PPGEE
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PARAÍBA

Processamento Digital de Sinais de Voz

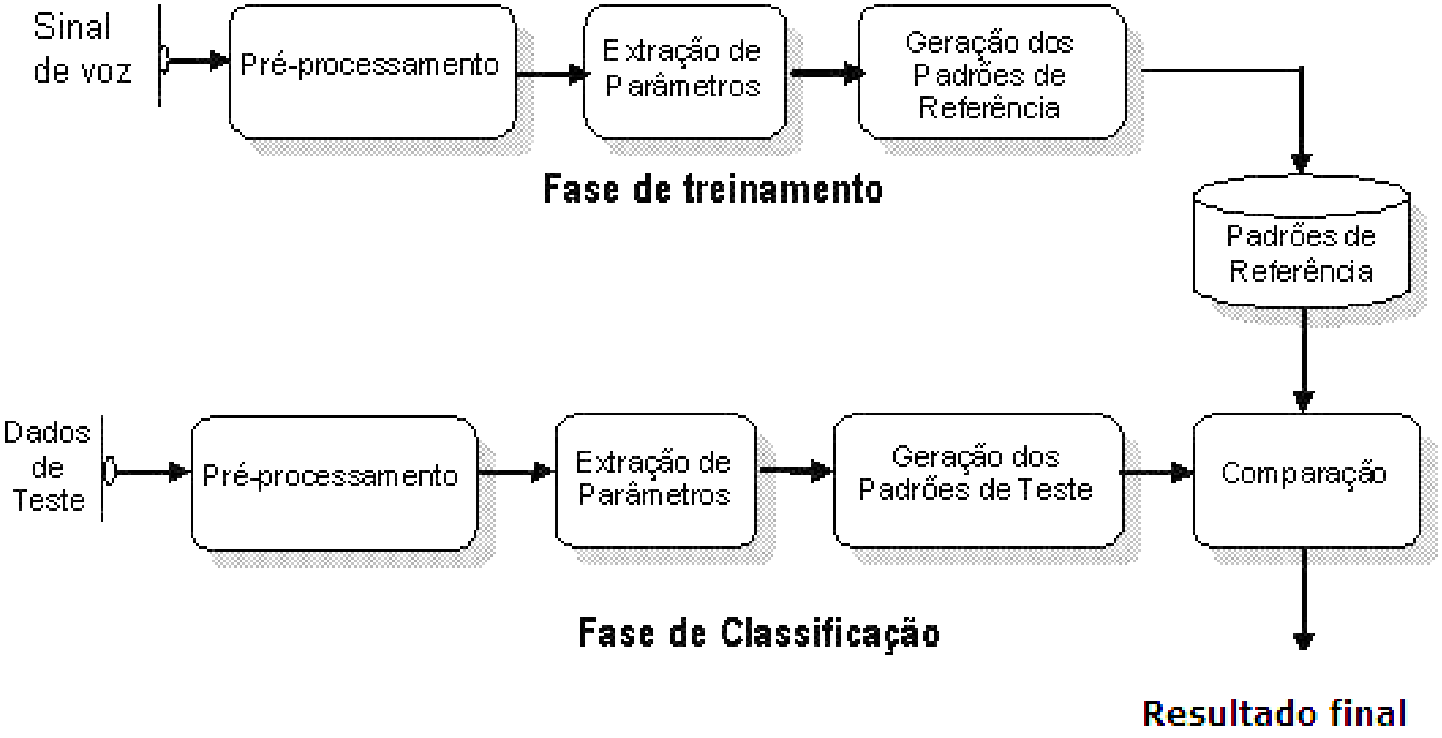
Pré-processamento de Sinais de Voz

☼

Análise de sinais de voz a curto intervalo de tempo

Profa. Silvana Luciene do N. Cunha Costa, D.Sc.

Sistema geral de Classificação



Pré-processamento

- Filtragem
- Divisão em quadros/segmentação
- Pré-ênfase
- Janelamento
  - Retangular
  - Hamming
  - Hanning
  - Blackman

Filtragem

- Limitação da largura de faixa → economia na energia espectral;
- Redução de ruído de fundo
- Realce de frequências
- Retirada de sinais indesejáveis -> sinais interferentes; eliminação dos 60 Hz (sinais biomédicos, por ex.)

Segmentação/Divisão em quadros

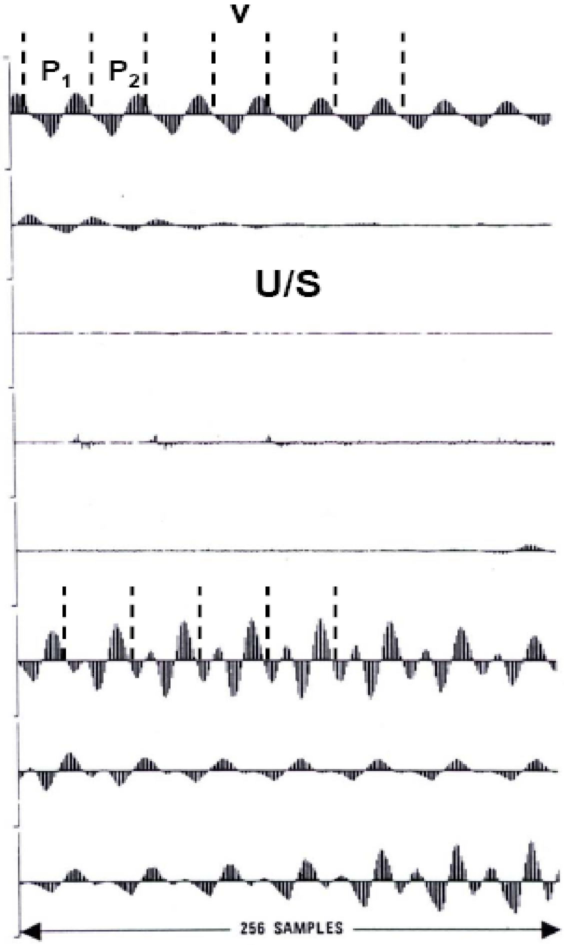
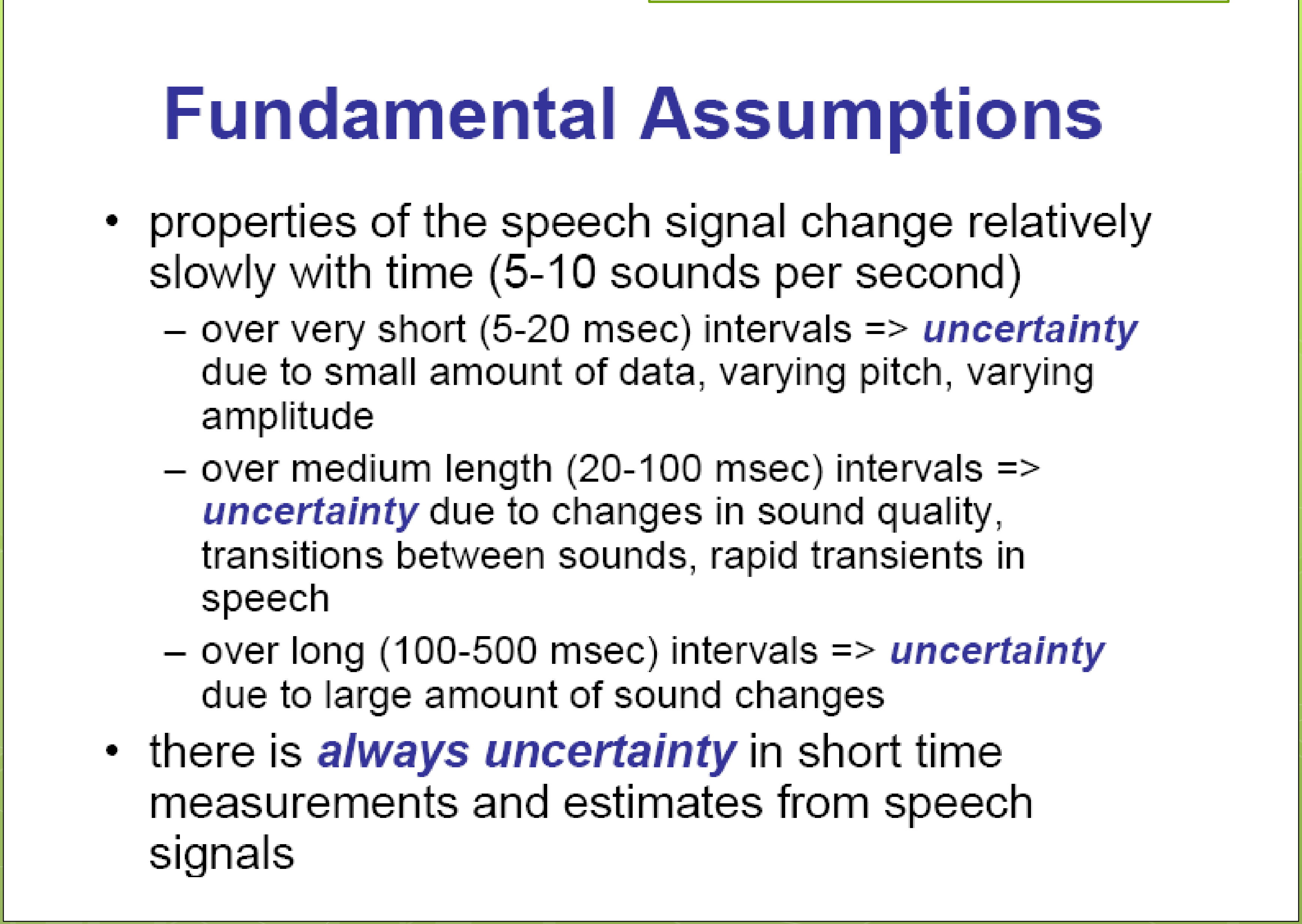


Fig. 4.1 Samples of a typical speech waveform (8 kHz sampling rate).

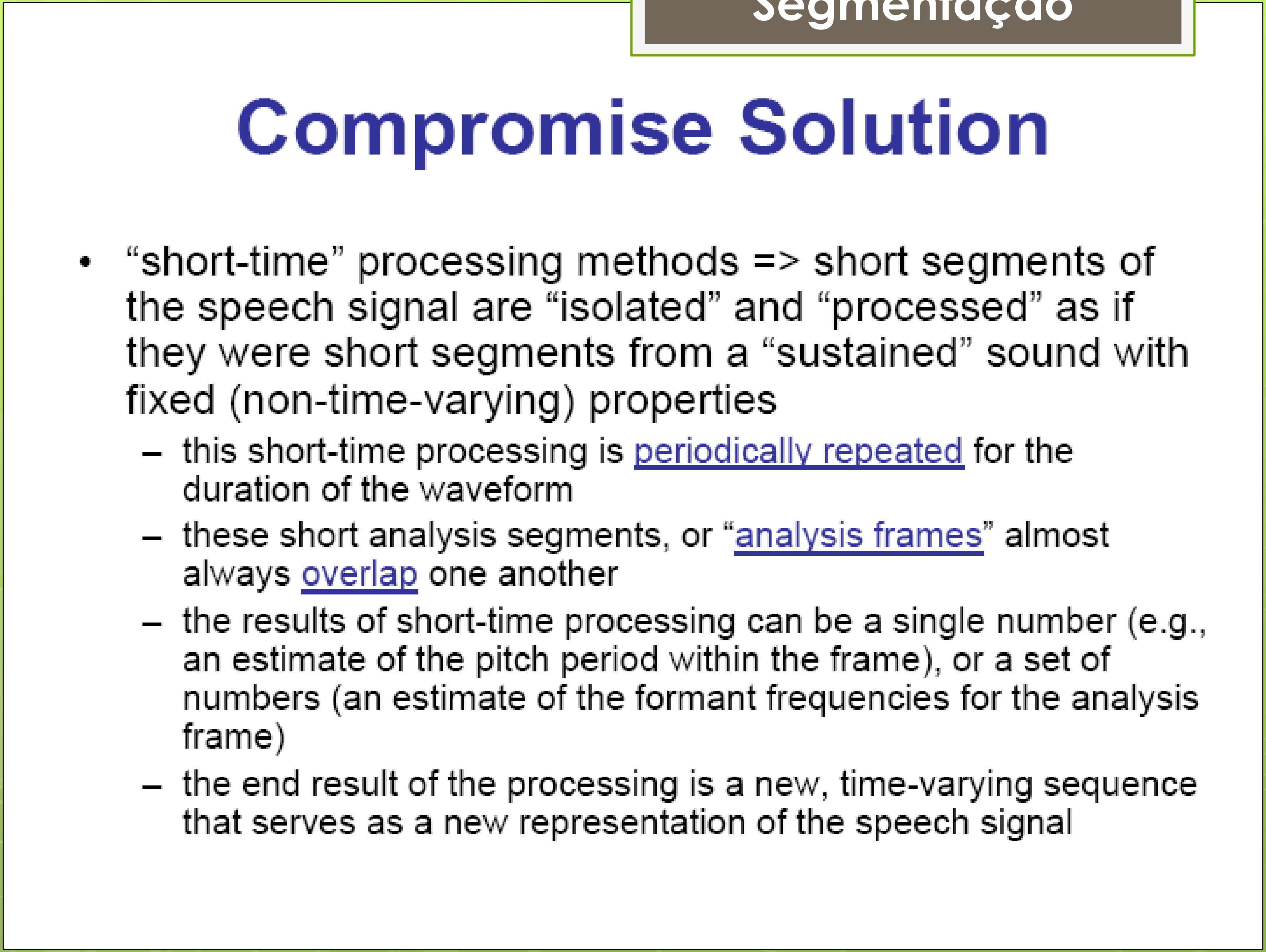
- 8 kHz sampled speech (bandwidth < 4 kHz)
- properties of speech change with time
  - excitation goes from voiced to unvoiced
  - peak amplitude varies with the sound being produced
  - pitch varies within and across voiced sounds
  - periods of silence where background signals are seen
- the key issue is whether we can create simple time-domain processing methods that enable us to measure/estimate speech *representations* reliably and accurately

Segmentação



Fundamental Assumptions

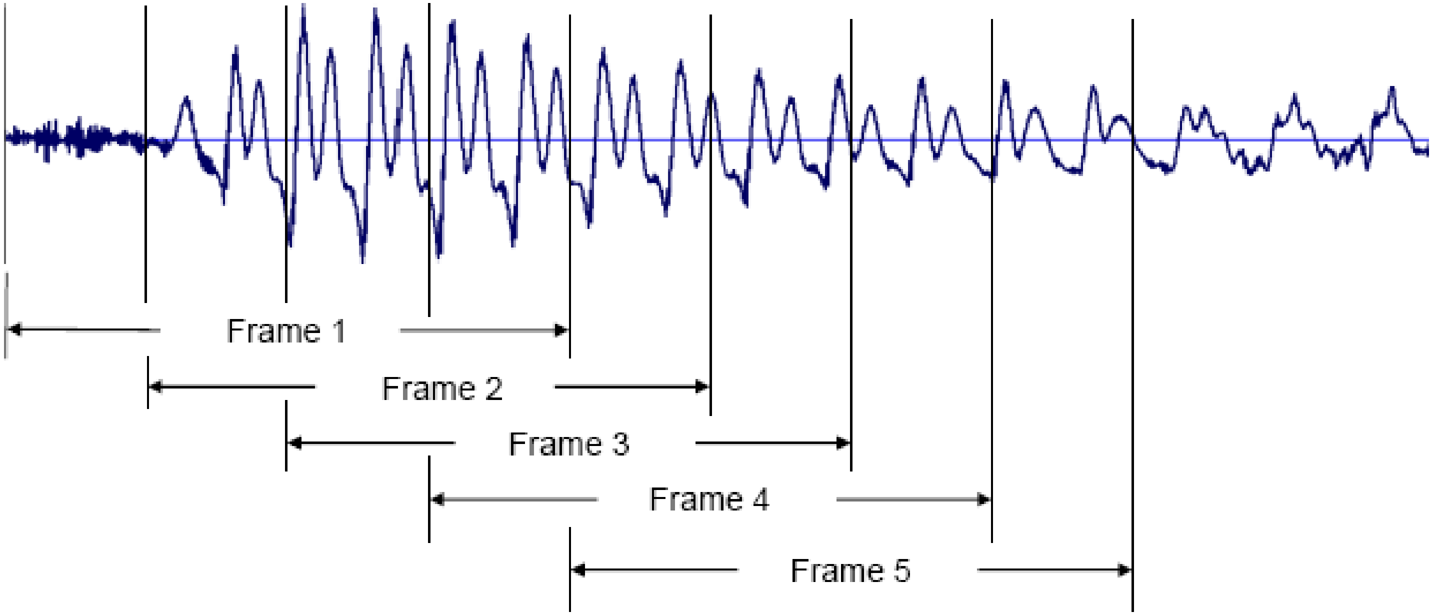
- properties of the speech signal change relatively slowly with time (5-10 sounds per second)
  - over very short (5-20 msec) intervals => ***uncertainty*** due to small amount of data, varying pitch, varying amplitude
  - over medium length (20-100 msec) intervals => ***uncertainty*** due to changes 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

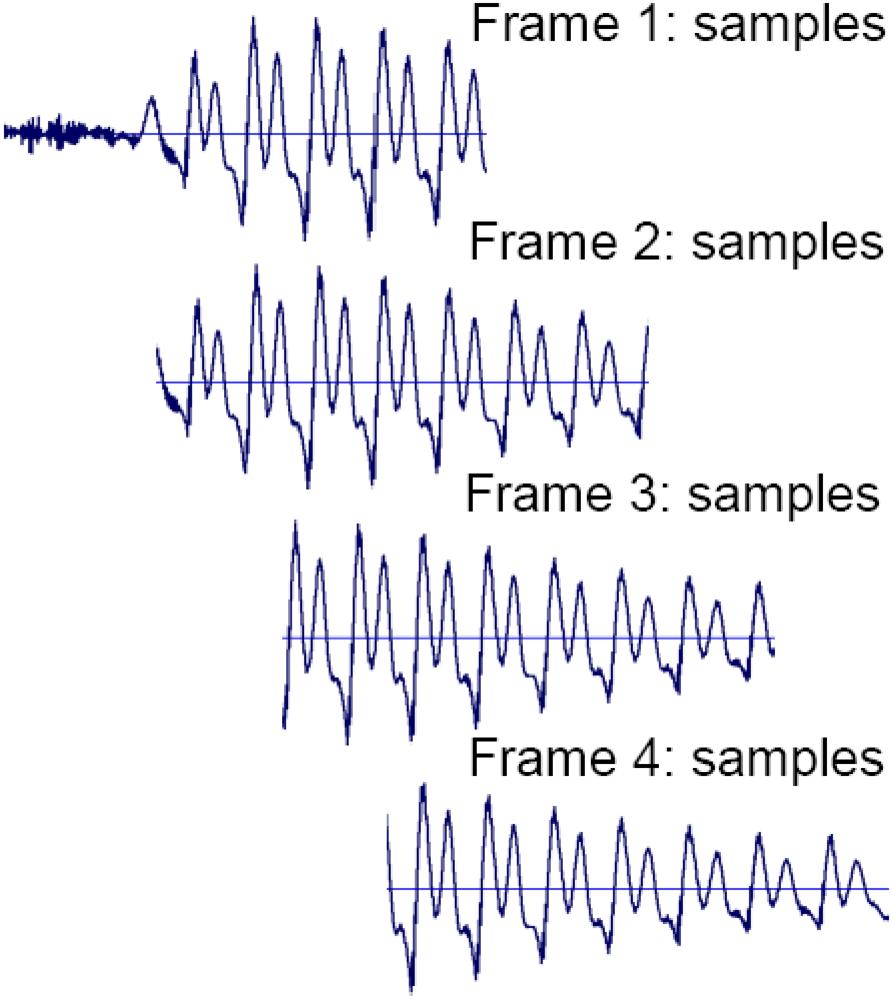
- "short-time" processing methods => short segments of the speech signal are "isolated" and "processed" as if they were short segments from a "sustained" sound with fixed (non-time-varying) properties
  - this short-time processing is periodically repeated for the duration of the waveform
  - these short analysis segments, or "analysis frames" almost always 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]}
…

Segmentação



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

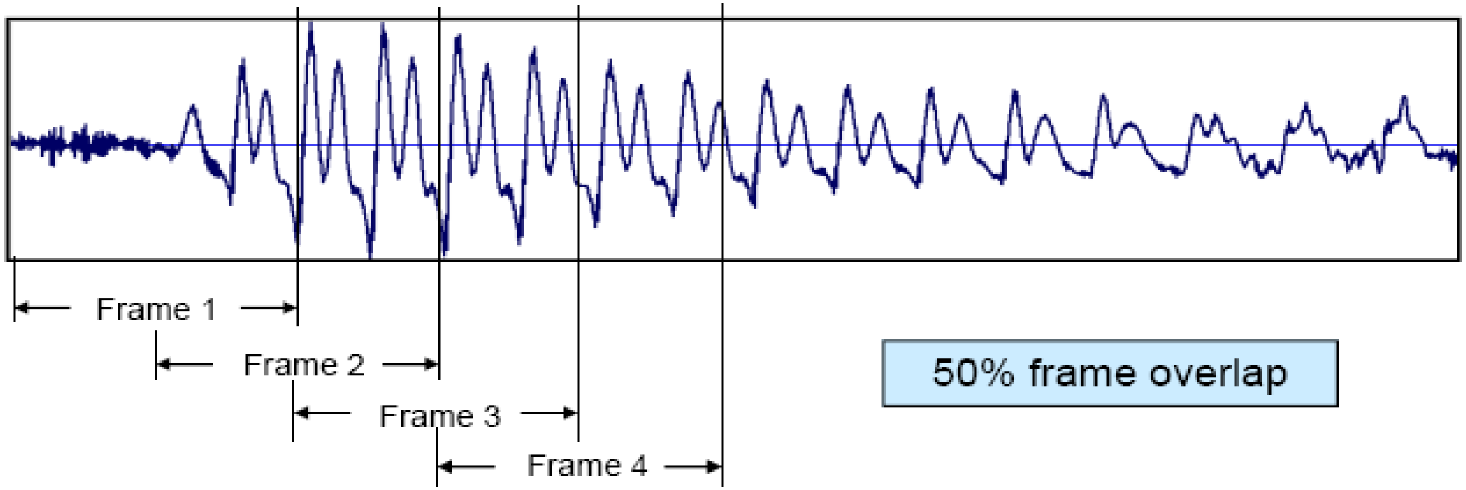
Frame 2: samples $R, R+1, \ldots, R+L-1$

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

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

Segmentação

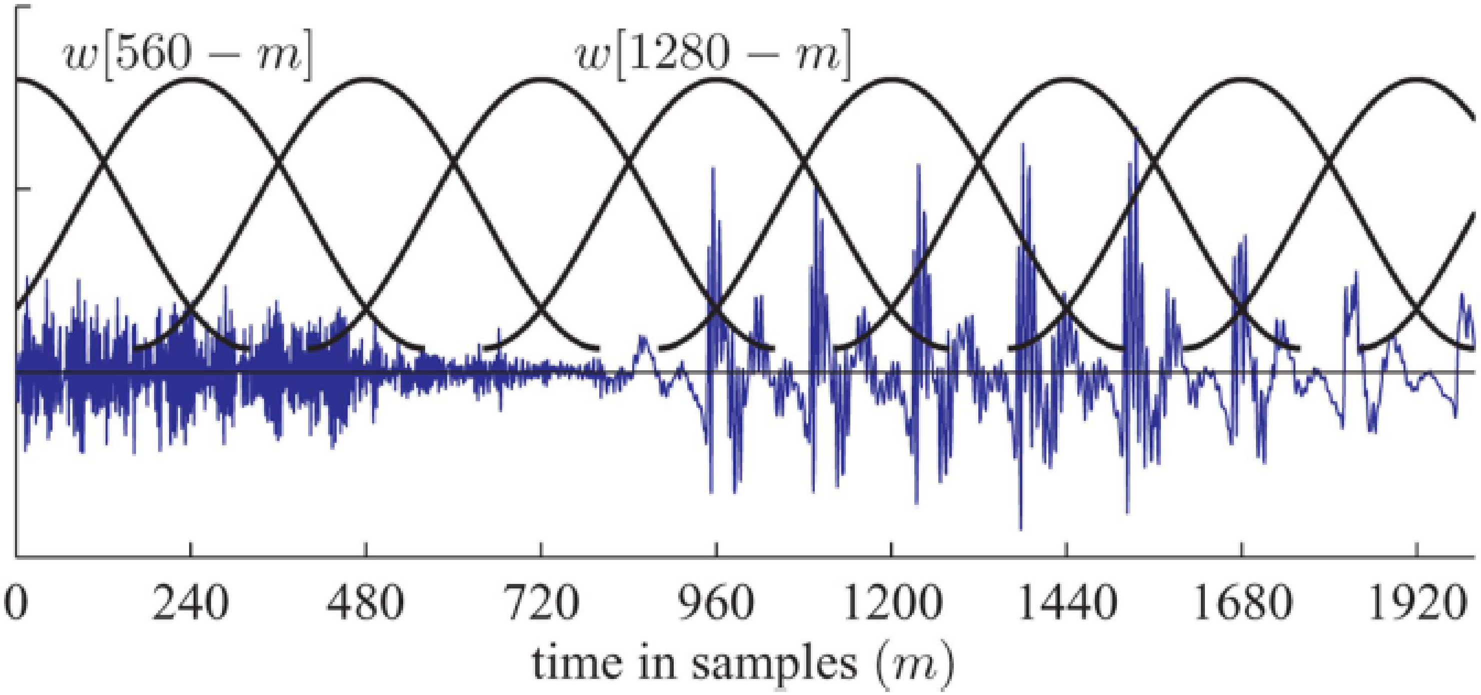
Frame-by-Frame Processing in Successive Windows



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

Pré-ênfase

- Proporciona compensação das perdas durante a passagem do sinal pelo trato vocal e pela radiação nos lábios (cerca de -6dB/oitava).

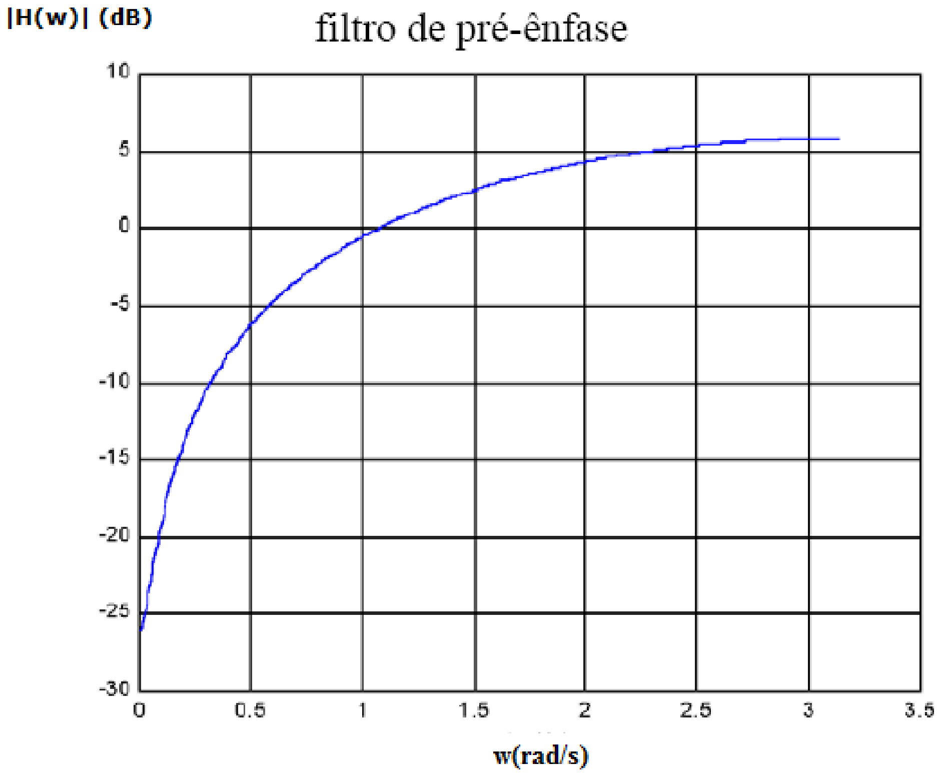
- Para solucionar esse problema é aplicado um filtro, de resposta de aproximadamente +6dB/oitava.

- Pode ser implementada, como uma operação digital no sinal amostrado, através de um filtro FIR de primeira ordem.

- Função de transferência do filtro: $H_p(z) = 1 - a_p\, z^{-1}$

$s_p(n) = s(n)$ - 0,95.$s(n$-1). (Valor típico de a_p = 0,95)

Função de transferência – filtro de pré-ênfase (a = 0,95)



Consiste de um filtro derivador, realçando as altas frequências.

Janelamento

- Sinal de voz - características estatísticas variam fortemente com o tempo, só podendo ser considerado estacionário em trechos muito pequenos - da ordem de dezenas de milissegundos - para efeito de obtenção de parâmetros.

- Segmentação – consiste em particionar o sinal de voz em segmentos, selecionados por janelas ou quadros de duração perfeitamente definida.

- A estimação espectral, na prática, é sempre feita em um trecho finito do sinal - Janelamento.

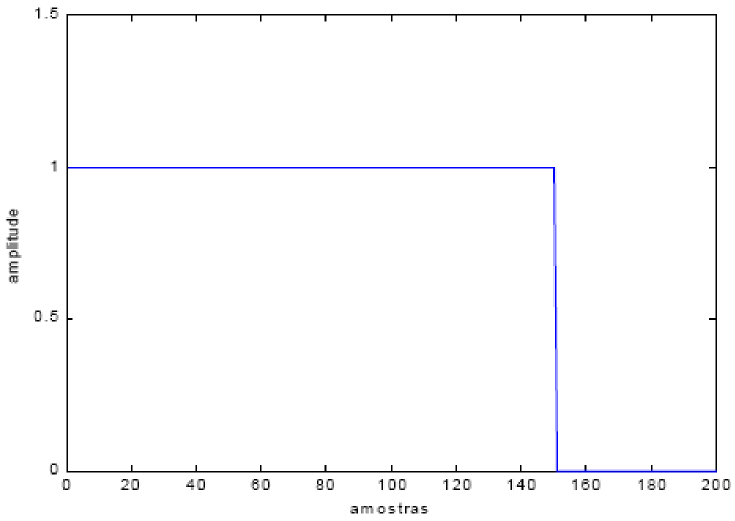
Janelamento

- Principais funções janela para a aplicação em processamento de voz:
  - Janela retangular: atribui igual peso a todas as amostras;
  - janelas Hamming e von Hann: atribuem pesos às amostras conforme a seguinte equação:

$$w[n] = \lambda + (1 - \lambda)\cos(2\pi n/(N-1))$$

$w[n]$ - função janela e N - número de pontos da janela.

Janela retangular



Janela retangular para N=150.

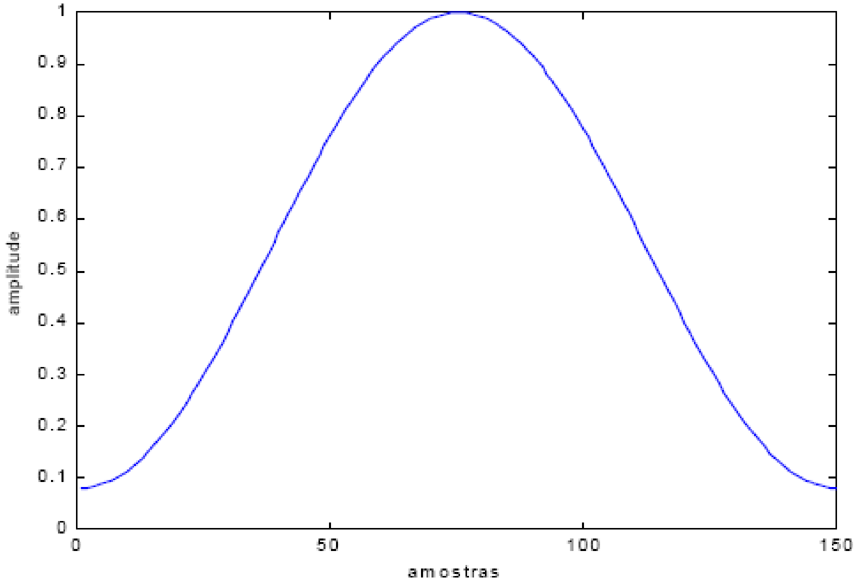
Dá o mesmo peso para todas as amostras.

Empregada no método da AMDF (Average Magnitude Difference Function), para detecção do *pitch*.

É o tipo de janela mais simples, sendo expressa pela seguinte função:

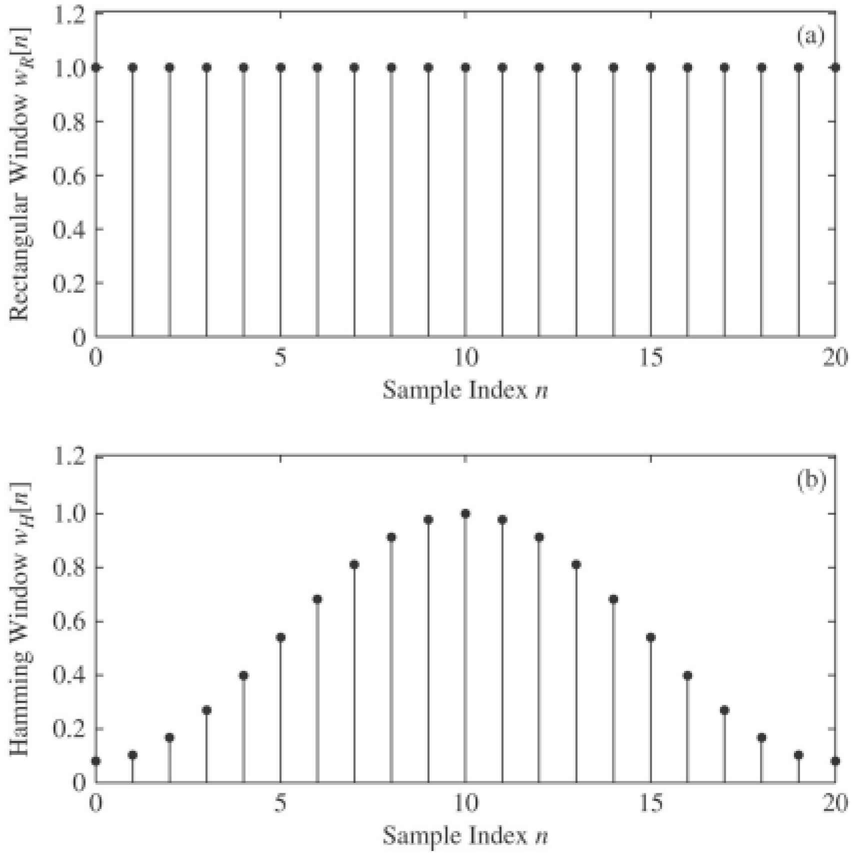
$$w(n) = \begin{cases} 1, \text{ para } 0<n\leq N \\ 0, \text{ para } n>N \end{cases}$$

Janela de Hamming



$$w(n) = \begin{cases} 0{,}54\text{-}0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right), \text{ para } 0 \le n \le N\text{-}1 \\ \\ 0, \text{ para } n \ge N \end{cases}$$

Janelas Retangular e Hamming



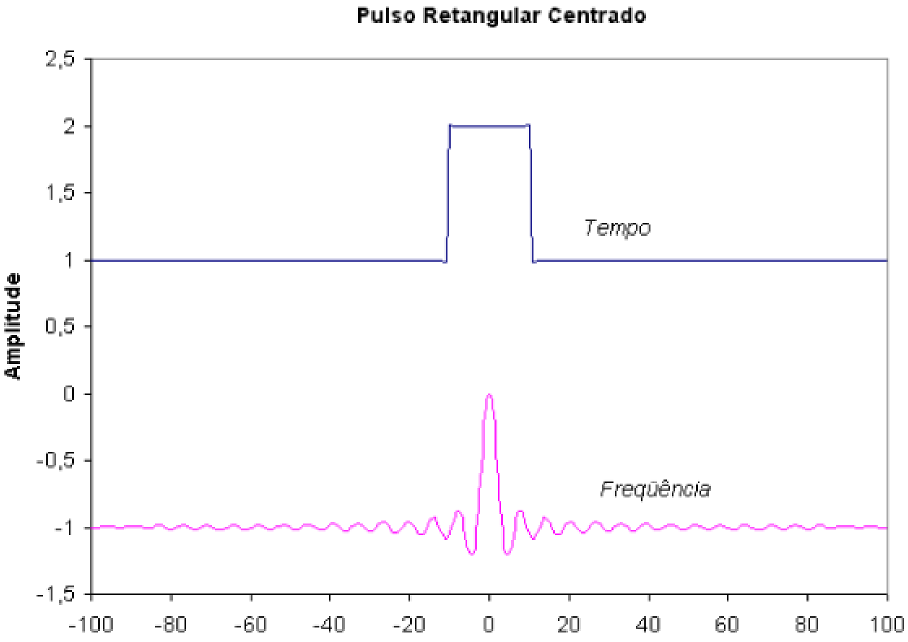
$L = 21$ samples

$$\tilde{w}_H[n] = 0.54\tilde{w}_R[n] - 0.46 * \cos(2\pi n / (L-1))\tilde{w}_R[n]$$

Janela de Hamming

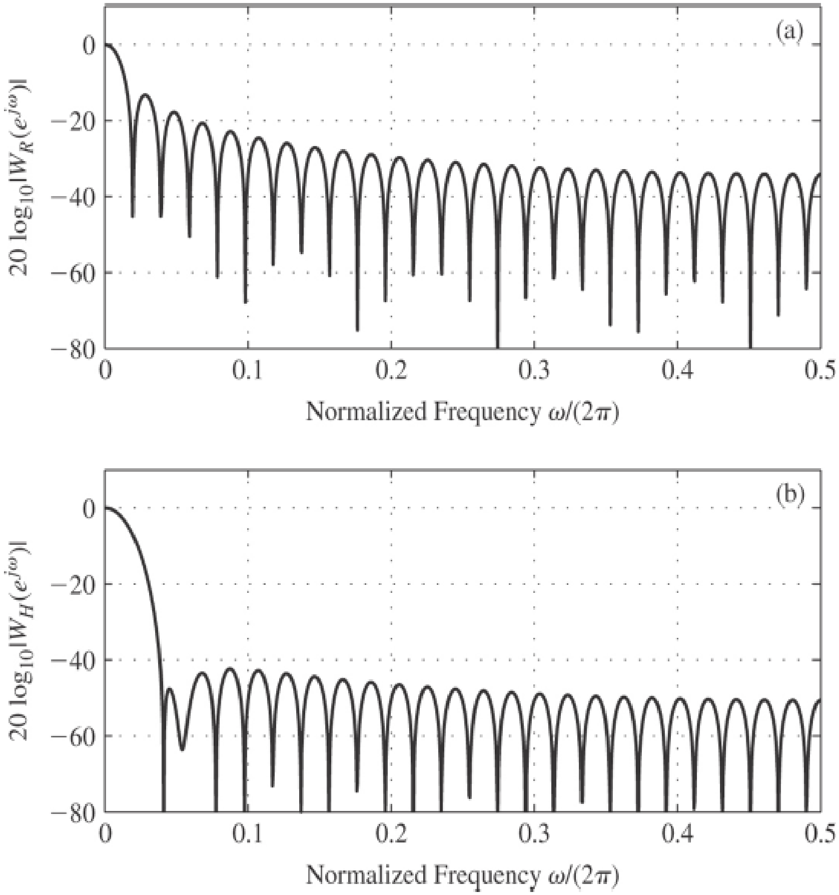
$$H(e^{j\Omega T}) = \frac{\sin(\Omega L T / 2)}{\sin(\Omega T / 2)} e^{-j\Omega T (L-1)/2}$$

O primeiro zero ocorre em *f=Fs/L=1/(LT)* (or *Ω=(2π)/(LT)*) → frequência de corte nominal do filtro 'passa-baixas' equivalente.

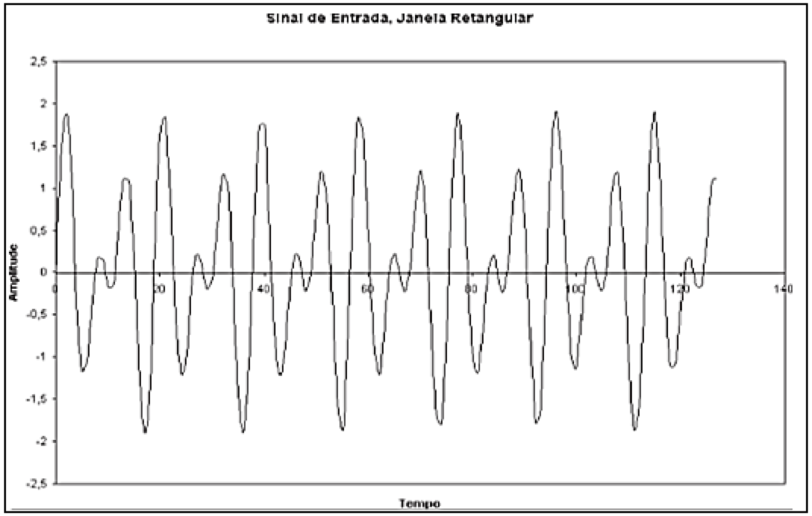


Resposta em frequência das Janelas Retangular (WR) e Hamming (WH)

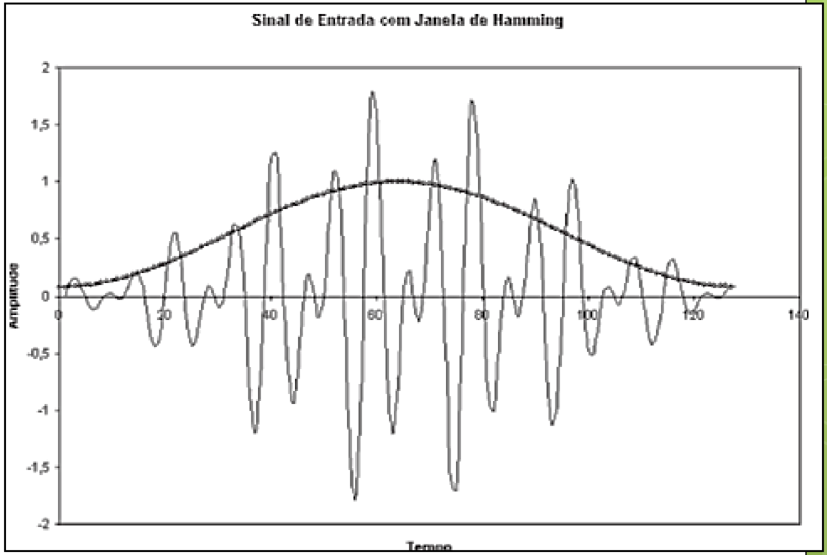
- Resposta em magnitude da WR e WH;
- Largura de faixa da WR é o dobro da WH;
- ***atenuação*** de mais de 40 dB para HW fora da faixa de passagem versus 14 dB para RW;
- Atenuação na faixa de rejeição é independente do comprimento (L) da janela;
- L → deve conter ao menos um período de pitch; deve manter a estacionaridade.



Exemplo

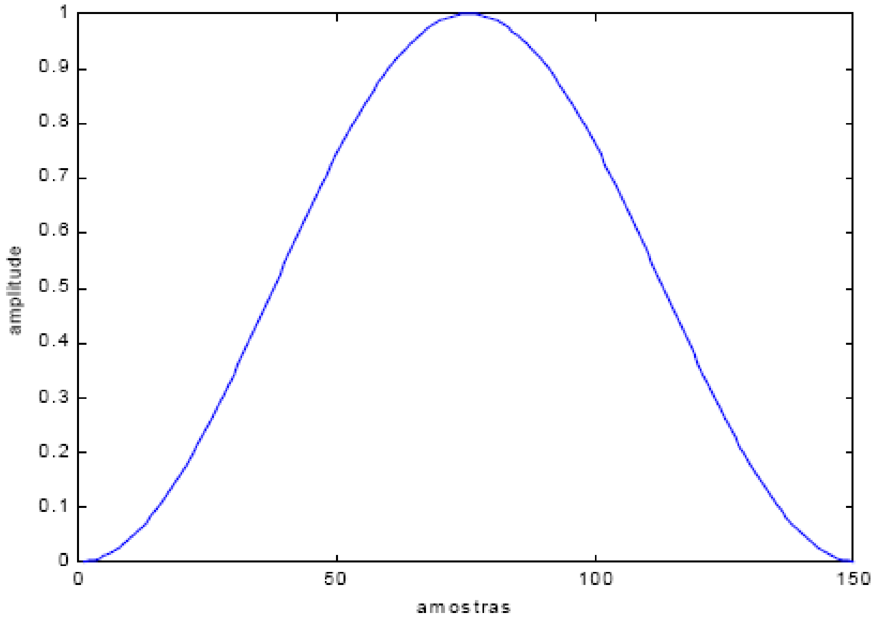


Janela retangular → interrupção repentina →vazamento no espectro



Janela de Hanning

$$w(n) = \begin{cases} 0{,}5\text{-}0{,}5\cos\left(\dfrac{2\pi n}{N+1}\right), \text{ para } 0 \leq n \leq N\text{-}1 \\ 0, \text{ para } n \geq N \end{cases}$$

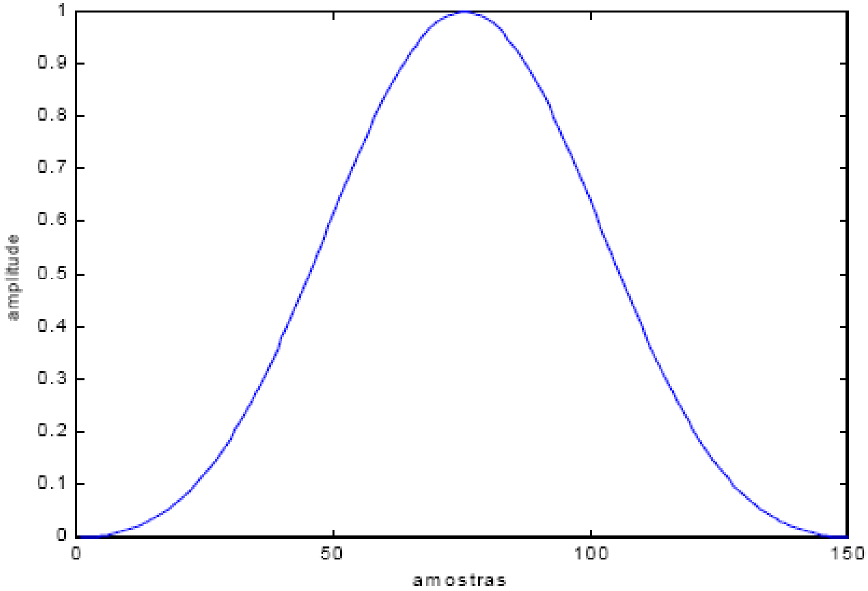


Janelamento

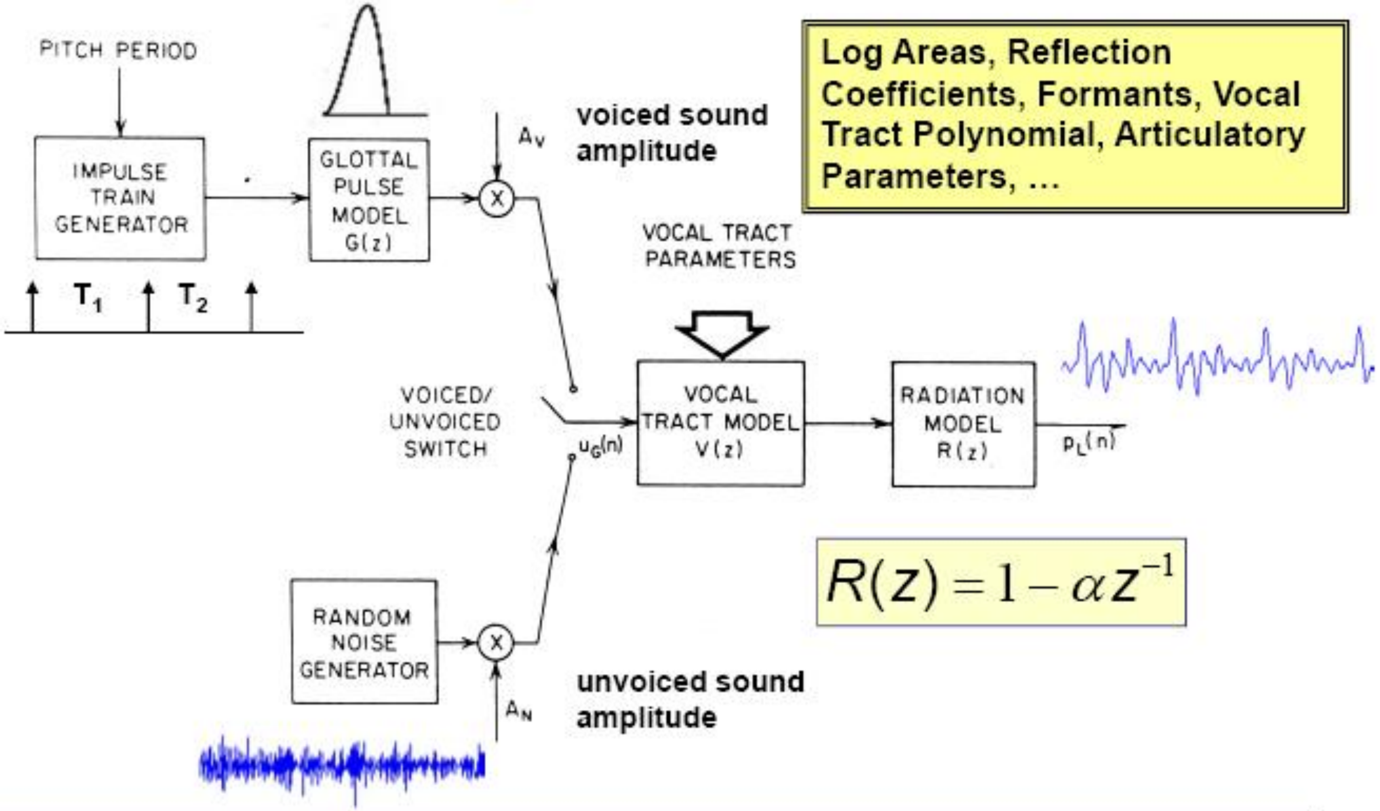
- **Janela Retangular**
  - Fugas espectrais alterando o espectro do sinal
- **Janela de Hamming**
  - Apresenta um lóbulo principal de amplitude bastante superior a dos lóbulos secundários – manutenção das características espectrais do centro do quadro e a eliminação das transições abruptas das extremidades.
- **Janela de Hanning**
  - Similar ao efeito da janela de Hamming, porém proporciona um reforço menor nas amostras do centro e uma suavização maior nas amostras da extremidade.

Janela de Blackman

$$w(n) = \begin{cases} 0{,}42\text{-}0{,}5\cos\left(\dfrac{2\pi n}{N-1}\right) + 0{,}8\ \cos\left(\dfrac{4\pi n}{N-1}\right),\ \text{para } 0 \leq n \leq N\text{-}1 \\ \\ 0,\ \text{para } n \geq N \end{cases}$$



General Synthesis Model



voiced sound amplitude

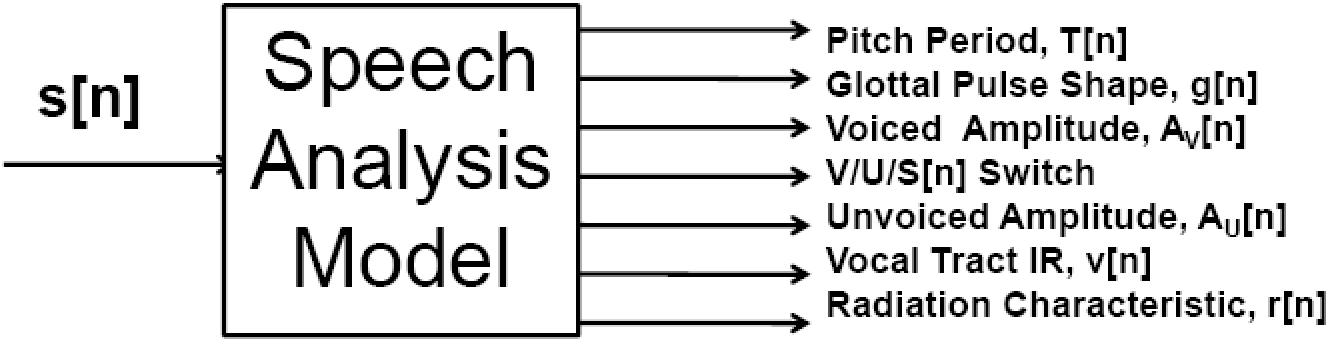
Log Areas, Reflection Coefficients, Formants, Vocal Tract Polynomial, Articulatory Parameters, …

$$R(z) = 1 - \alpha z^{-1}$$

unvoiced sound amplitude

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

General Analysis Model



s[n]

Speech Analysis Model

Pitch Period, T[n]

Glottal Pulse Shape, g[n]

Voiced Amplitude, $A_V[n]$

V/U/S[n] Switch

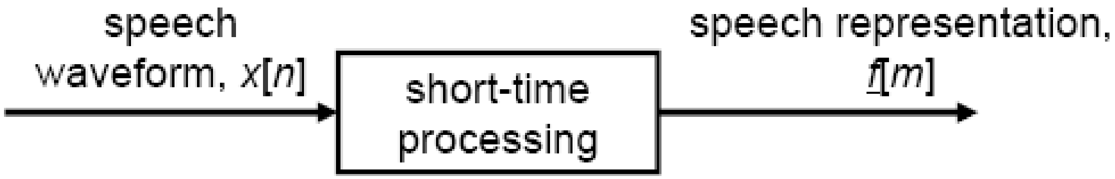
Unvoiced Amplitude, $A_U[n]$

Vocal Tract IR, v[n]

Radiation Characteristic, r[n]

- 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

Short-Time Processing

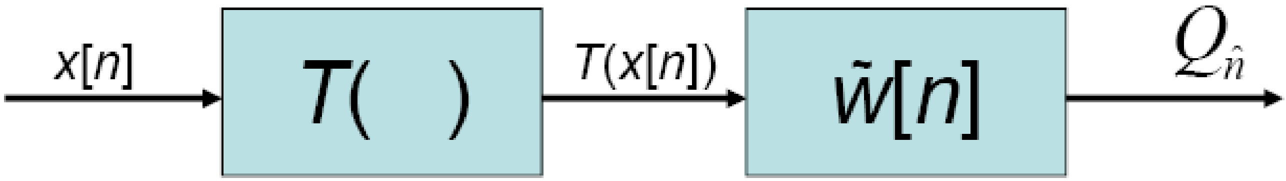


- $x[n]$ = samples at 8000/sec rate; (e.g. 2 seconds of 4 kHz bandlimited speech, $x[n]$, $0 \le n \le 16000$)

- $\vec{f}[m] = \{f_1[m], f_2[m], \ldots, f_L[m]\}$ = vectors at 100/sec rate, $1 \le m \le 200$, L is the size of the analysis vector (e.g., 1 for pitch period estimate, 12 for autocorrelation estimates, etc)

Generic Short-Time Processing

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



linear or non-linear transformation

window sequence (usually finite length)

- $Q_{\hat{n}}$ 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

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

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

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

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

Computation of Short-Time Energy

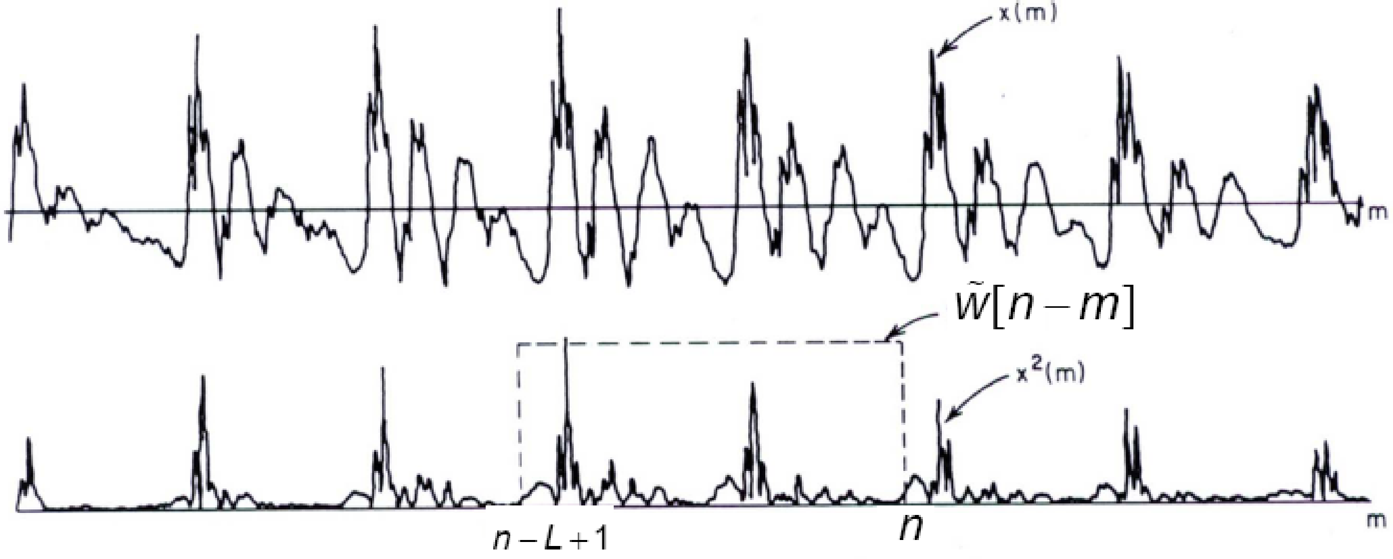
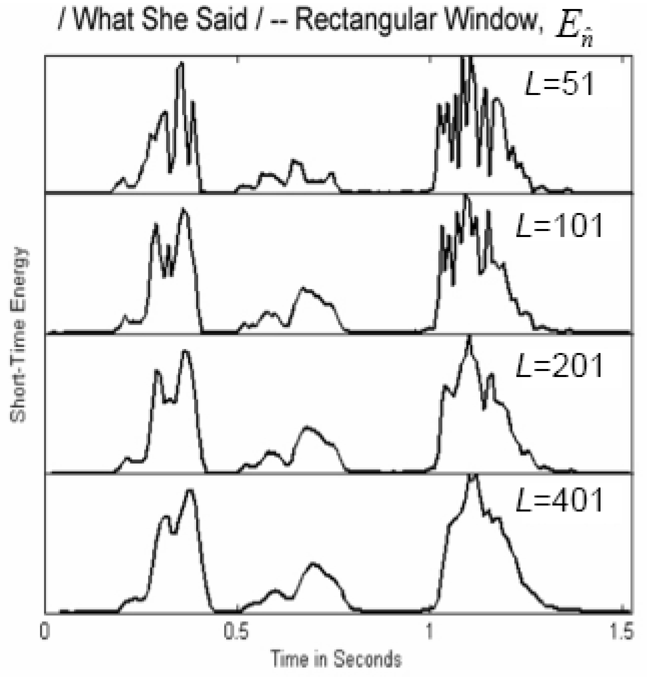
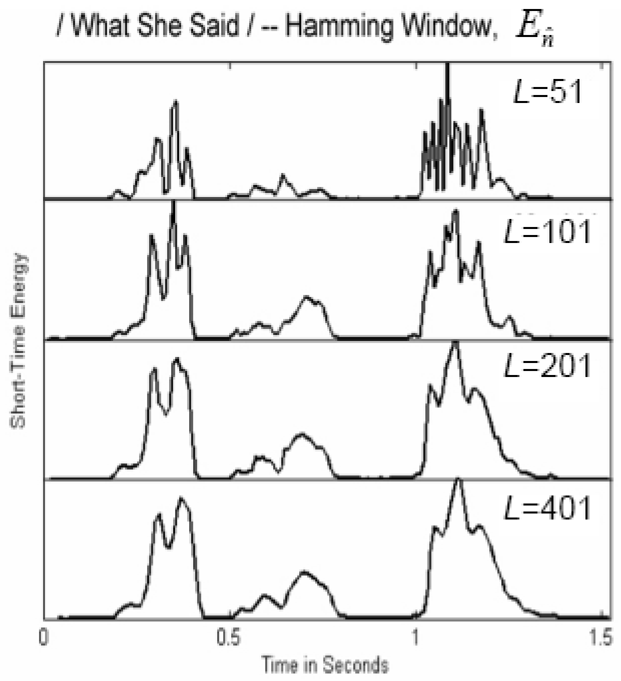


Fig. 4.2 Illustration of the computation of snort-time energy.

- **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 using RW/HW





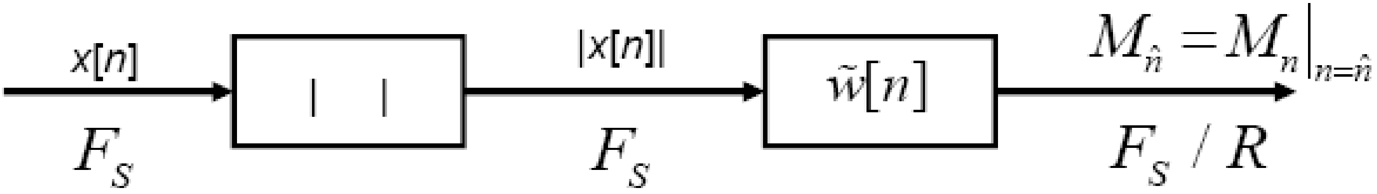
- as *L* increases, the plots tend to converge (however you are smoothing sound energies)
- short-time energy provides the basis for distinguishing voiced from unvoiced speech regions, and for medium-to-high SNR recordings, can even be used to find regions of silence/background signal

Short-Time Magnitude

- 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)

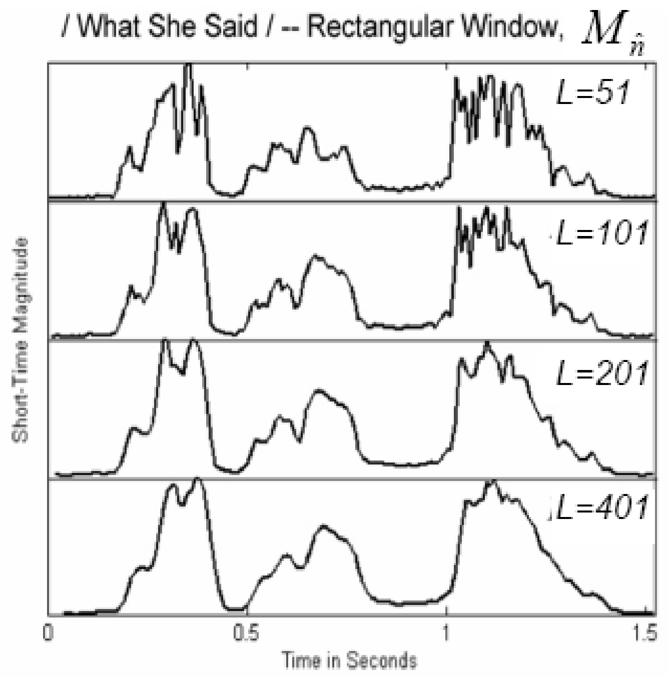
$$M_{\hat{n}} = \sum_{m=-\infty}^{\infty} |x[m]| \, \tilde{w}[\hat{n} - m]$$

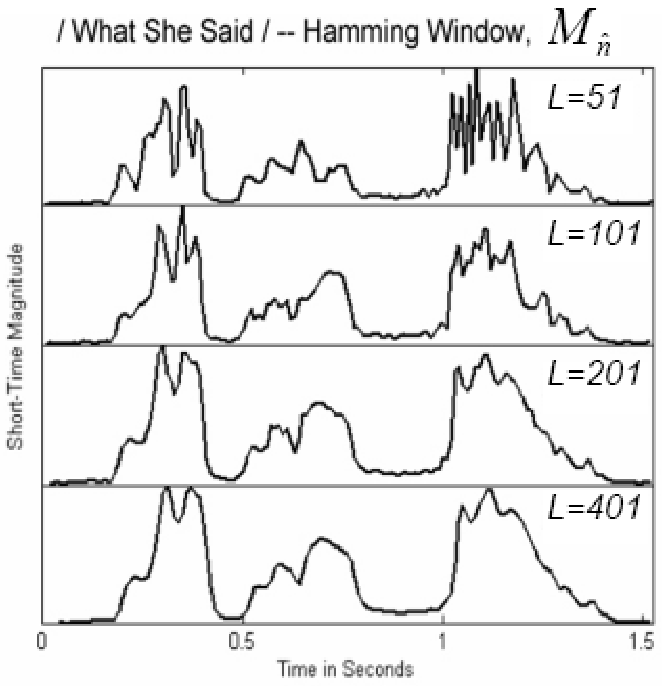
  - weighted sum of magnitudes, rather than weighted sum of squares



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

Short-Time Magnitudes



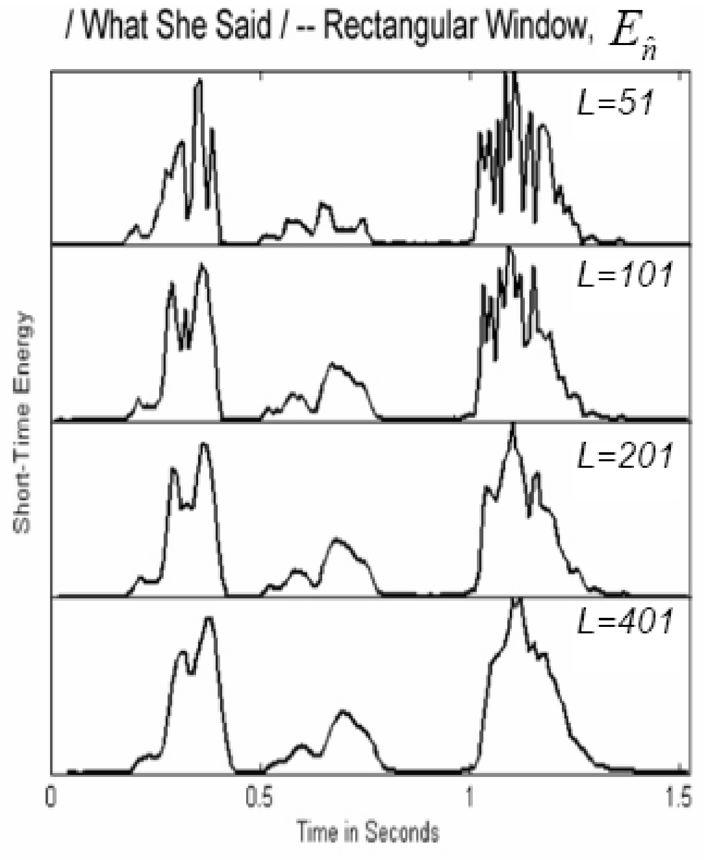


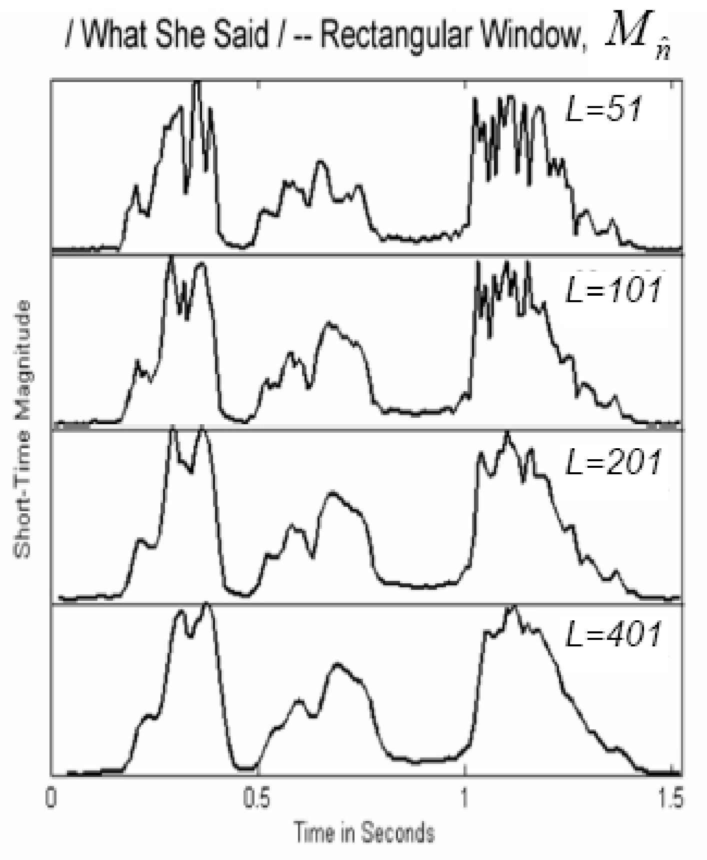
- differences between E_n and M_n noticeable in unvoiced regions

- dynamic range of M_n ~ square root (dynamic range of E_n) => level differences between voiced and unvoiced segments are smaller

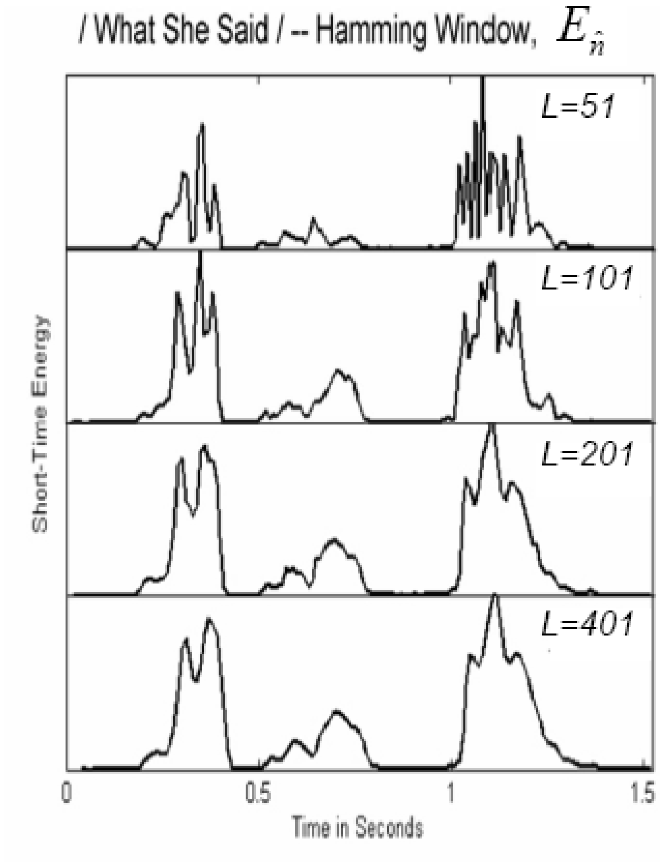
- E_n and M_n can be sampled at a rate of 100/sec for window durations of 20 msec or so => efficient representation of signal energy/magnitude

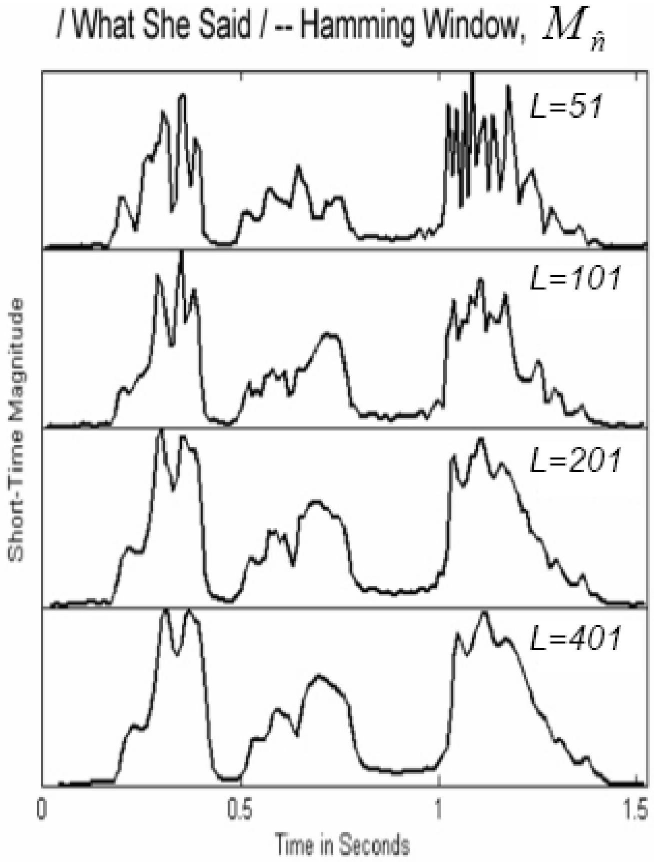
Short Time Energy and Magnitude—Rectangular Window



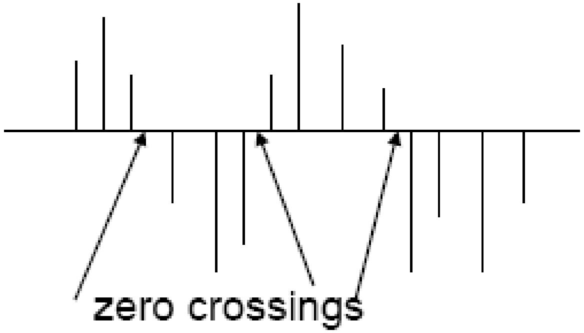


Short Time Energy and Magnitude—Hamming Window





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

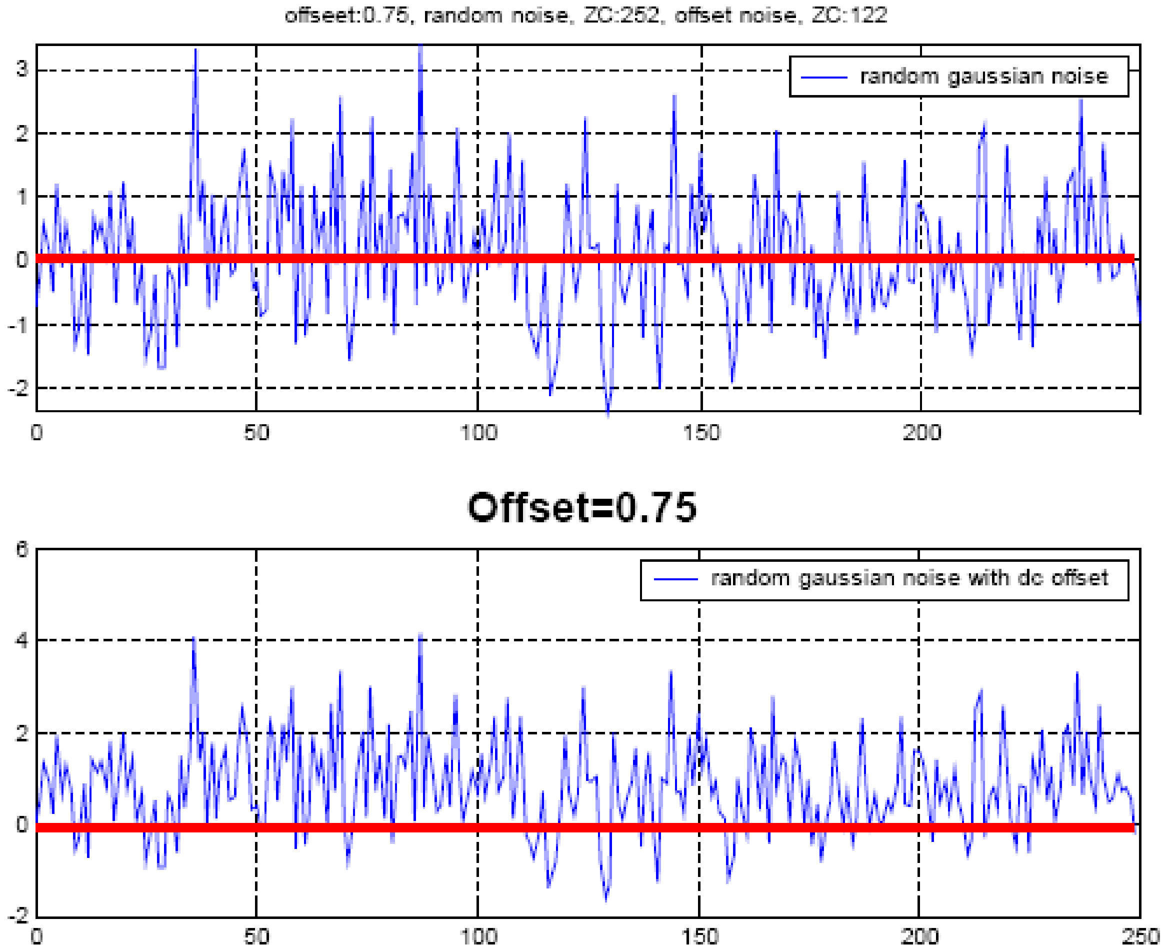
Assume the sampling rate is $F_S = 10,000$ Hz

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 Crossings for Noise

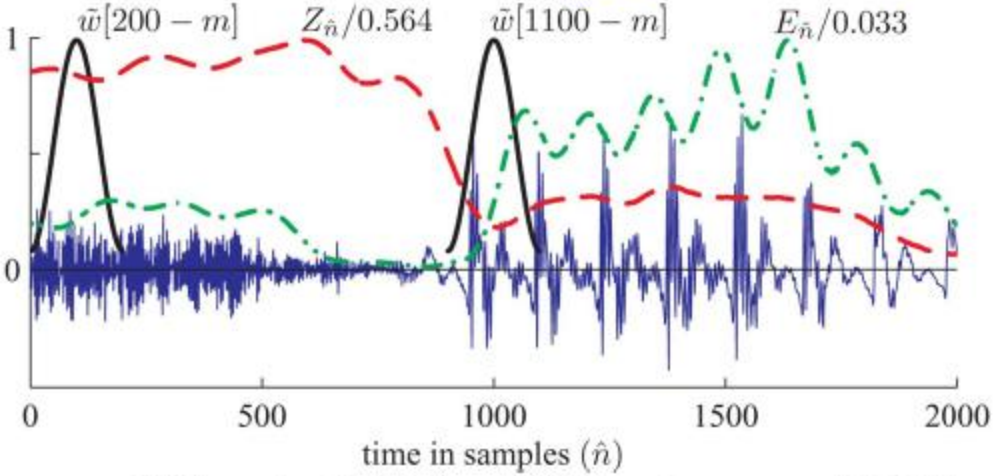


ZC=252

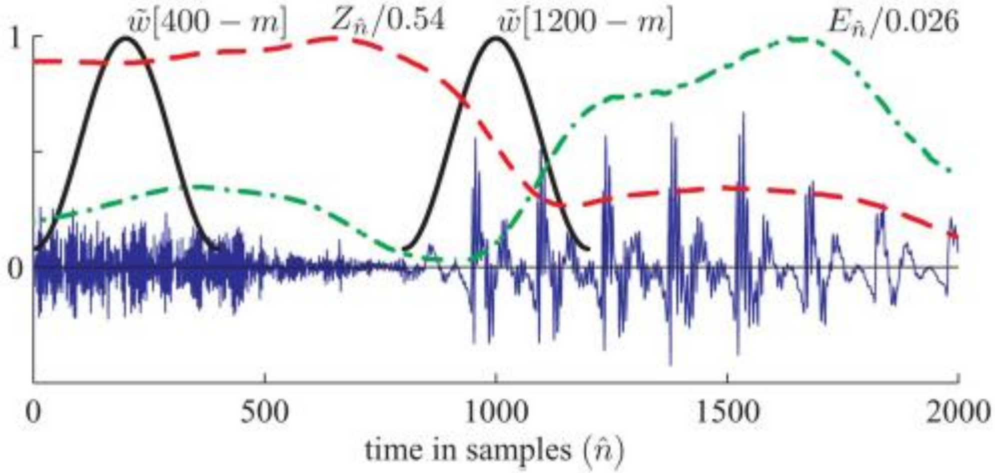
Offset=0.75

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)

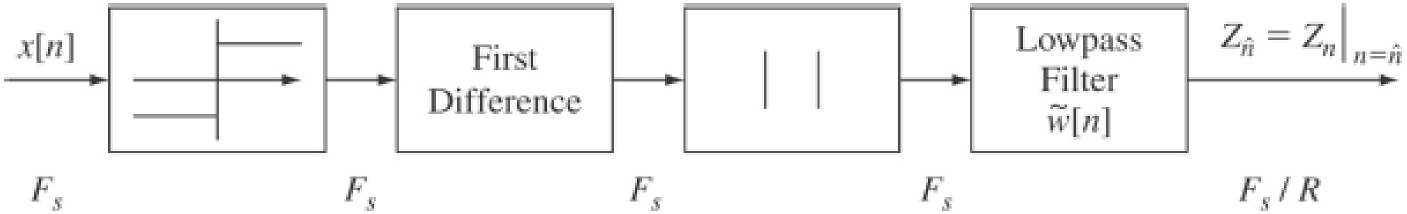
ZC Rate Definitions

$$Z_{\hat{n}} = \frac{1}{2L_{\text{eff}}} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\,\text{sgn}(x[m]) - \text{sgn}(x[m-1])\,|\,\tilde{w}[\hat{n}-m]$$

$$\begin{aligned} \text{sgn}(x[n]) &= 1 \qquad & x[n] \geq 0 \\ &= -1 \qquad & x[n] < 0 \end{aligned}$$

- simple rectangular window:

$$\begin{aligned} \tilde{w}[n] &= 1 \qquad & 0 \leq n \leq L-1 \\ &= 0 \qquad & \text{otherwise} \\ L_{\text{eff}} &= L & \end{aligned}$$



Same form for $Z_{\hat{n}}$ as for $E_{\hat{n}}$ or $M_{\hat{n}}$

ZC Rate Distributions

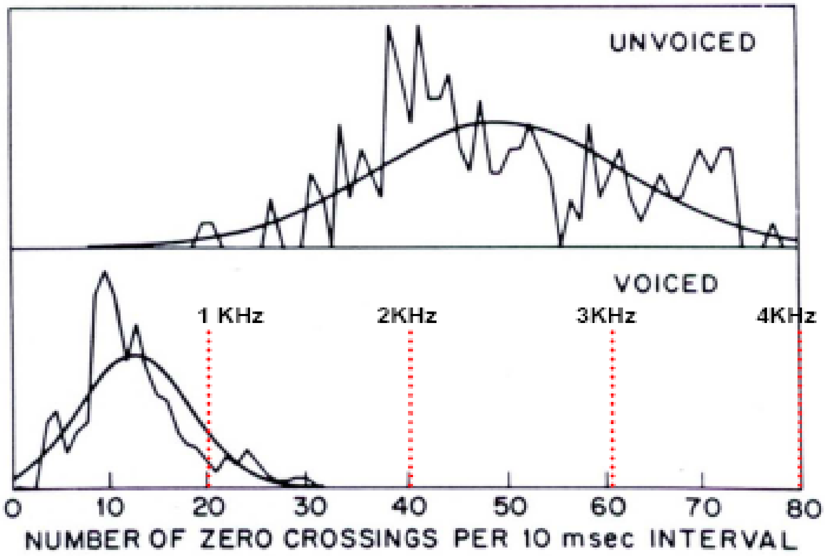


Fig. 4.11 Distribution of zero-crossings for unvoiced and voiced speech.

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

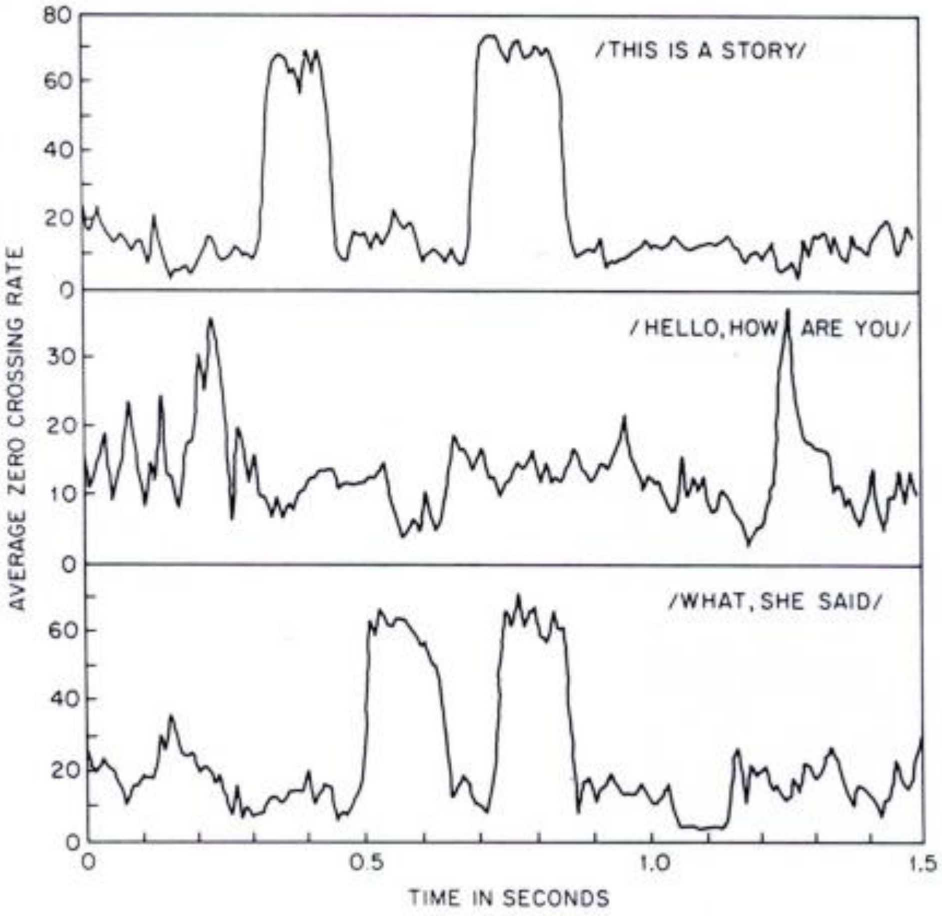


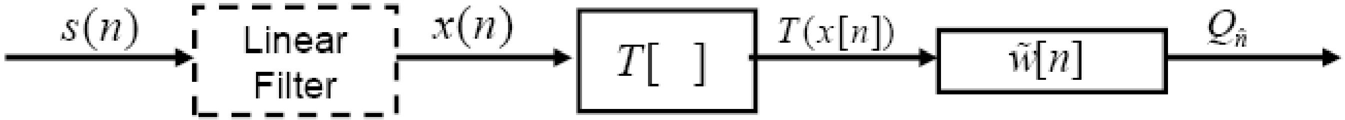
Fig. 4.12 Average zero-crossing rate for three different utterances.

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

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:

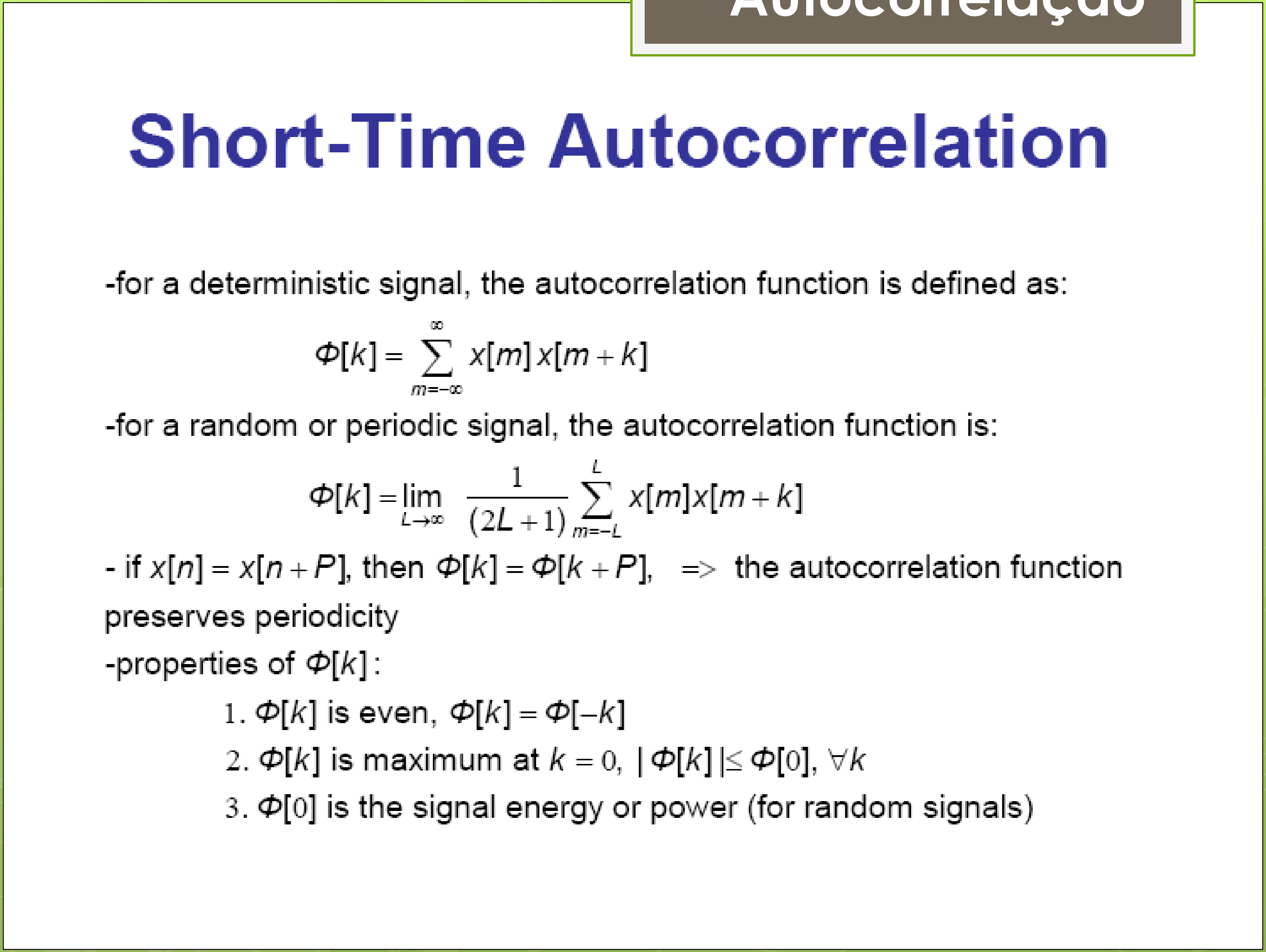
$$M_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} |x[m]|\tilde{w}[\hat{n}-m]$$

3. Zero Crossing Rate:

$$Z_{\hat{n}} = z_1 = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\mathrm{sgn}(x[m]) - \mathrm{sgn}(x[m-1])|\tilde{w}[\hat{n}-m]$$

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

Autocorrelação



Short-Time Autocorrelation

-for a deterministic signal, the autocorrelation function is defined as:

$$\Phi[k] = \sum_{m=-\infty}^{\infty} x[m]\,x[m+k]$$

-for a random or periodic signal, the autocorrelation function is:

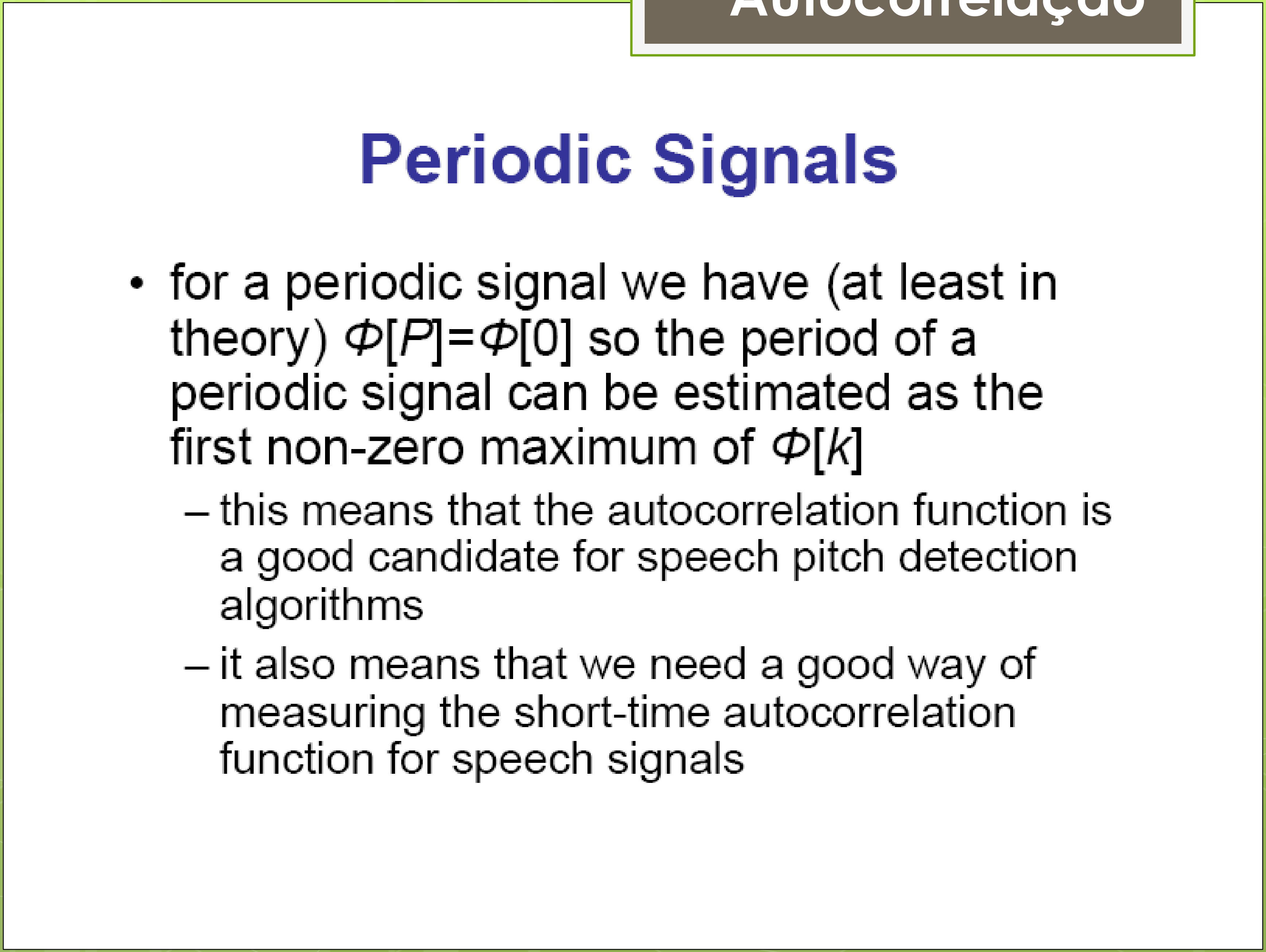
$$\Phi[k] = \lim_{L\to\infty} \frac{1}{(2L+1)} \sum_{m=-L}^{L} x[m]x[m+k]$$

- if $x[n] = x[n+P]$, then $\Phi[k] = \Phi[k+P]$, => the autocorrelation function preserves periodicity

-properties of $\Phi[k]$:

1. $\Phi[k]$ is even, $\Phi[k] = \Phi[-k]$
2. $\Phi[k]$ is maximum at $k = 0$, $|\Phi[k]| \le \Phi[0],\ \forall k$
3. $\Phi[0]$ is the signal energy or power (for random signals)

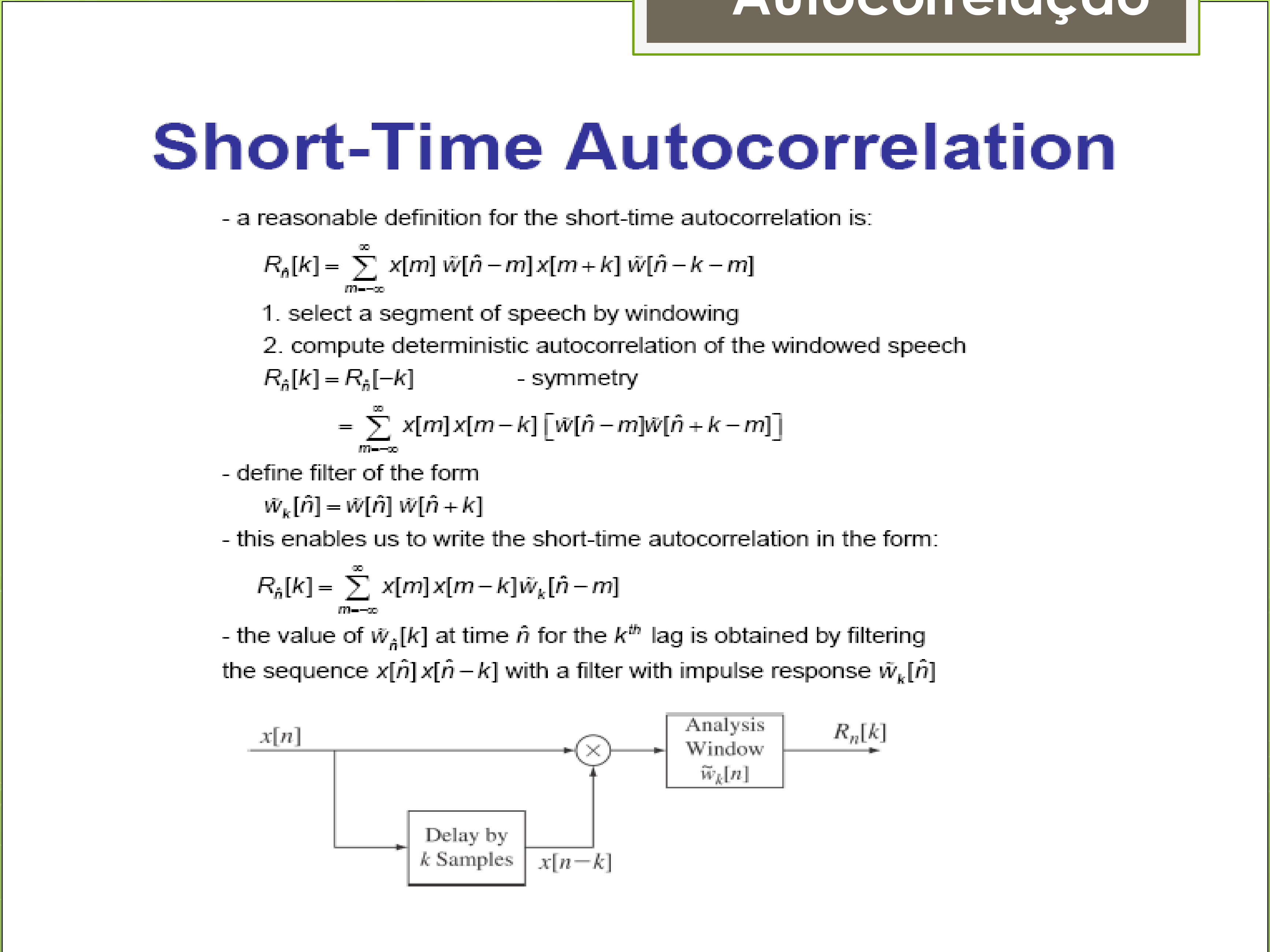
Autocorrelação



Periodic Signals

- for a periodic signal we have (at least in theory) $\Phi[P]=\Phi[0]$ so the period of a periodic signal can be estimated as the first non-zero maximum of $\Phi[k]$
  - this means that the autocorrelation function is a good candidate for speech pitch detection algorithms
  - it also means that we need a good way of measuring the short-time autocorrelation function for speech signals

Autocorrelação



Short-Time Autocorrelation

- a reasonable definition for the short-time autocorrelation is:

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} x[m]\,\tilde{w}[\hat{n}-m]\,x[m+k]\,\tilde{w}[\hat{n}-k-m]$$

1. select a segment of speech by windowing
2. compute deterministic autocorrelation of the windowed speech

$R_{\hat{n}}[k] = R_{\hat{n}}[-k]$ - symmetry

$$= \sum_{m=-\infty}^{\infty} x[m]\,x[m-k]\left[\tilde{w}[\hat{n}-m]\tilde{w}[\hat{n}+k-m]\right]$$

- define filter of the form

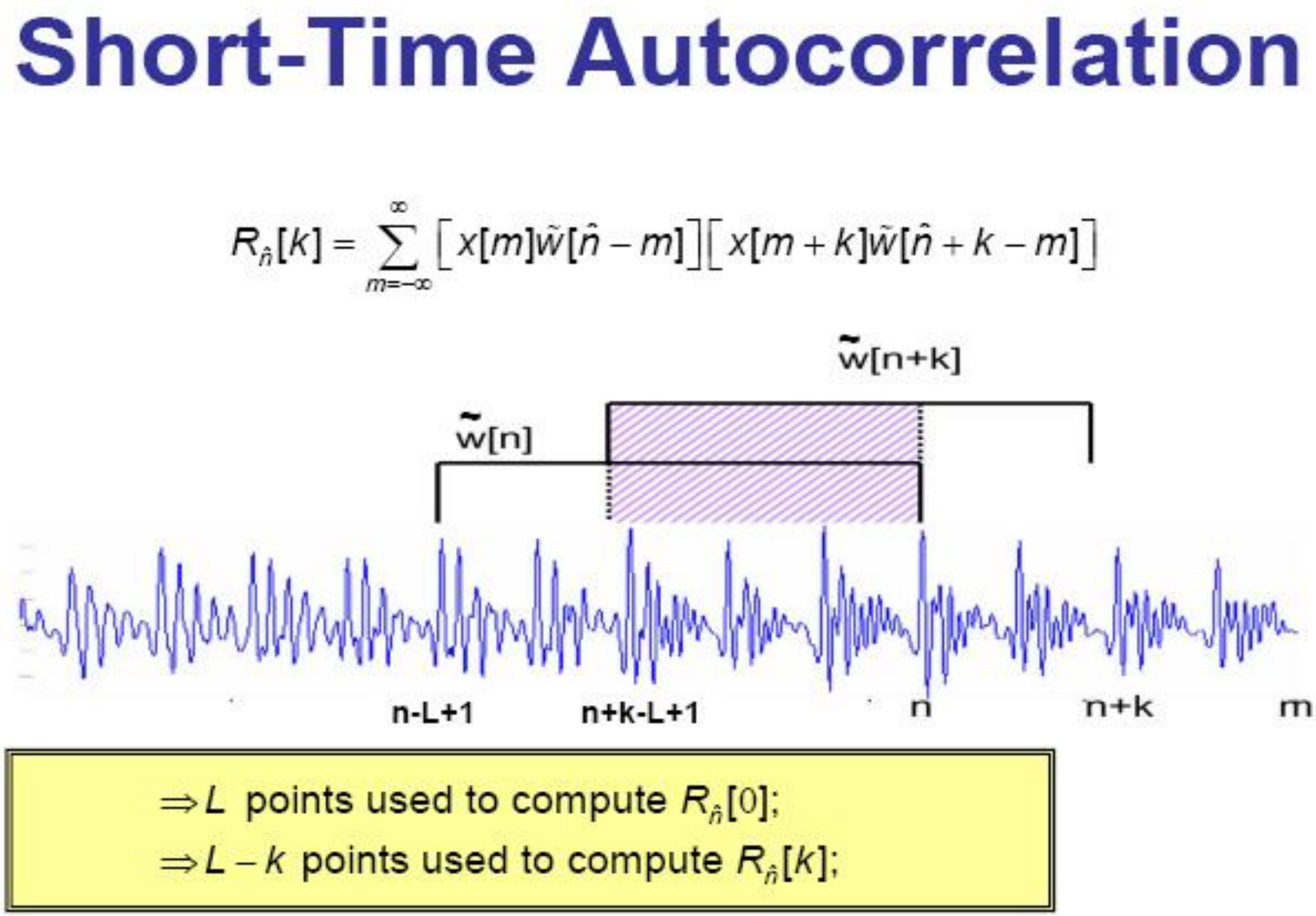
$$\tilde{w}_k[\hat{n}] = \tilde{w}[\hat{n}]\,\tilde{w}[\hat{n}+k]$$

- this enables us to write the short-time autocorrelation in the form:

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} x[m]\,x[m-k]\,\tilde{w}_k[\hat{n}-m]$$

- the value of $\tilde{w}_{\hat{n}}[k]$ at time $\hat{n}$ for the k^{th} lag is obtained by filtering the sequence $x[\hat{n}]\,x[\hat{n}-k]$ with a filter with impulse response $\tilde{w}_k[\hat{n}]$

Autocorrelação



Short-Time Autocorrelation

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} \left[x[m]\tilde{w}[\hat{n}-m]\right]\left[x[m+k]\tilde{w}[\hat{n}+k-m]\right]$$

$\Rightarrow L$ points used to compute $R_{\hat{n}}[0]$;

$\Rightarrow L-k$ points used to compute $R_{\hat{n}}[k]$;

Autocorrelação

Examples of Autocorrelations

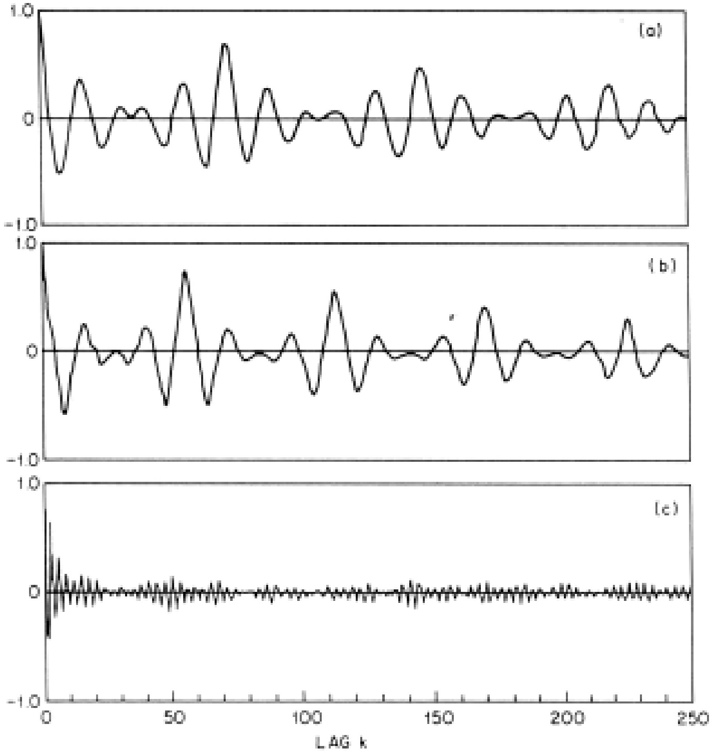


Fig. 4.24 Autocorrelation function for (a) and (b) voiced speech; and (c) unvoiced speech, using a rectangular window with $N = 401$.

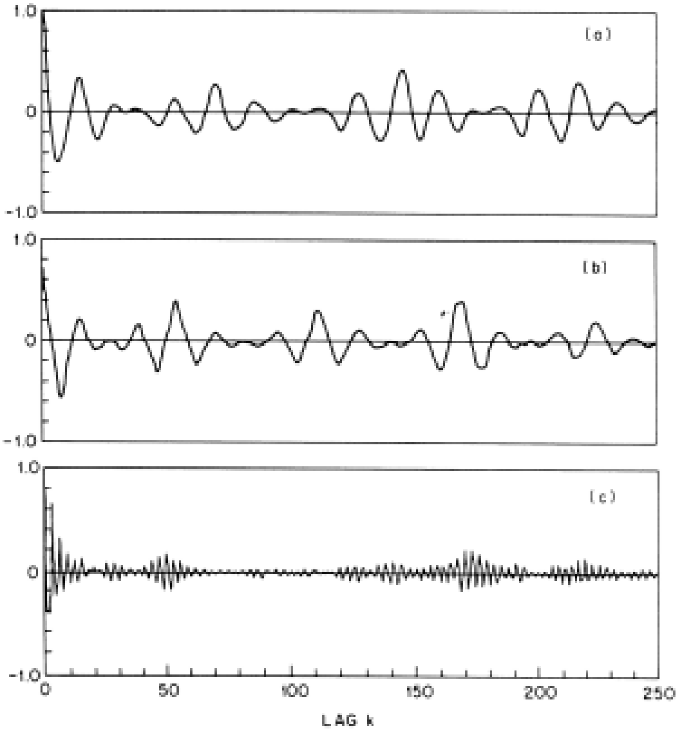


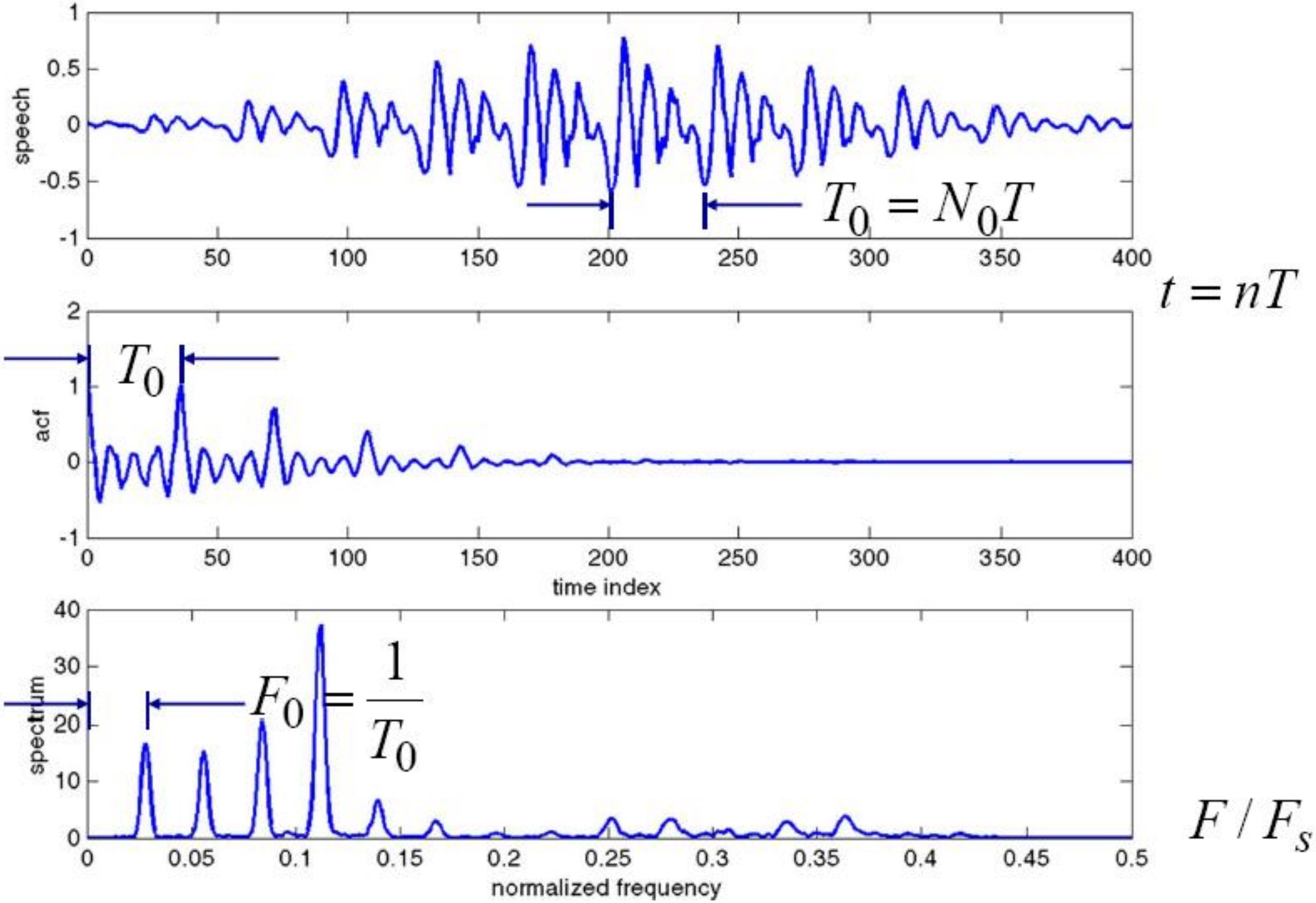
Fig. 4.25 Autocorrelation functions for (a) and (b) voiced speech; and (c) unvoiced speech, using a Hamming window with $N = 401$.

- autocorrelation peaks occur at k=72, 144, ... => 140 Hz pitch
- $\Phi(P)<\Phi(0)$ since windowed speech is not perfectly periodic
- over a 401 sample window (40 msec of signal), pitch period changes occur, so P is not perfectly defined

- much less clear estimates of periodicity since HW tapers signal so strongly, making it look like a non-periodic signal
- no strong peak for unvoiced speech

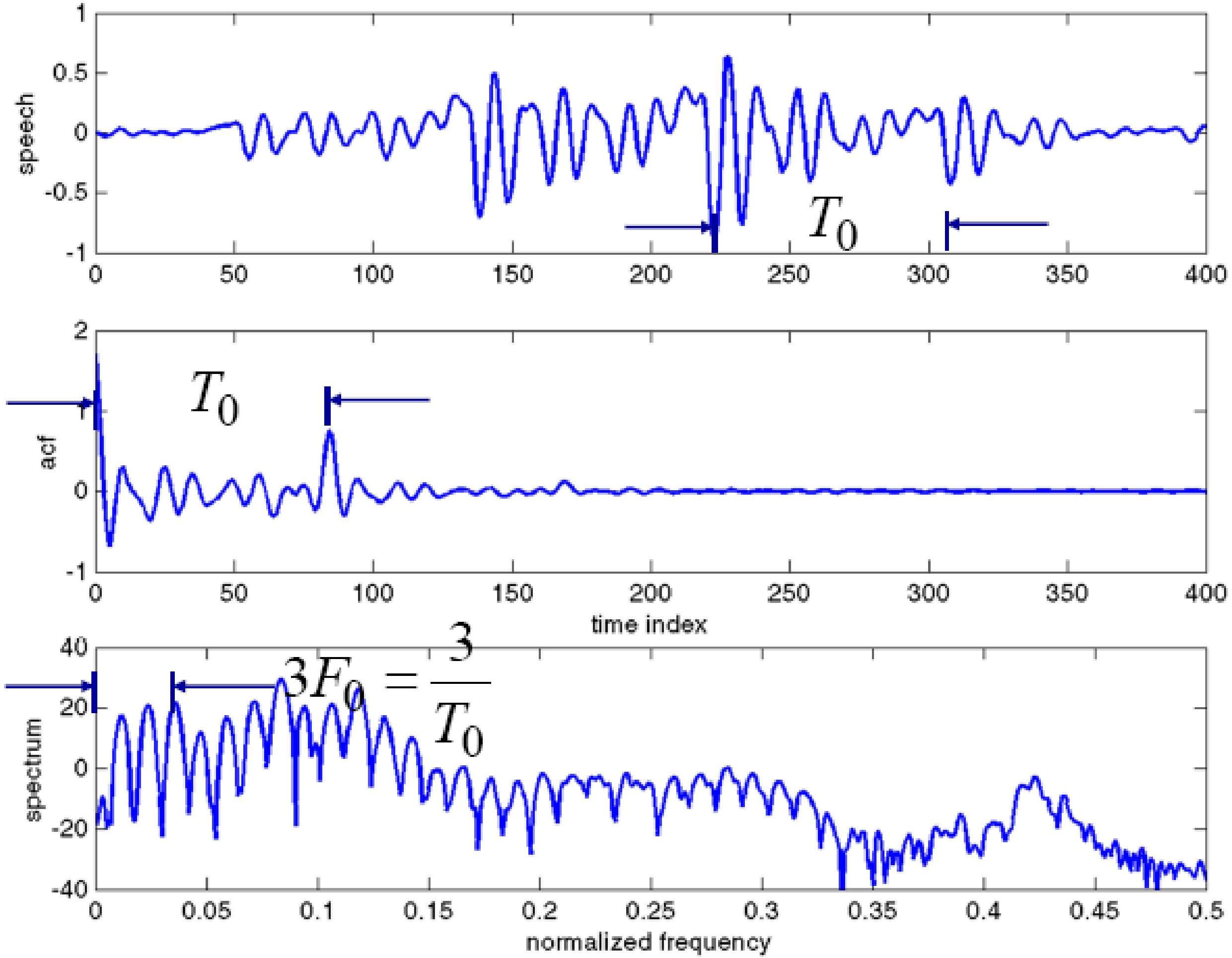
Autocorrelação

Voiced (female) *L=401* (magnitude)

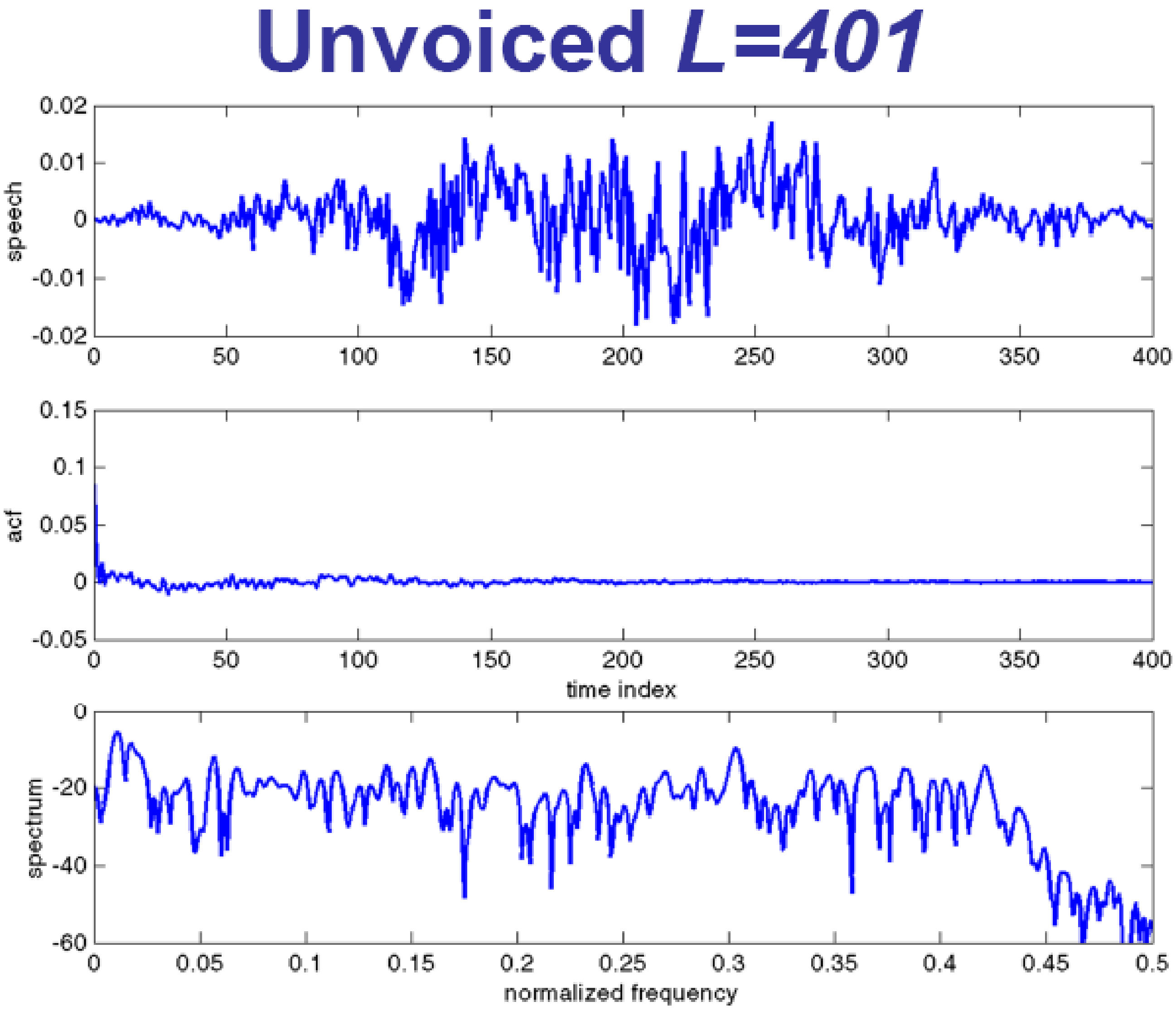


Autocorrelação

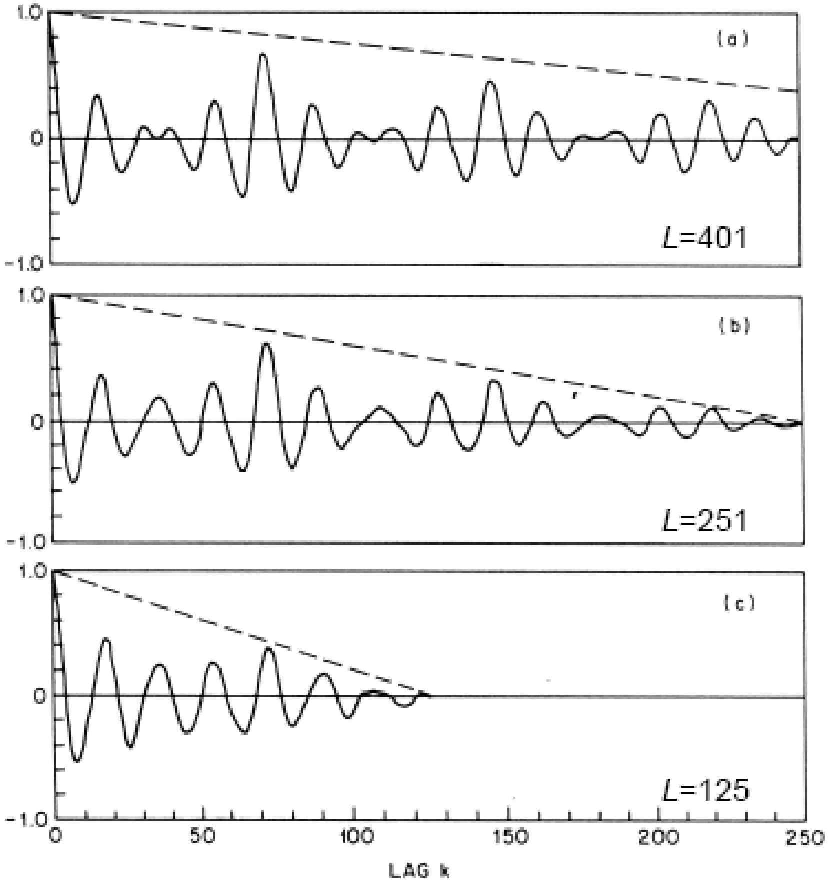
Voiced (male) *L=401*



Autocorrelação



Effects of Window Size



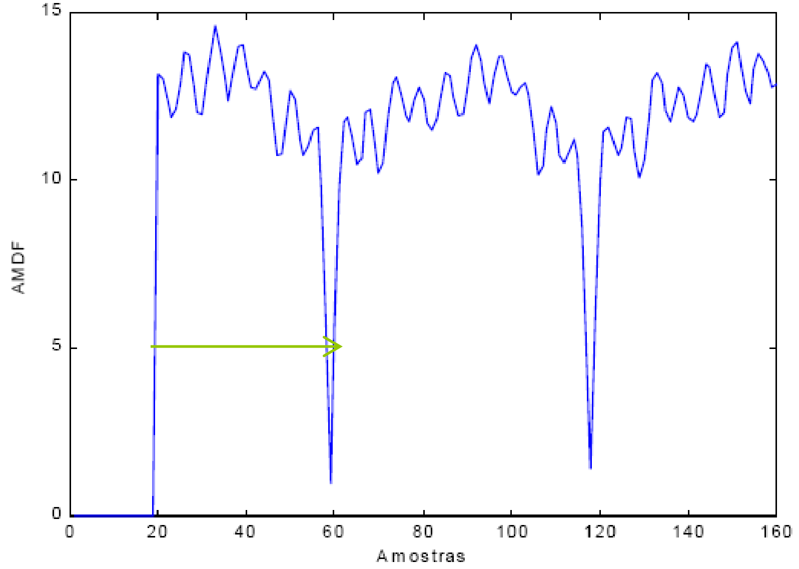
- choice of *L*, window duration
  - small *L* so pitch period almost constant in window
  - large *L* so clear periodicity seen in window
  - as *k* increases, the number of window points decrease, reducing the accuracy and size of $R_n(k)$ for large *k* => have a taper of the type $R(k)=1-k/L$, $|k|<L$ shaping of autocorrelation (this is the autocorrelation of size *L* rectangular window)
- allow *L* to vary with detected pitch periods (so that at least 2 full periods are included)

AMDF - Average Magnitude Difference Function

- Diferença entre o sinal original e o sinal deslocado de τ amostras.

$$AMDF(\tau) = \frac{1}{N}\sum_{j=1}^{k} | s(j) - s(j+\tau) | ,$$

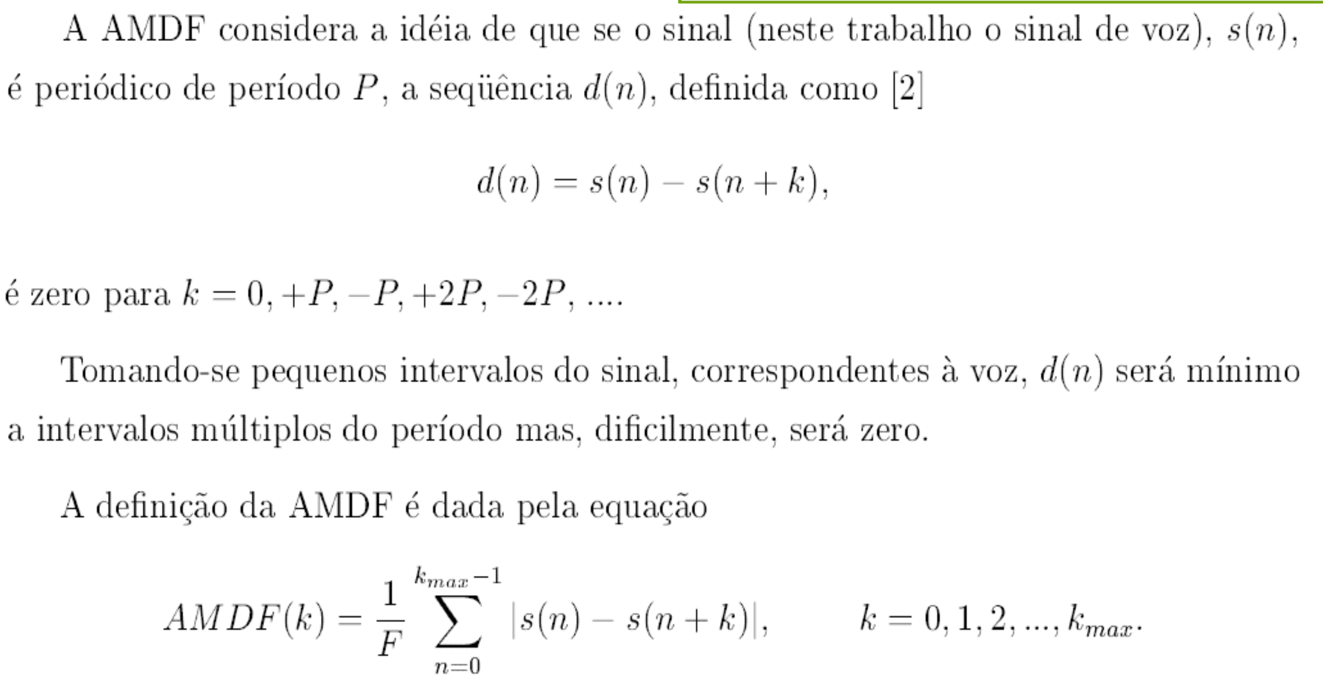
AMDF - Exemplo



AMDF da vogal “e”.

- Para o cálculo do pitch, usa-se a janela retangular, filtra passa-baixas em 800 Hz, para eliminar sinais de alta frequência.
- Identifica se o sinal é vozeado ou não.
- Identifica os valores mínimos da AMDF.

AMDF



A AMDF considera a idéia de que se o sinal (neste trabalho o sinal de voz), $s(n)$, é periódico de período P, a seqüência $d(n)$, definida como [2]

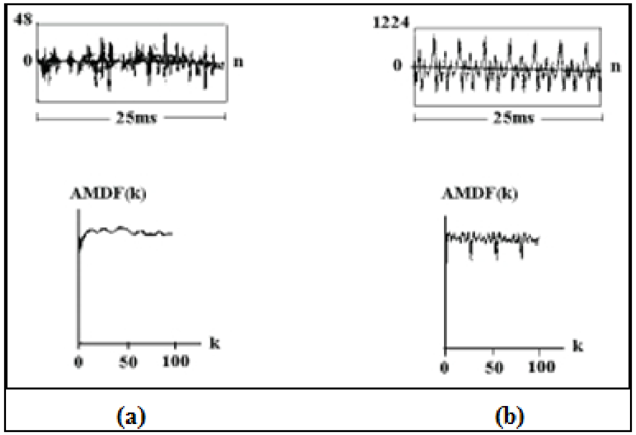
$$d(n) = s(n) - s(n+k),$$

é zero para $k = 0, +P, -P, +2P, -2P, \ldots$.

Tomando-se pequenos intervalos do sinal, correspondentes à voz, $d(n)$ será mínimo a intervalos múltiplos do período mas, dificilmente, será zero.

A definição da AMDF é dada pela equação

$$AMDF(k) = \frac{1}{F} \sum_{n=0}^{k_{max}-1} |s(n) - s(n+k)|, \qquad k = 0, 1, 2, \ldots, k_{max}.$$



AMDF para segmento (a) surdo; (b) sonoro.

O período de pitch será o primeiro mínimo da função AMDF>

Short-Time AMDF

- belief that for periodic signals of period P, the difference function

$$d[n] = x[n] - x[n-k]$$

- will be approximately zero for $k = 0, \pm P, \pm 2P, \ldots$ For realistic speech signals, $d[n]$ will be small at $k = P$--but not zero. Based on this reasoning. the short-time Average Magnitude Difference Function (AMDF) is defined as:

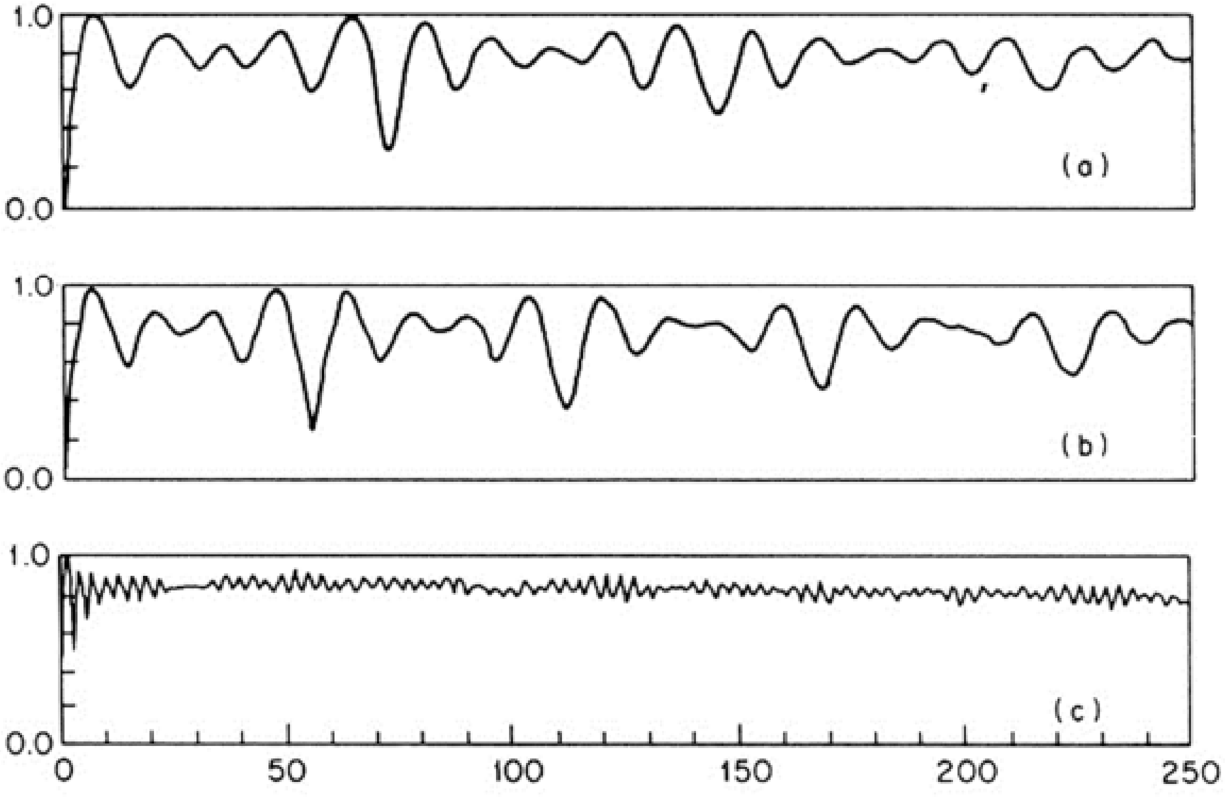
$$\gamma_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} | x[\hat{n}+m]\tilde{w}_1[m] - x[\hat{n}+m-k]\tilde{w}_2[m-k] |$$

- with $\tilde{w}_1[m]$ and $\tilde{w}_2[m]$ being rectangular windows. If both are the same length, then $\gamma_{\hat{n}}[k]$ is similar to the short-time autocorrelation, whereas if $\tilde{w}_2[m]$ is longer than $\tilde{w}_1[m]$, then $\gamma_{\hat{n}}[k]$ is similar to the modified short-time autocorrelation (or covariance) function. In fact it can be shown that

$$\gamma_{\hat{n}}[k] \approx \sqrt{2}\beta[k]\left[\hat{R}_{\hat{n}}[0] - \hat{R}_{\hat{n}}[k]\right]^{1/2}$$

-where $\beta[k]$ varies between 0.6 and 1.0 for different segments of speech.

AMDF for Speech Segments



Referências Bibliográficas Básicas

FECHINE, J. M. *Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística*. Tese de Doutorado. Universidade Federal da Paraíba, 2000.

Kutwak, André B. Análise da codificação LPC para sinais de fala. Universidade Federal do Rio de janeiro, 1999. Disponível em: https://www.lps.ufrj.br/arquivos/0909090c772f.pdf. Data de acesso: 19/03/2013.

Rabiner, Lawrence. *Digital Speech Processing—Lectures 7-8. Time Domain Methods in Speech Processing. Digital Speech Processing Course (Winter 2013).* http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lectures%207-8_winter_2012.pdf

FECHINE, J. M. Notas de aula. Disponível em: http://www.dsc.ufcg.edu.br/~joseana/PDSV.html/Notas de Aula.

http://ensino.univates.br/~chaet/Materiais/Cap%EDtulo_11_Bloch.pdf