Digital Speech Processing—
Lectures 7-8

Time Domain Methods in Speech Processing
General Synthesis Model

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

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

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

- **Signal Processing**
  - Representation of speech
    - Speech or music \( A(x,t) \)
    - Formants
    - Reflection coefficients
    - Voiced-unvoiced-silence
    - Pitch
    - Sounds of language
    - Speaker identification
    - Emotions

- **Time domain processing** => direct operations on the speech waveform
- **Frequency domain processing** => direct operations on a spectral representation of the signal

- **System**
  - \( x[n] \)
  - Zero crossing rate
  - Level crossing rate
  - Energy
  - Autocorrelation

- **Simple processing**
- Enables various types of feature estimation
Basics

- 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

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

Frame 2

Frame 3

Frame 4

Frame 5
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 $3R, 3R + 1, \ldots, 3R + L - 1$
• 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 \( f[m], m = 0, 1, 2, \cdots \) where \( n \) is the time index and \( m \) is the frame index.
Frames and Windows

\[ w[560 - m] \quad w[1280 - m] \]

\[
F_s = 16,000 \text{ samples/second}
\]
\[
L = 641 \text{ samples (equivalent to 40 msec frame (window) length)}
\]
\[
R = 240 \text{ samples (equivalent to 15 msec frame (window) shift)}
\]

Frame rate of 66.7 frames/second
Short-Time Processing

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

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

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

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

**Diagram:**
- Input: \( x[n] \)
- Linear or non-linear transformation: \( T(\ ) \)
- Output: \( \tilde{w}[n] \)
- Output: \( Q_{\hat{n}} \)

\( T(x[n]) \) is a window sequence (usually finite length)
**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 \( \hat{n} \)

\[
T(x) = x^2
\]

\[
\tilde{w}[n] = \begin{cases} 1 & 0 \leq n \leq N - 1 \\ 0 & \text{otherwise} \end{cases}
\]
Computation of Short-Time Energy

- **window jumps/slides across sequence of squared values**, selecting interval for processing

- what happens to $E_{\tilde{n}}$ as sequence jumps by 2,4,8,...,$L$ samples ($E_{\tilde{n}}$ is a lowpass function—so it can be decimated without lost of information; why is $E_{\tilde{n}}$ lowpass?)

- effects of decimation depend on $L$; if $L$ is small, then $E_{\tilde{n}}$ is a lot more variable than if $L$ is large (window bandwidth changes with $L$!)

*Fig. 4.2 Illustration of the computation of short-time energy.*
Effects of Window

\[ Q_{\hat{n}} = T(x[n]) * \tilde{w}[n] \bigg|_{n=\hat{n}} \]
\[ = x'[n] * \tilde{w}[n] \bigg|_{n=\hat{n}} \]

- \( \tilde{w}[n] \) serves as a lowpass filter on \( T(x[n]) \) which often has a lot of high frequencies (most non-linearities introduce significant high frequency energy—think of what \( (x[n] \cdot x[n]) \) does in frequency).

- Often we extend the definition of \( Q_{\hat{n}} \) to include a pre-filtering term so that \( x[n] \) itself is filtered to a region of interest.
Short-Time Energy

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

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

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

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

\[ h[n] = \tilde{w}^2[n] \]  
short time energy

\[ E_{\hat{n}} = E_n \bigg|_{n=\hat{n}} \]

\[ \frac{F_s}{F_s / R} \]
Short-Time Energy Properties

• depends on choice of \( h[n] \), or equivalently, window \( \tilde{w}[n] \)
  – if \( w[n] \) duration very long and constant amplitude
    \((\tilde{w}[n]=1, n=0,1,...,L-1)\), \( E_n \) would not change much over time, and would not reflect the short-time amplitudes of the sounds of the speech
  – very long duration windows correspond to narrowband lowpass filters
  – want \( E_n \) to change at a rate comparable to the changing sounds of the speech \( \Rightarrow \) this is the essential conflict in all speech processing, namely we need short duration window to be responsive to rapid sound changes, but short windows will not provide sufficient averaging to give smooth and reliable energy function
**Windows**

- consider two windows, $\tilde{w}[n]$
  - rectangular window:
    - $h[n]=1, 0\leq n \leq L-1$ and 0 otherwise
  - Hamming window (raised cosine window):
    - $h[n]=0.54-0.46 \cos(2\pi n/(L-1)), 0\leq n \leq L-1$ 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

\[ L = 21 \text{ samples} \]
Window Frequency Responses

- rectangular window

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

- first zero occurs at \( f = F_s / L = 1 / (LT) \) (or \( \Omega = (2\pi) / (LT) \)) => nominal cutoff frequency of the equivalent “lowpass” filter

- Hamming window

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

- can decompose Hamming Window FR into combination of three terms
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 $\Rightarrow$ 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,

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

- 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

\[
h[n] = \alpha^{n-1}u[n-1] = \alpha^{n-1} \quad n \geq 1
\]

\[
= 0 \quad n < 1
\]

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

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

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

\[
G[n] = \frac{G_0}{\sigma[n]}
\]
Recursive Short-Time Energy

\[
\sigma^2[n] = \alpha \cdot \sigma^2[n-1] + x^2[n-1](1 - \alpha)
\]
Recursive Short-Time Energy

(a) Exponential Window; $\alpha = 0.9$

(b) Discrete-Time Fourier Transform

log magnitude in dB

normalized frequency $\omega / \pi$
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) $x(n) \cdot G(n)$. 
Use of Short-Time Energy for AGC

\[ \alpha = 0.9 \]

\[ \alpha = 0.99 \]
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 $M_n \sim \text{square root (dynamic range of } E_n)$ => 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

/ What She Said / -- Rectangular Window, $E_{\hat{n}}$

/ What She Said / -- Rectangular Window, $M_{\hat{n}}$

Short-Time Energy

Short-Time Magnitude

Time in Seconds

$L=51$

$L=101$

$L=201$

$L=401$

$L=101$

$L=201$

$L=401$
Short Time Energy and Magnitude—Hamming Window
Other Lowpass Windows

- can replace RW or HW with any lowpass filter
- 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.,

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

\[
    H(z) = \frac{1}{1 - az^{-1}} \quad |z| > |a|
\]
Other Lowpass Windows

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

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

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

• 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

\[ z_1 = (2) \text{ crossings/cycle} \times \left( \frac{F_0}{F_S} \right) \text{ cycles/sample} \]

\[ z_1 = \frac{2F_0}{F_S} \text{ crossings/sample} \text{ (i.e., } z_1 \text{ proportional to } F_0 \text{) } \]

\[ z_M = M \left( \frac{2F_0}{F_S} \right) \text{ crossings}/(M \text{ samples}) \]
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 \( z_{100} = 2 / 100 \times 100 = 2 \) 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 \( z_{100} = 2 / 10 \times 100 = 20 \) 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 \( z_{100} = 2 / 2 \times 100 = 100 \) crossings/10 msec interval
Zero Crossing for Sinusoids

offset: 0.75, 100 Hz sinewave, ZC: 9, offset sinewave, ZC: 8

Offset = 0.75

ZC = 9

ZC = 8
Zero Crossings for Noise

Offset: 0.75, random noise, ZC: 252, offset noise, ZC: 122

Offset = 0.75

ZC = 252

ZC = 122
ZC Rate Definitions

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

\[
sgn(x[n]) = \begin{cases} 
1 & x[n] \geq 0 \\
-1 & x[n] < 0 
\end{cases}
\]

- simple rectangular window:
  \[
  \tilde{w}[n] = \begin{cases} 
1 & 0 \leq n \leq L - 1 \\
0 & \text{otherwise} 
\end{cases}
\]

\[
L_{\text{eff}} = L
\]

Same form for \(Z_{\hat{n}}\) as for \(E_{\hat{n}}\) or \(M_{\hat{n}}\)
ZC Normalization

• The formal definition of $z_n$ is:

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

is interpreted as the number of zero crossings per sample.

• For most practical applications, we need the rate of zero crossings per fixed interval of $M$ samples, which is

$$z_M = z_1 \cdot M = \text{rate of zero crossings per } M \text{ sample interval}$$

Thus, for an interval of $\tau$ sec., corresponding to $M$ samples we get

$$z_M = z_1 \cdot M; \quad M = \tau F_S = \tau / T$$
ZC Normalization

- For a 1000 Hz sinewave as input, using a 40 msec window length ($L$), with various values of sampling rate ($F_S$), we get the following:

<table>
<thead>
<tr>
<th>$F_S$</th>
<th>$L$</th>
<th>$z_1$</th>
<th>$M$</th>
<th>$z_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8000</td>
<td>320</td>
<td>1/4</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>10000</td>
<td>400</td>
<td>1/5</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>16000</td>
<td>640</td>
<td>1/8</td>
<td>160</td>
<td>20</td>
</tr>
</tbody>
</table>

- Thus we see that the normalized (per interval) zero crossing rate, $z_M$, is independent of the sampling rate and can be used as a measure of the dominant energy in a band.
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

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

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

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

- 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

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

Fig. 4.12 Average zero-crossing rate for three different utterances.
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:
\[ Z_\hat{n} = z_1 = \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}(x[m]) - \text{sgn}(x[m-1])|\tilde{w}[\hat{n} - m] \]
   where \( \text{sgn}(x[m]) = 1 \quad x[m] \geq 0 \)
   \[ = -1 \quad x[m] < 0 \]
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:

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

- 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 $k = 0$, $|\Phi[k]| \leq \Phi[0]$, $\forall k$
  3. $\Phi[0]$ is the signal energy or power (for random signals)
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
Short-Time Autocorrelation

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

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

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

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

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

- 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 \( \tilde{w}_k[\hat{n}] \)
Short-Time Autocorrelation

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

\[ R_{\hat{n}}[k] = \sum_{m=0}^{L-1-k} \left[ x[\hat{n} + m] \tilde{w}'[m] \right] \left[ x[\hat{n} + m + k] \tilde{w}'[k + m] \right] \]

⇒ \( L - k \) points used to compute \( R_{\hat{n}}[k] \)
Short-Time Autocorrelation

(a) $x[m]$ and $w[\hat{n} - m]$

(b) $x[\hat{n} + m]w'[m]$

(c) $x[\hat{n} + m + k]w'[m + k]$
Examples of Autocorrelations

- 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
Voiced (female) \( L=401 \) (magnitude)

\[
T_0 = N_0 T
\]

\[
t = nT
\]

\[
F_0 = \frac{1}{T_0}
\]
Voiced (female) $L=401$ (log mag)

$T_0$

$t = nT$

$F_0 = \frac{1}{T_0}$

$F \frac{T}{F_s}$
Voiced (male) $L=401$
Unvoiced $L=401$
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)
Modified Autocorrelation

• want to solve problem of differing number of samples for each different $k$ term in $R_n[k]$, so modify definition as follows:

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

- where $\tilde{w}_1$ is standard $L$-point window, and $\tilde{w}_2$ is extended window of duration $L + K$ samples, where $K$ is the largest lag of interest
- we can rewrite modified autocorrelation as:

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

- where
  $$\hat{w}_1[m] = \tilde{w}_1[-m] \quad \text{and} \quad \hat{w}_2[m] = \tilde{w}_2[-m]$$
- for rectangular windows we choose the following:
  $$\hat{w}_1[m] = 1, \quad 0 \leq m \leq L - 1$$
  $$\hat{w}_2[m] = 1, \quad 0 \leq m \leq L - 1 + K$$

- giving

$$\hat{R}_n[k] = \sum_{m=0}^{L-1} x[\hat{n} + m] x[\hat{n} + m + k], \quad 0 \leq k \leq K$$

- always use $L$ samples in computation of $\hat{R}_n[k] \forall k$
Examples of Modified Autocorrelation

- $\hat{R}_n[k]$ is a cross-correlation, not an auto-correlation
- $\hat{R}_n[k] \neq \hat{R}_n[-k]$
- $\hat{R}_n[k]$ will have a strong peak at $k = P$ for periodic signals and will not fall off for large $k$
Examples of Modified Autocorrelation
Examples of Modified AC

Modified Autocorrelations – fixed value of $L=401$

Modified Autocorrelations – values of $L=401, 251, 125$
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:

  $$\gamma_\hat{n}[k] = \sum_{m=-\infty}^{\infty} |x[\hat{n} + m]w_1[m] - x[\hat{n} + m - k]w_2[m - k]|$$

- with $w_1[m]$ and $w_2[m]$ being rectangular windows. If both are the same
  length, then $\gamma_\hat{n}[k]$ is similar to the short-time autocorrelation, whereas
  if $w_2[m]$ is longer than $w_1[m]$, then $\gamma_\hat{n}[k]$ is similar to the modified short-time
  autocorrelation (or covariance) function. In fact it can be shown that

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

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

(a)

(b)

(c)

LAG k
Summary

• Short-time parameters in the time domain:
  – short-time energy
  – short-time average magnitude
  – short-time zero crossing rate
  – short-time autocorrelation
  – modified short-time autocorrelation
  – Short-time average magnitude difference function