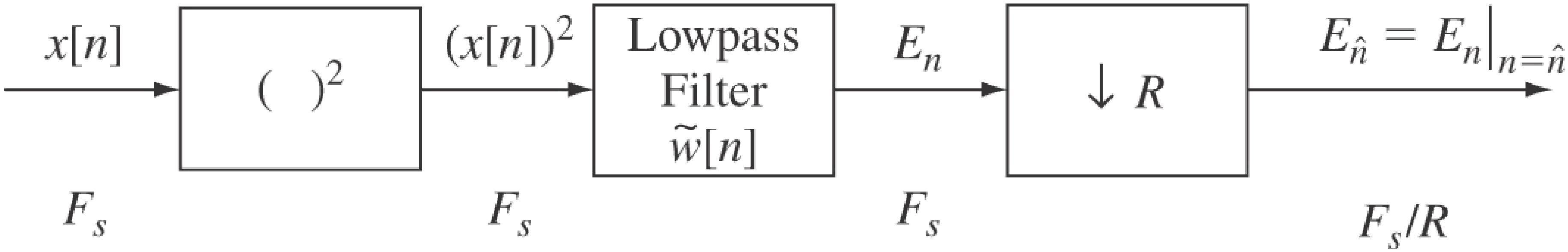




Rectangular Windows, L=21,41,61,81,101

Hamming Windows, L=21,41,61,81,101

Short-Time Energy



- Short-time energy computation:

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])^2$$

$$= \sum_{m=-\infty}^{\infty} (x[m])^2 \tilde{w}[\hat{n}-m]$$

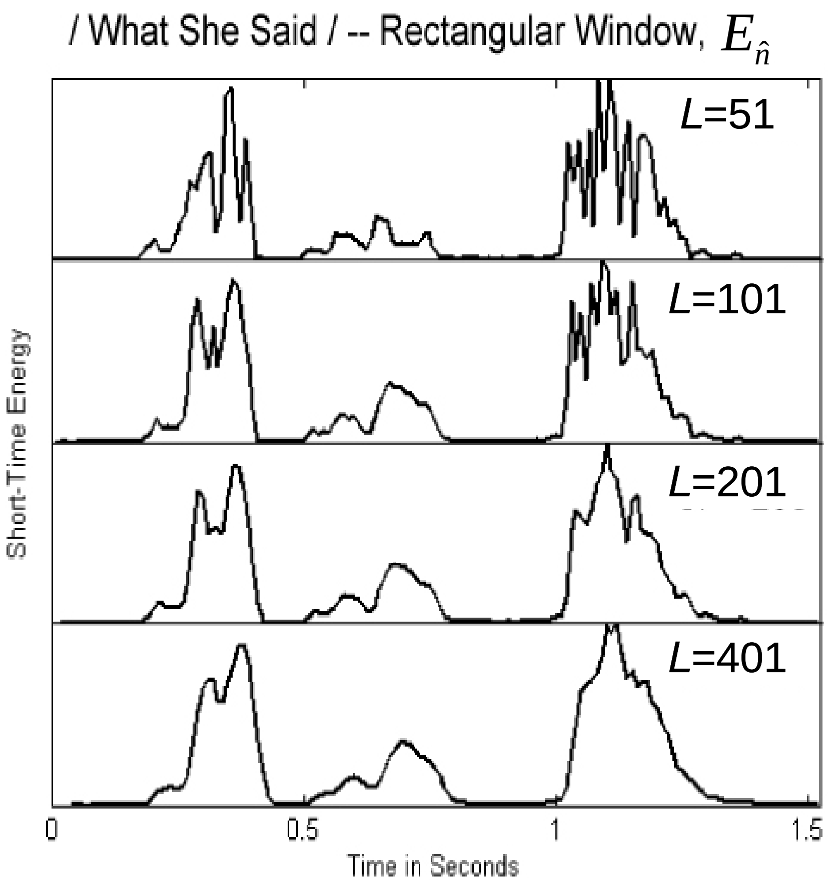
- For L-point rectangular window,

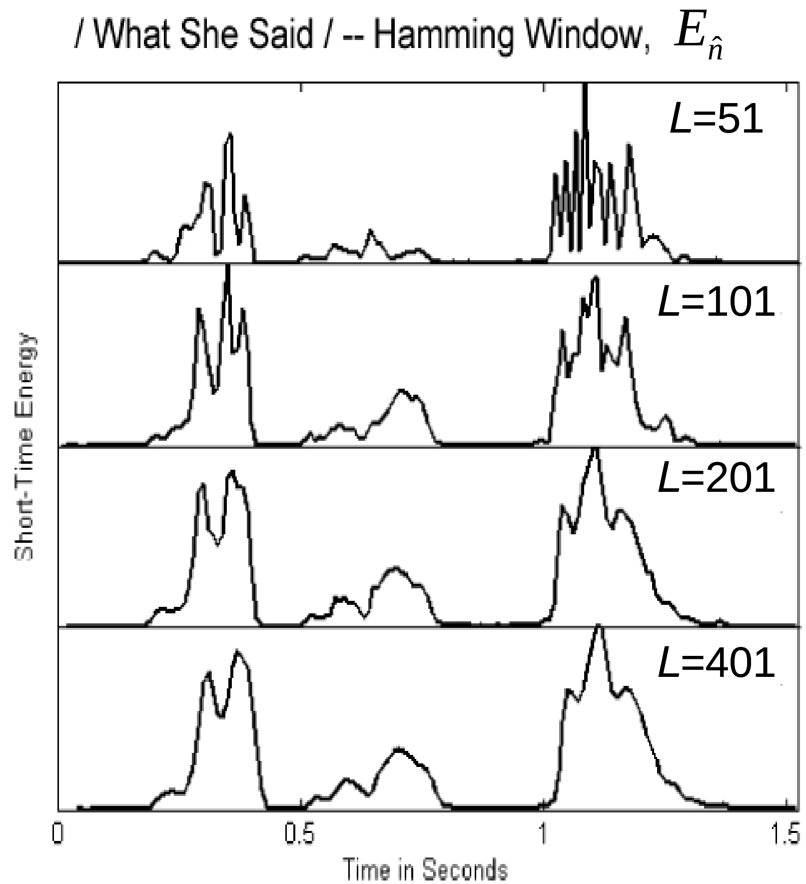
$$\tilde{w}[m] = 1, \quad m = 0,1,\ldots,L-1$$

- giving

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} (x[m])^2$$

Short-Time Energy using RW/HW





- as *L* increases, the plots tend to converge (however you are smoothing sound energies)
- short-time energy provides the basis for distinguishing voiced from unvoiced speech regions, and for medium-to-high SNR recordings, can even be used to find regions of silence/background signal

Short-Time Energy for AGC

Can use an IIR filter to define short-time energy, e.g.,

- time-dependent energy definition

$$\sigma^2[n] = \sum_{m=-\infty}^{\infty} x^2[m]h[n-m] / \sum_{m=0}^{\infty} h[m]$$

- consider impulse response of filter of form

$$\begin{aligned} h[n] = \alpha^{n-1}u[n-1] &= \alpha^{n-1} \quad n \geq 1 \\ &= 0 \qquad\quad n < 1 \end{aligned}$$

$$\sigma^2[n] = \sum_{m=-\infty}^{\infty} (1-\alpha)\, x^2[m]\alpha^{n-m-1}u[n-m-1]$$

Recursive Short-Time Energy

- $u[n-m-1]$ implies the condition $n-m-1 \geq 0$

  or $m \leq n-1$ giving

$$\sigma^2[n] = \sum_{m=-\infty}^{n-1} (1-\alpha)\, x^2[m]\, \alpha^{n-m-1} = (1-\alpha)(x^2[n-1] + \alpha x^2[n-2] + \ldots)$$

- for the index $n-1$ we have

$$\sigma^2[n-1] = \sum_{m=-\infty}^{n-2} (1-\alpha)\, x^2[m]\, \alpha^{n-m-2} = (1-\alpha)(x^2[n-2] + \alpha x^2[n-3] + \ldots)$$

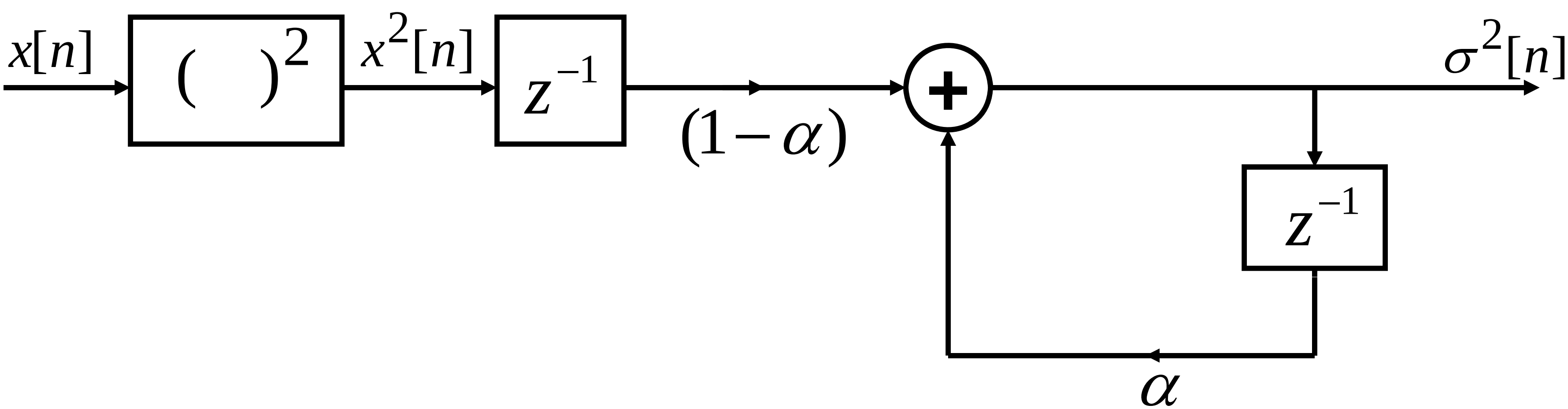
- thus giving the relationship

$$\boxed{\sigma^2[n] = \alpha \cdot \sigma^2[n-1] + x^2[n-1](1-\alpha)}$$

- and defines an Automatic Gain Control (AGC) of the form

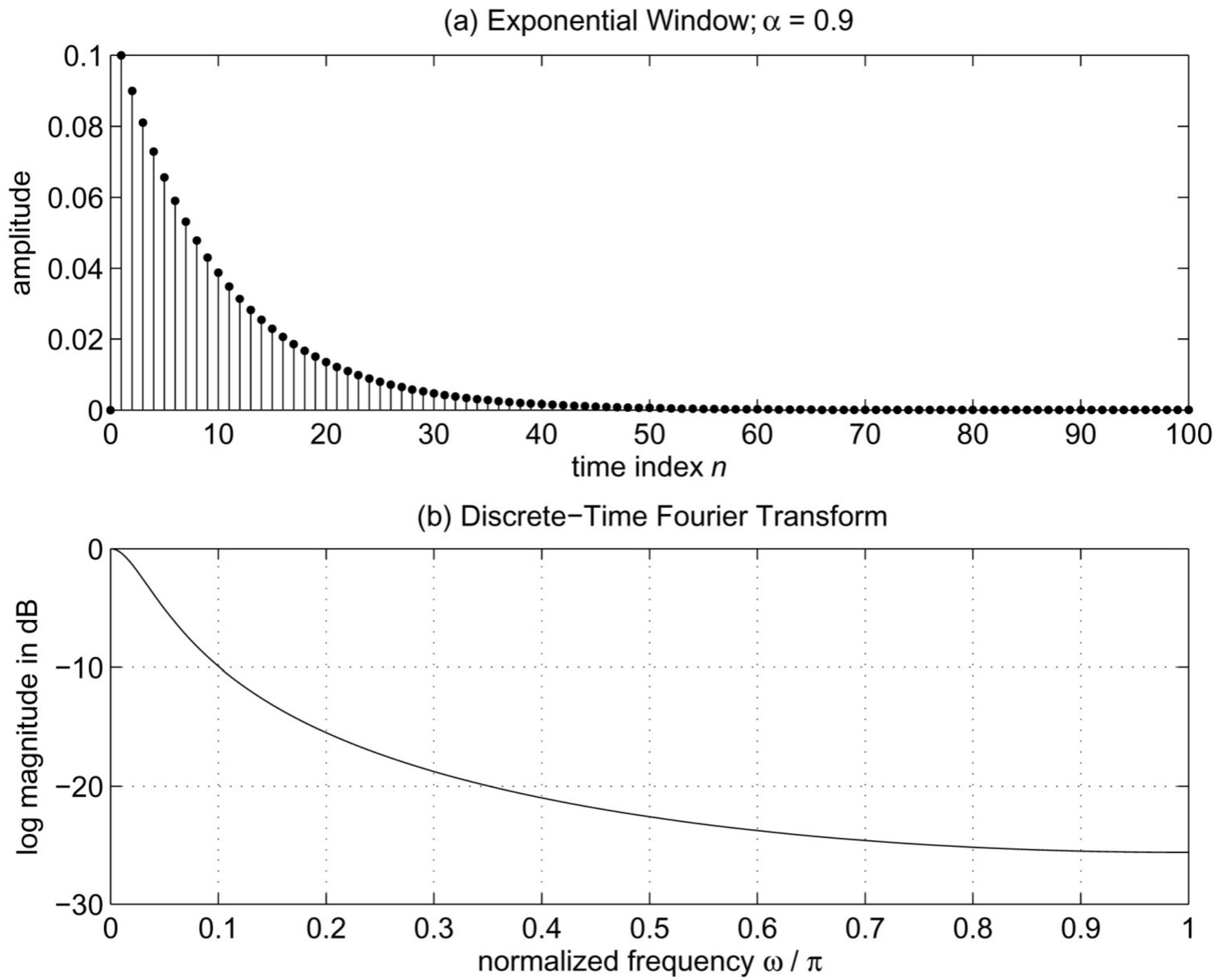
$$G[n] = \frac{G_0}{\sigma[n]}$$

Recursive Short-Time Energy



$$\boxed{\sigma^2[n] = \alpha \cdot \sigma^2[n-1] + x^2[n-1](1-\alpha)}$$

Recursive Short-Time Energy



Use of Short-Time Energy for AGC

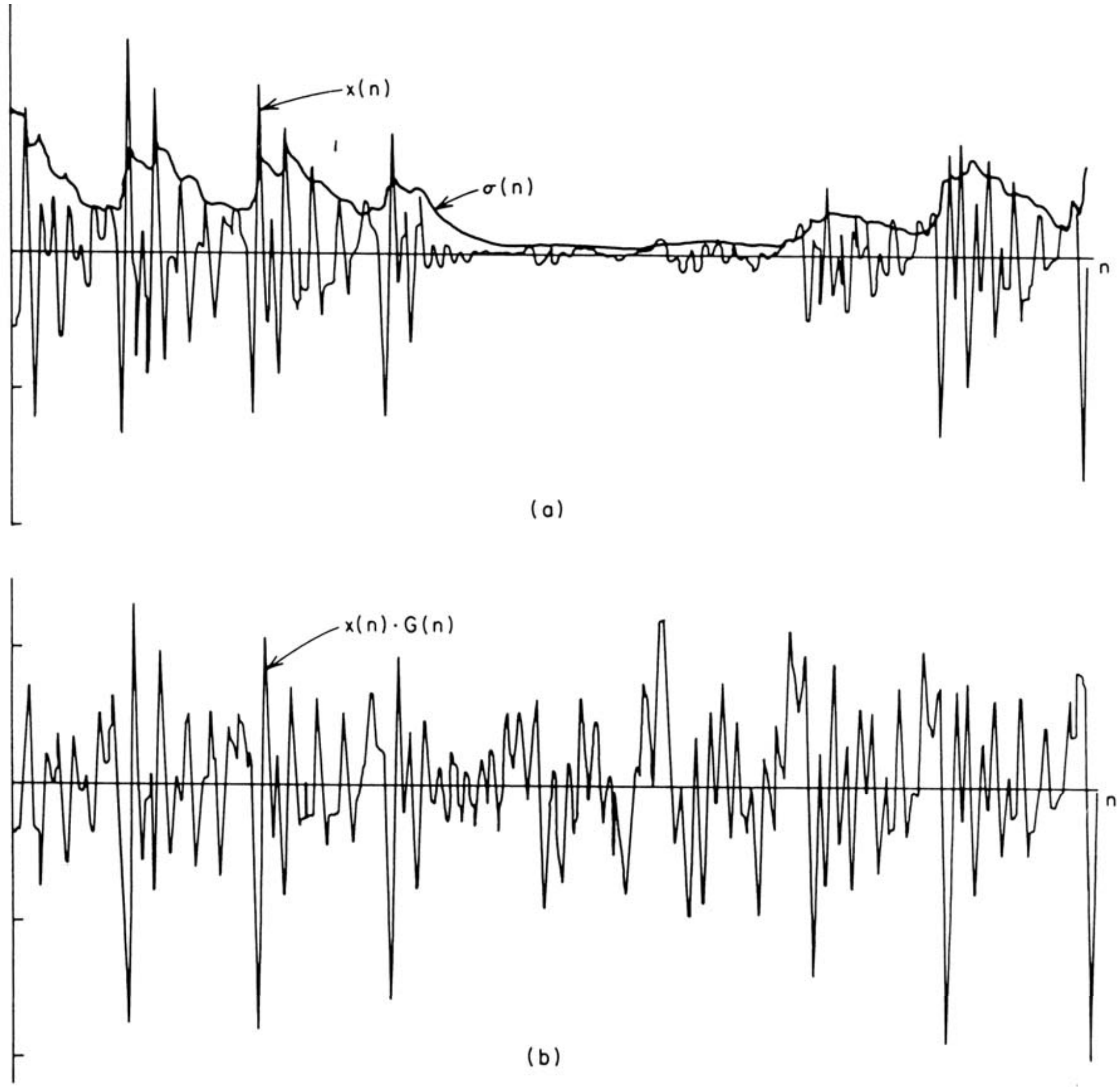
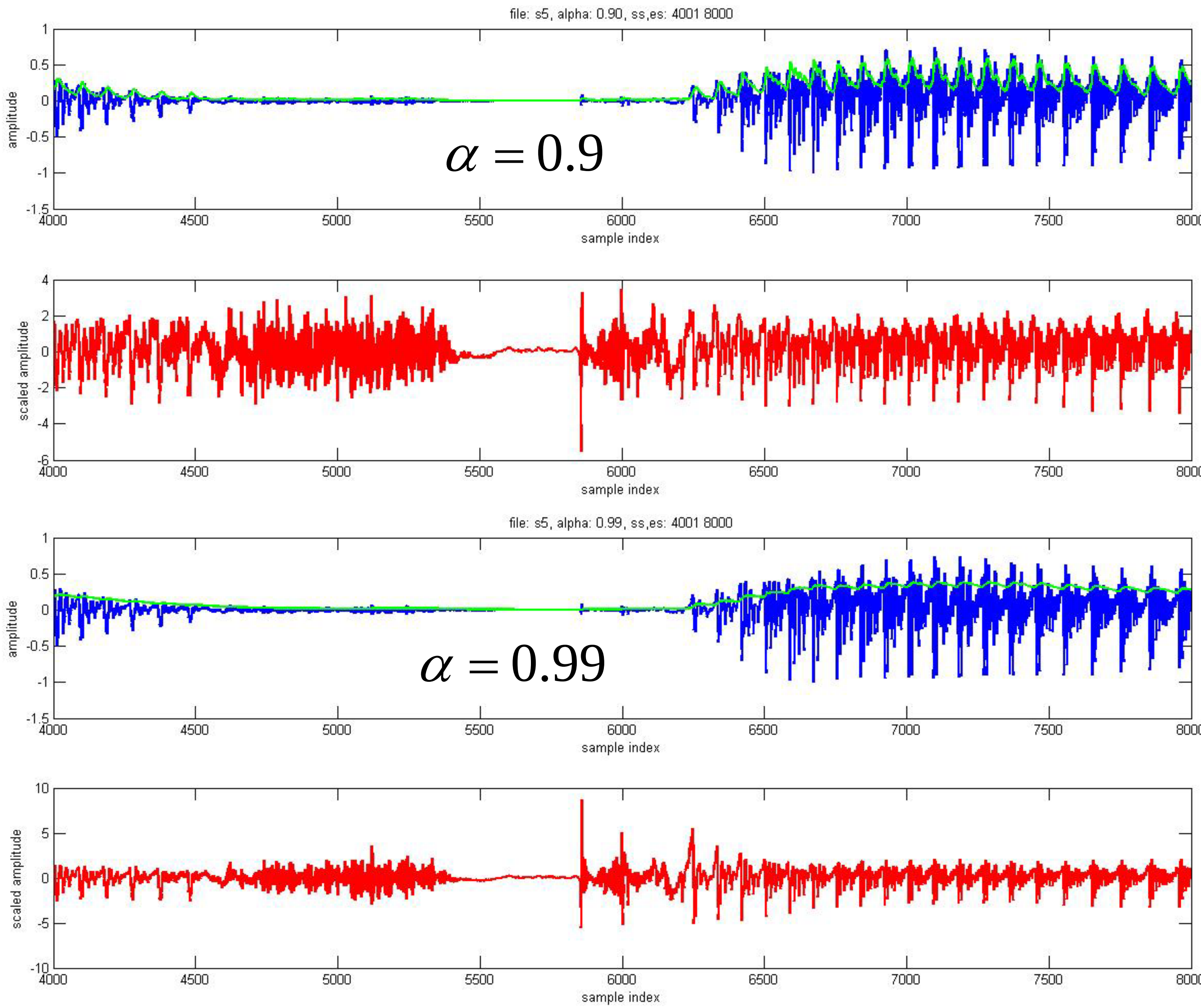


Fig. 5.26 Variance estimate using Eq. (5.56); (a) $x(n)$ and $\sigma(n)$ for $\alpha = 0.9$; (b) $x(n)\ G(n)$.

Use of Short-Time Energy for AGC



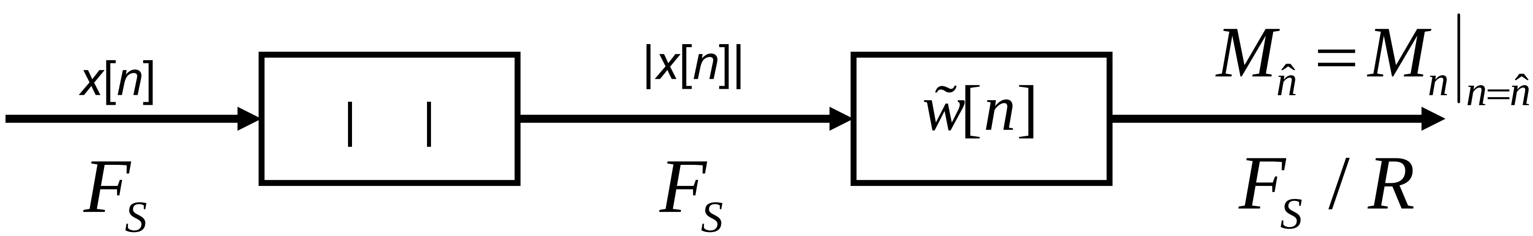
Short-Time Magnitude

Power

- short-time energy is very sensitive to large signal levels due to $x^2[n]$ terms
  - consider a new definition of 'pseudo-energy' based on average signal magnitude (rather than energy)

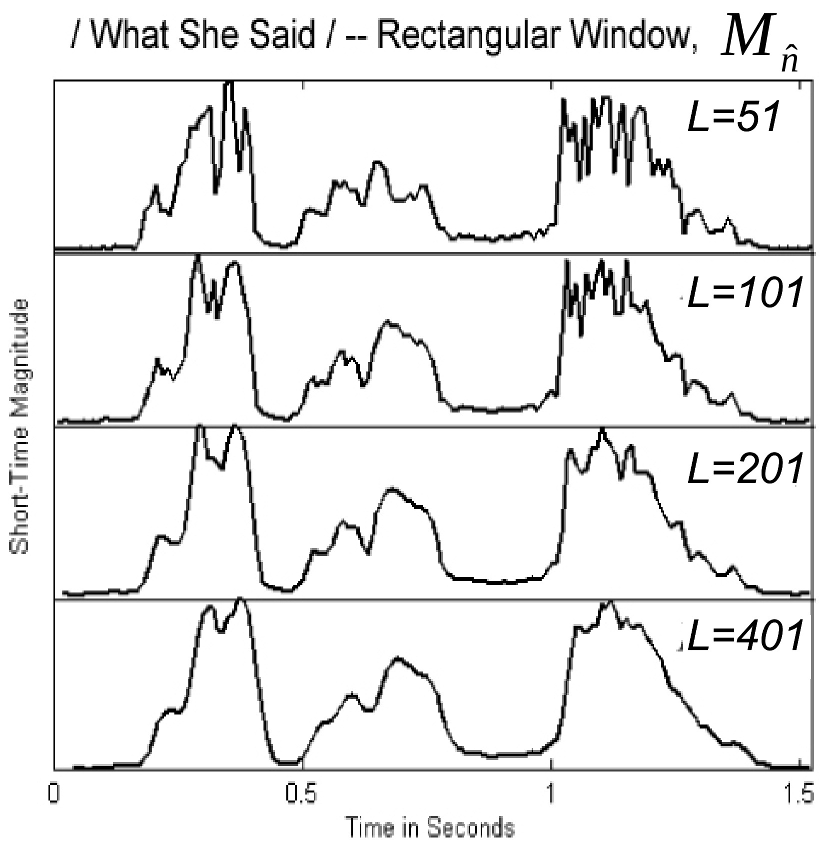
$$M_{\hat{n}} = \sum_{m=-\infty}^{\infty} |x[m]|\, \tilde{w}[\hat{n}-m]$$

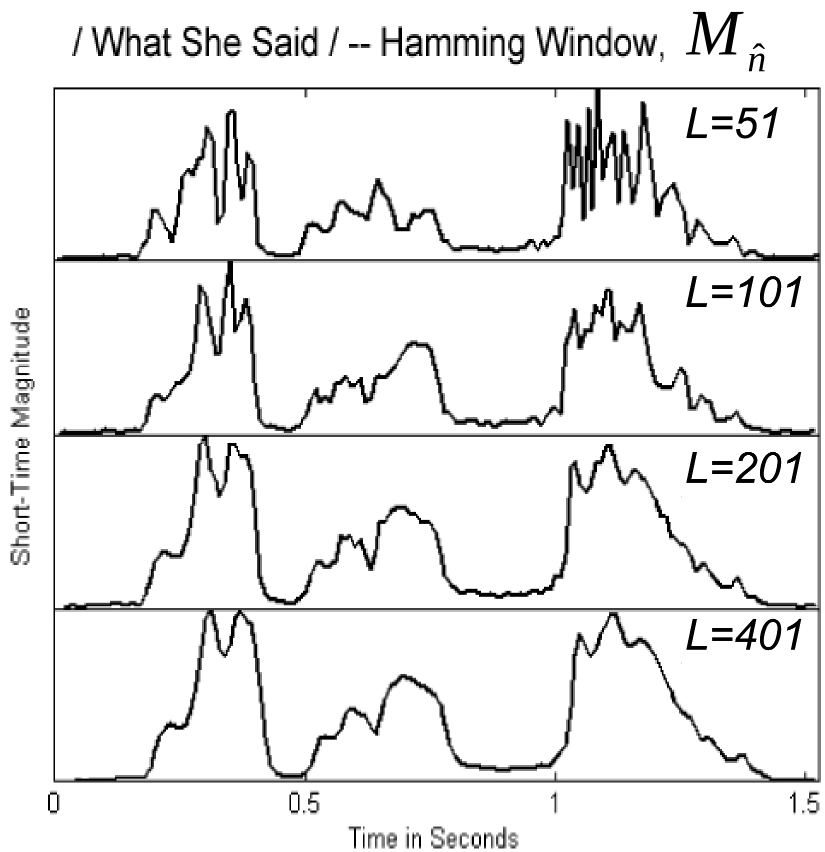
  - weighted sum of magnitudes, rather than weighted sum of squares



• computation avoids multiplications of signal with itself (the squared term)

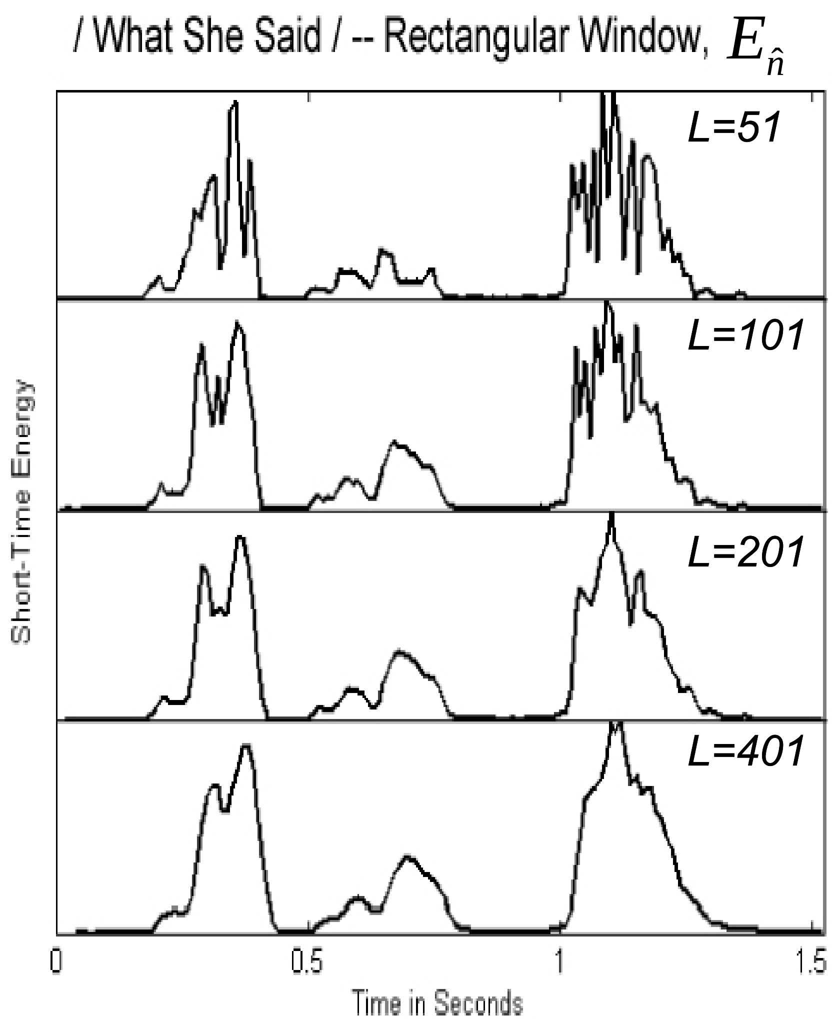
Short-Time Magnitudes

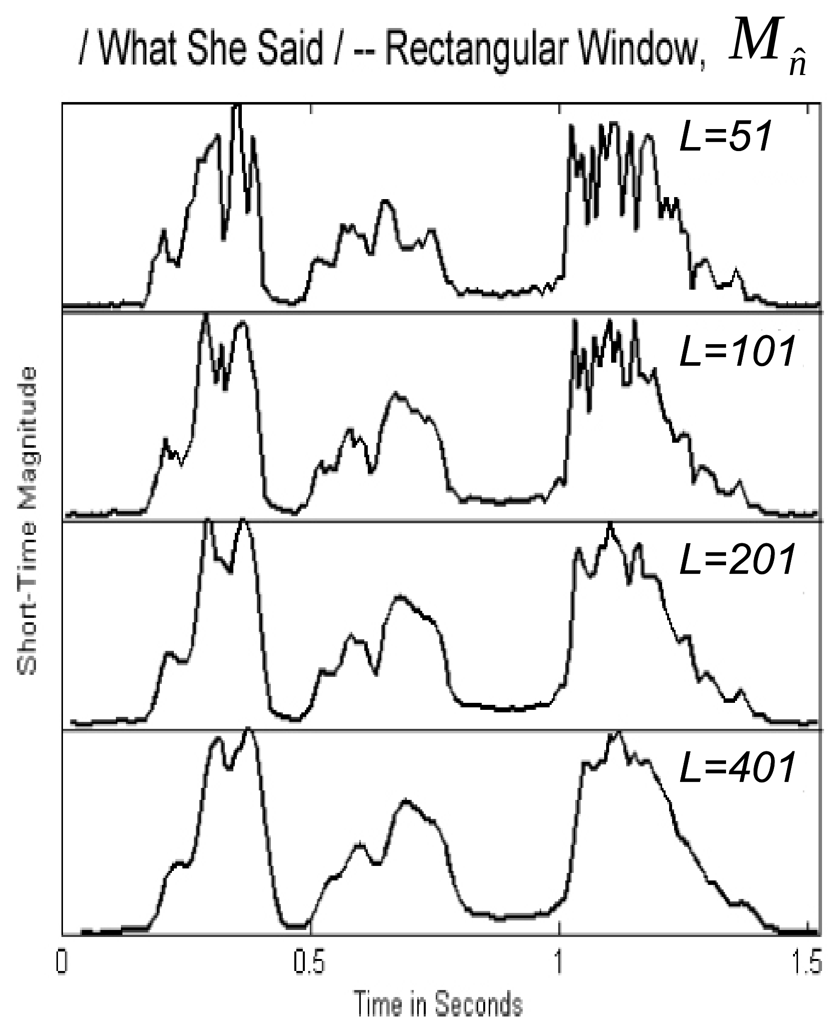




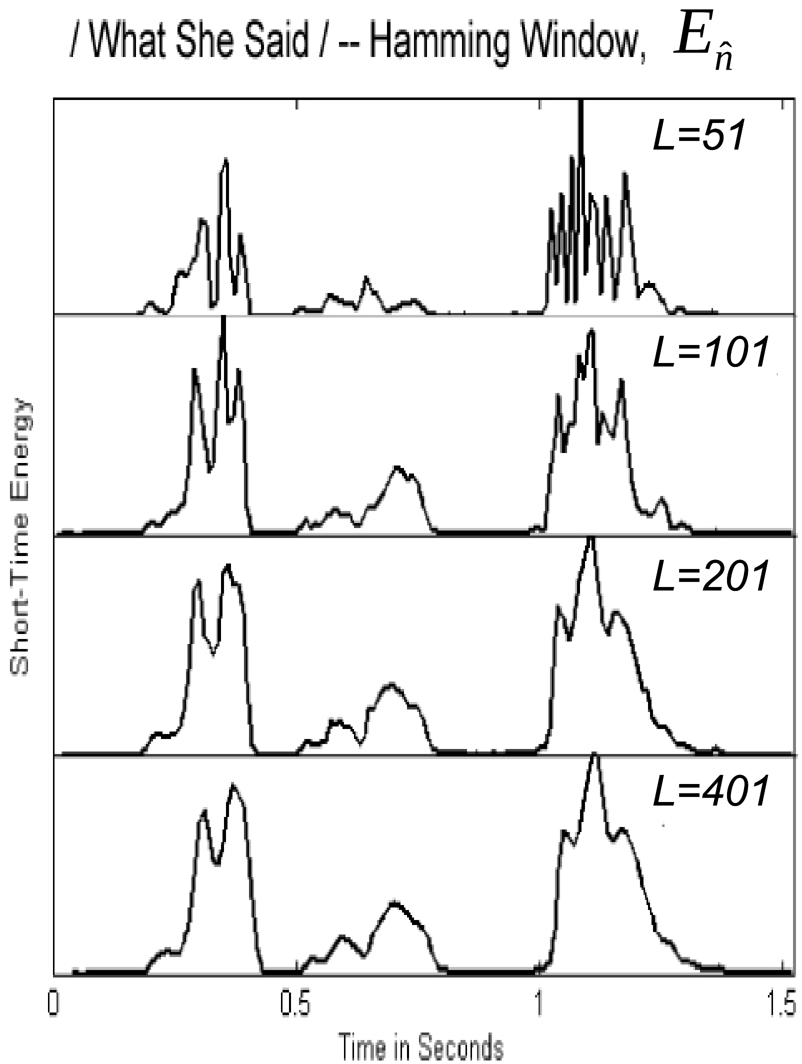
- differences between E_n and M_n noticeable in unvoiced regions
- dynamic range of M_n ~ square root (dynamic range of E_n) => level differences between voiced and unvoiced segments are smaller
- E_n and M_n can be sampled at a rate of 100/sec for window durations of 20 msec or so => efficient representation of signal energy/magnitude

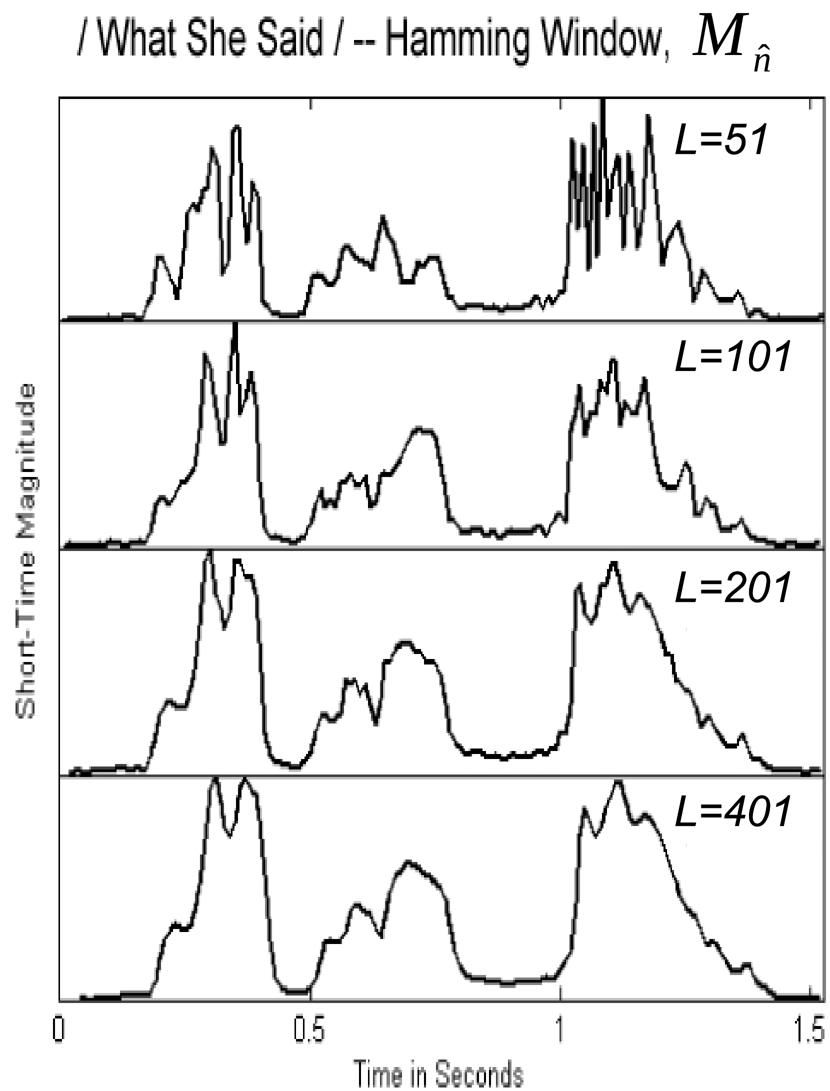
Short Time Energy and Magnitude—Rectangular Window





Short Time Energy and Magnitude—Hamming Window





Other Lowpass Windows

- can replace RW or HW with any lowpass filer
- window should be positive since this guarantees E_n and M_n will be positive
- FIR windows are efficient computationally since they can slide by L samples for efficiency with no loss of information (what should L be?)
- can even use an infinite duration window if its z-transform is a rational function, i.e.,

$$h[n] = a^n, \quad n \geq 0, \quad 0 < a < 1$$

$$= 0 \qquad n < 0$$

$$H(z) = \frac{1}{1 - az^{-1}} \qquad |z| > |a|$$

Other Lowpass Windows

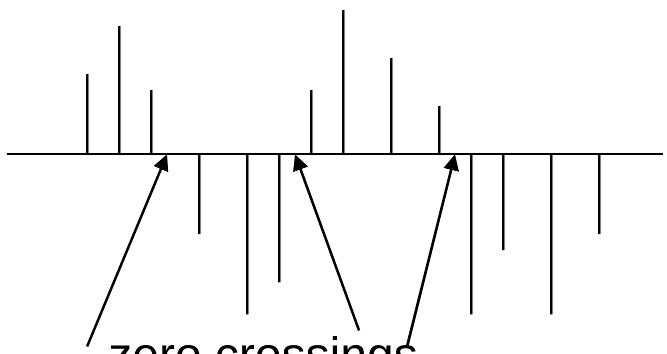
- this simple lowpass filter can be used to implement E_n and M_n recursively as:

$$E_n = aE_{n-1} + (1-a)x^2[n] - \text{short-time energy}$$

$$M_n = aM_{n-1} + (1-a)|x[n]| - \text{short-time magnitude}$$

- need to compute E_n or M_n every sample and then down-sample to 100/sec rate
- recursive computation has a non-linear phase, so delay cannot be compensated exactly

Short-Time Average ZC Rate



zero crossing => successive samples have different algebraic signs

• zero crossing rate is a simple measure of the 'frequency content' of a signal—especially true for narrowband signals (e.g., sinusoids)

• sinusoid at frequency F_0 with sampling rate F_S has F_S/F_0 samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of

$z_1=(2)$ crossings/cycle x (F_0 / F_S) cycles/sample

$z_1=2F_0 / F_S$ crossings/sample (i.e., z_1 proportional to F_0)

$z_M=M\ (2F_0 /F_S)$ crossings/(M samples)

Sinusoid Zero Crossing Rates

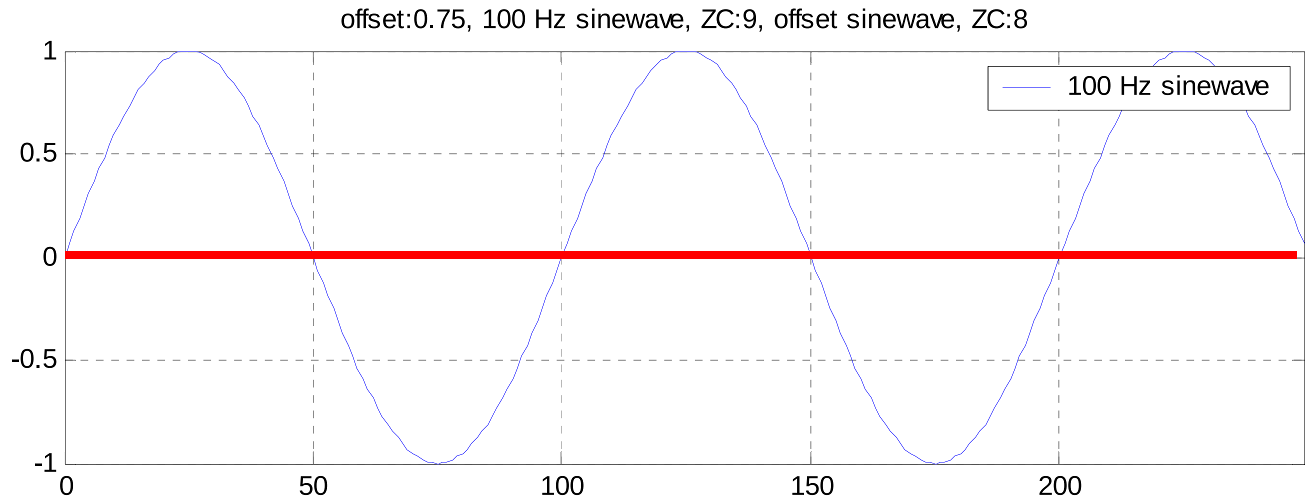
Assume the sampling rate is $F_S = 10{,}000$ Hz

1. $F_0 = 100$ Hz sinusoid has $F_S / F_0 = 10{,}000/100 = 100$ samples/cycle;
   or $z_1 = 2/100$ crossings/sample, or $z_{100} = 2/100 * 100 =$
   2 crossings/10 msec interval

2. $F_0 = 1000$ Hz sinusoid has $F_S / F_0 = 10{,}000/1000 = 10$ samples/cycle;
   or $z_1 = 2/10$ crossings/sample, or $z_{100} = 2/10 * 100 =$
   20 crossings/10 msec interval

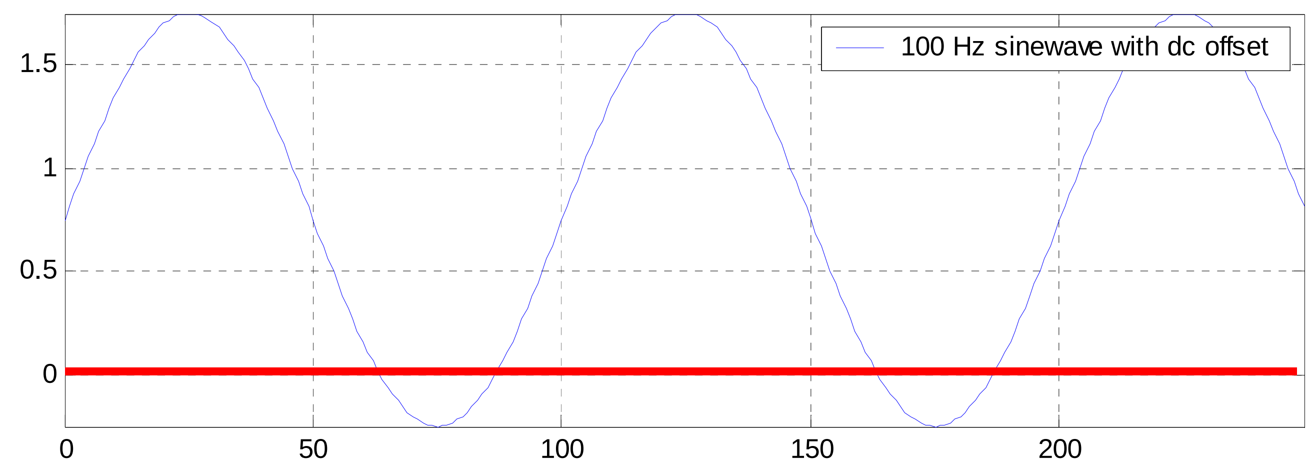
3. $F_0 = 5000$ Hz sinusoid has $F_S / F_0 = 10{,}000/5000 = 2$ samples/cycle;
   or $z_1 = 2/2$ crossings/sample, or $z_{100} = 2/2 * 100 =$
   100 crossings/10 msec interval

Zero Crossing for Sinusoids



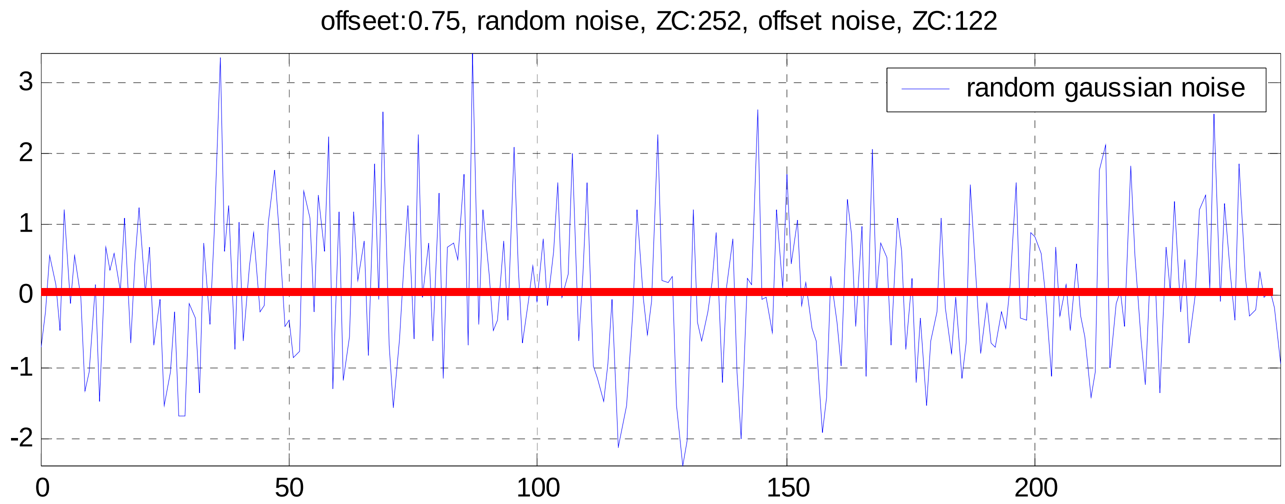
ZC=9

Offset=0.75



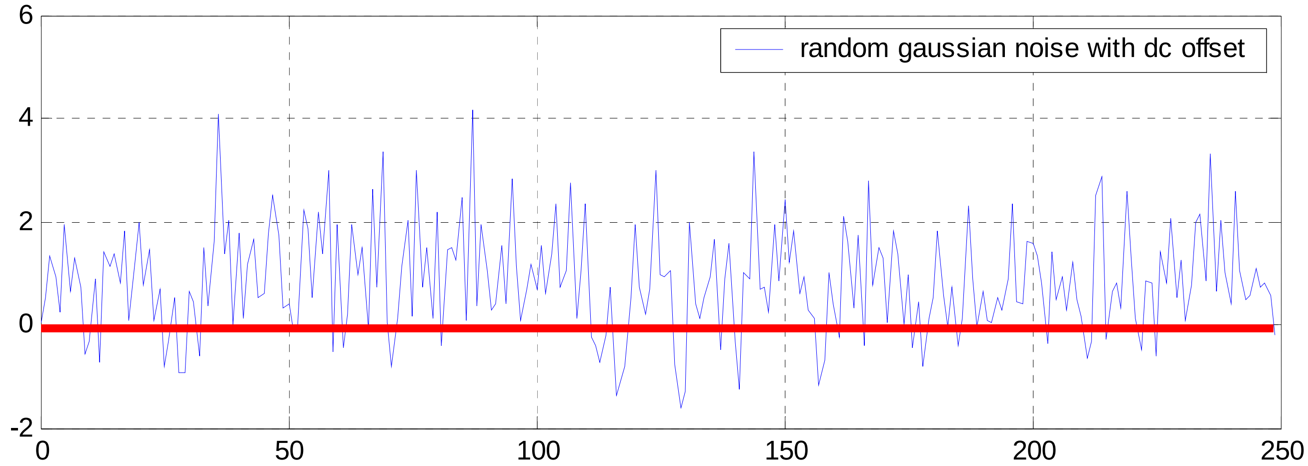
ZC=8

Zero Crossings for Noise



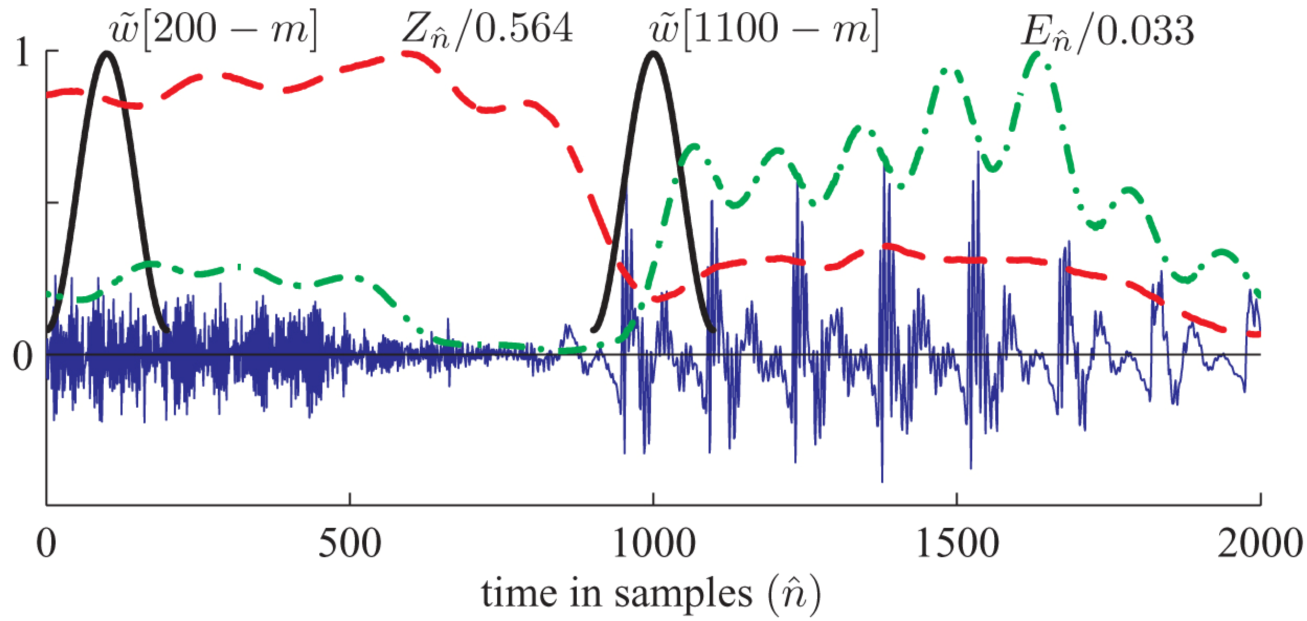
ZC=252

Offset=0.75

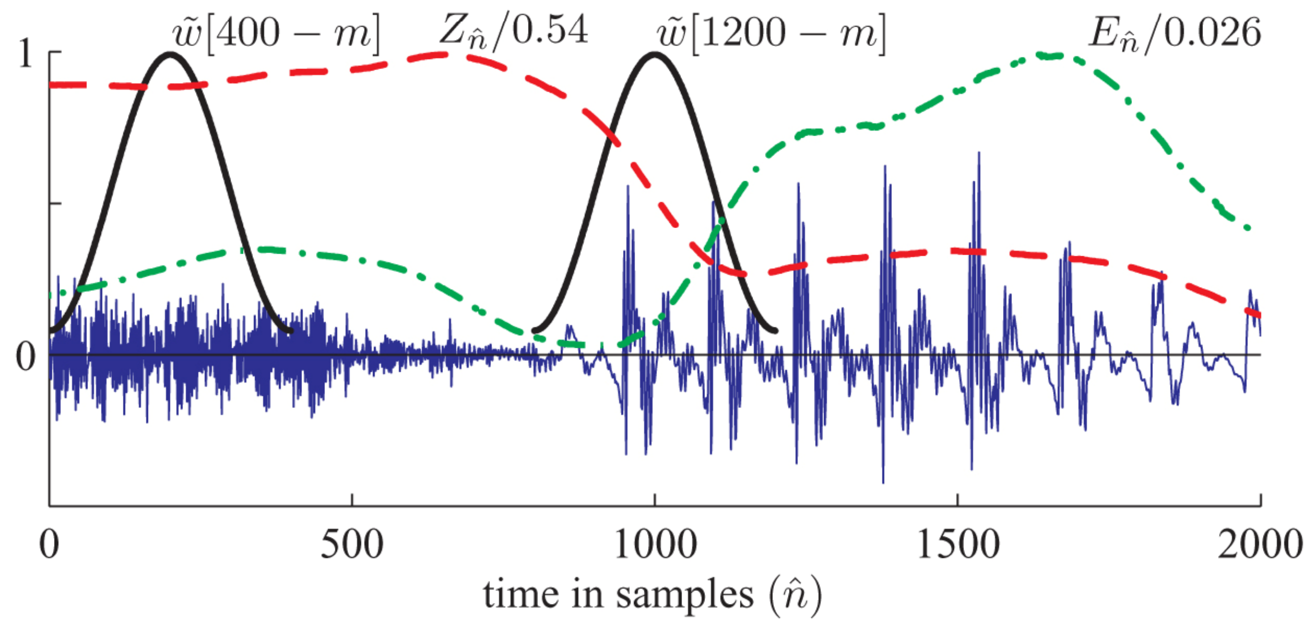


ZC=122

ZC and Energy Computation



Hamming window with duration *L=201* samples (12.5 msec at *Fs*=16 kHz)



Hamming window with duration *L=401* samples (25 msec at *Fs*=16 kHz)

ZC Rate Distributions

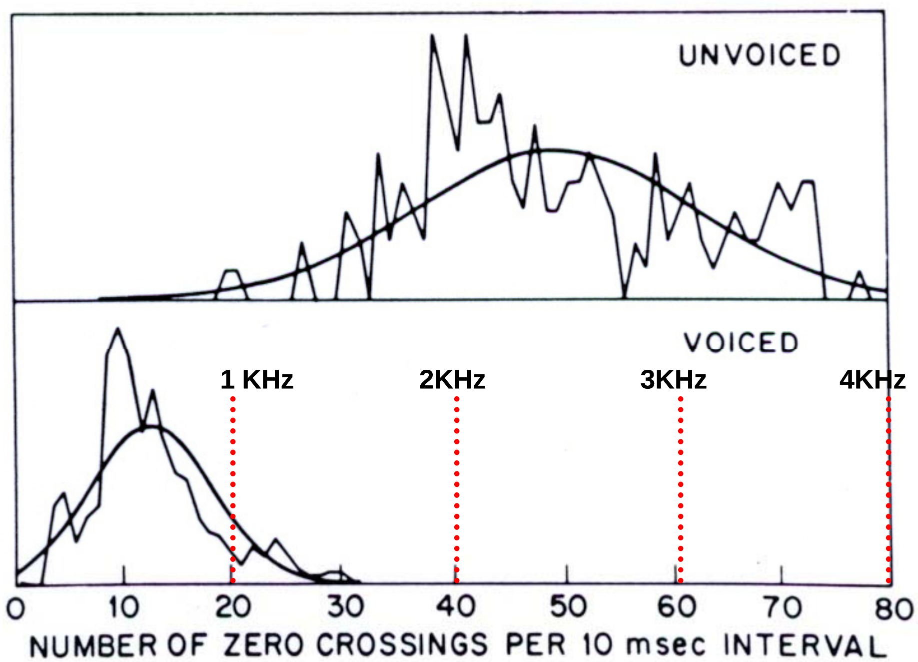


Fig. 4.11 Distribution of zero-crossings for unvoiced and voiced speech.

Unvoiced Speech: the dominant energy component is at about 2.5 kHz

Voiced Speech: the dominant energy component is at about 700 Hz

- for voiced speech, energy is mainly below 1.5 kHz
- for unvoiced speech, energy is mainly above 1.5 kHz
- mean ZC rate for unvoiced speech is 49 per 10 msec interval
- mean ZC rate for voiced speech is 14 per 10 msec interval

ZC Rates for Speech

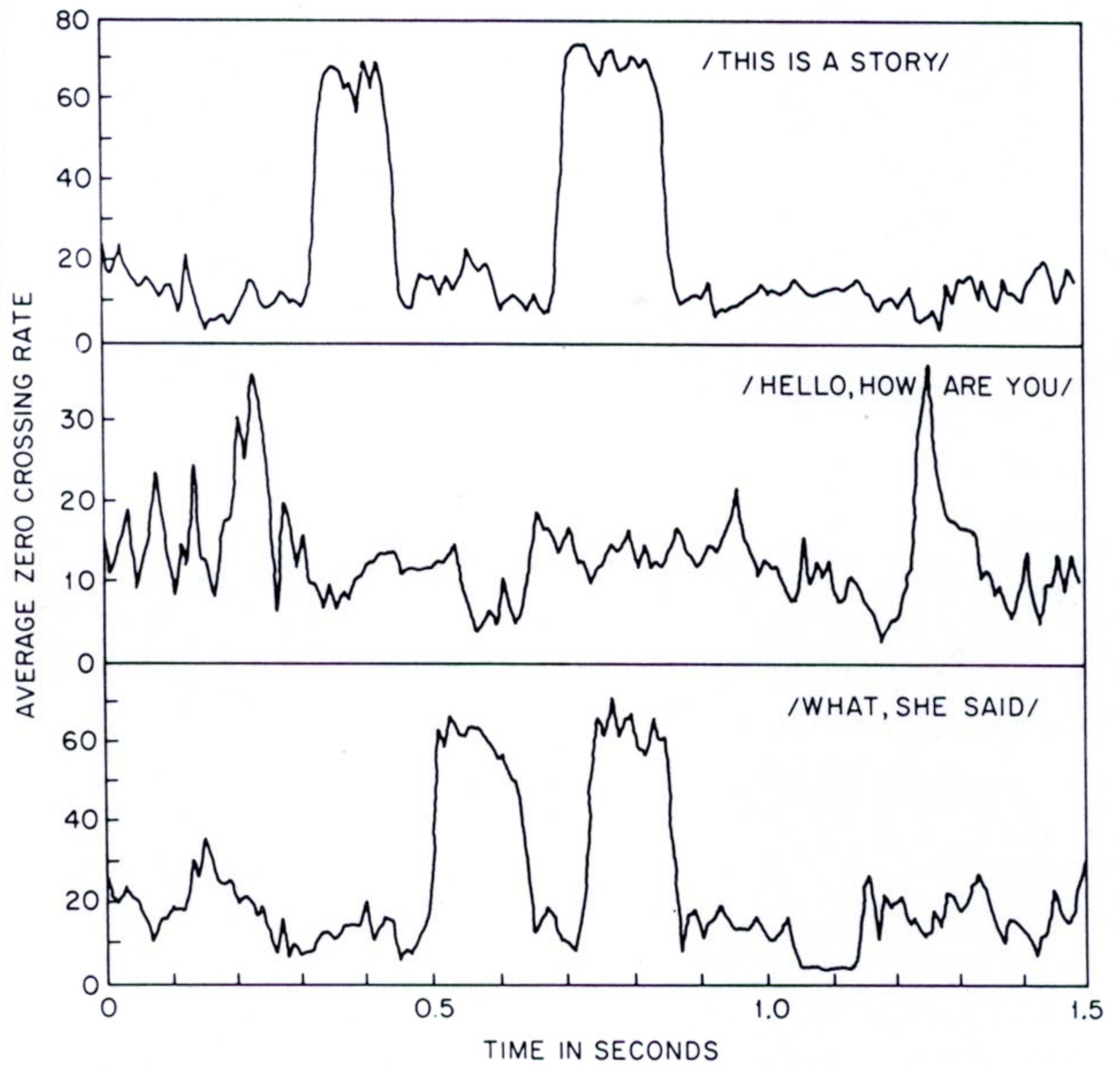
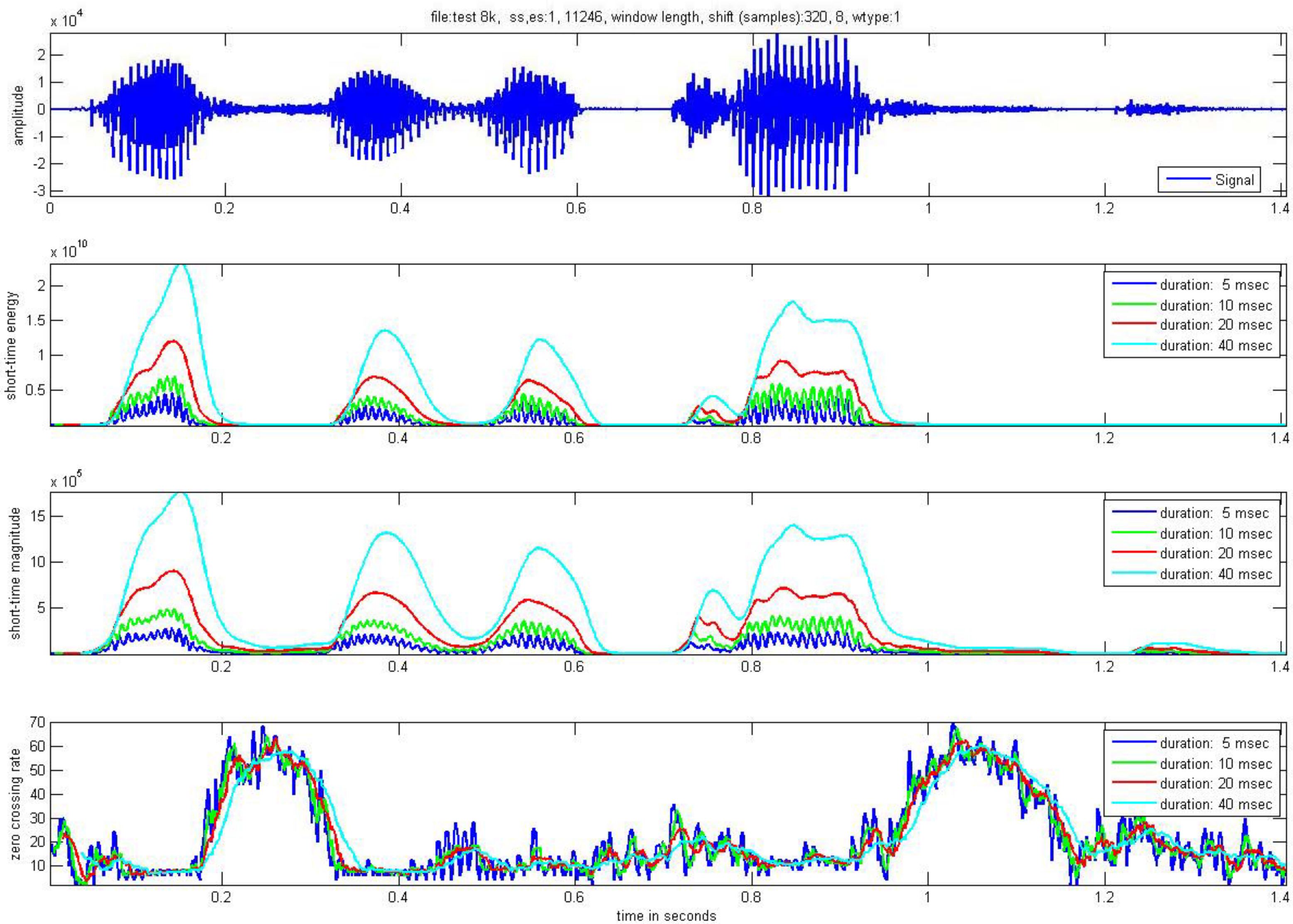


Fig. 4.12 Average zero-crossing rate for three different utterances.

- **15 msec windows**
- **100/sec sampling rate on ZC computation**

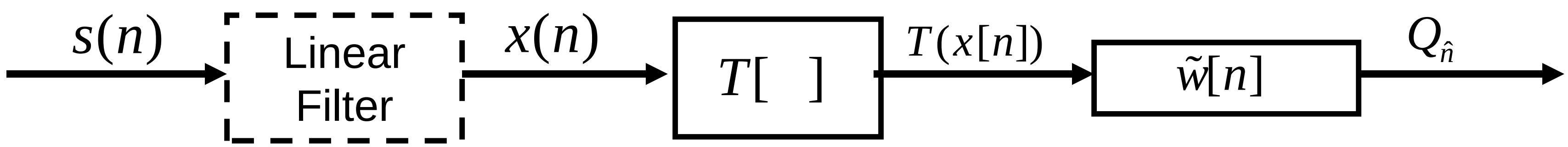
Short-Time Energy, Magnitude, ZC



Issues in ZC Rate Computation

- for zero crossing rate to be accurate, need zero DC in signal => need to remove offsets, hum, noise => use bandpass filter to eliminate DC and hum
- can quantize the signal to 1-bit for computation of ZC rate
- can apply the concept of ZC rate to bandpass filtered speech to give a ‘crude’ spectral estimate in narrow bands of speech (kind of gives an estimate of the strongest frequency in each narrow band of speech)

Summary of Simple Time Domain Measures



$$Q_{\hat{n}} = \sum_{m=-\infty}^{\infty} T(x[m])\tilde{w}[\hat{n}-m]$$

1. Energy:

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} x^2[m]\tilde{w}[\hat{n}-m]$$

- can downsample $E_{\hat{n}}$ at rate commensurate with window bandwidth

2. Magnitude:

$$M_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} |x[m]|\tilde{w}[\hat{n}-m]$$

3. Zero Crossing Rate:

$$Z_{\hat{n}} = z_1 = \sum_{m=\hat{n}-L+1}^{\hat{n}} |\mathrm{sgn}(x[m]) - \mathrm{sgn}(x[m-1])|\tilde{w}[\hat{n}-m]$$

where $\mathrm{sgn}(x[m]) = 1 \quad x[m] \geq 0$

$\qquad\qquad\qquad\quad = -1 \quad x[m] < 0$

Summary

- Short-time parameters in the domain
  - short-time energy time energy
  - short-time average magnitude
  - Short-time zero crossing rate (ZCR)

- Can be used in distinguishing fore/background

DEEE725 Speech Signal Processing Lab

Gil-Jin Jang

END OF LECTURE 06
CHAPTER 6. TIME-DOMAIN METHODS FOR SPEECH PROCESSING