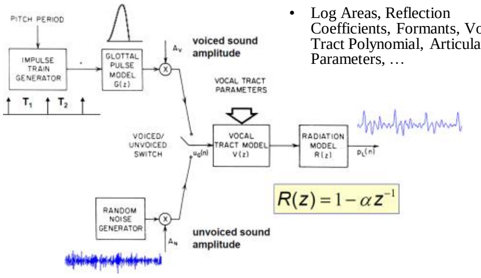
Digital Audio and Speech Processing
(Sayısal Ses ve Konuşma İşleme)

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr
nizamettinaydin@gmail.com
http://www3.yildiz.edu.tr/~naydin

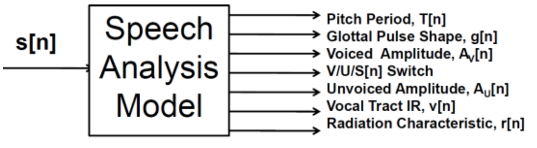
Time Domain Methods in Speech Processing

General Synthesis Model



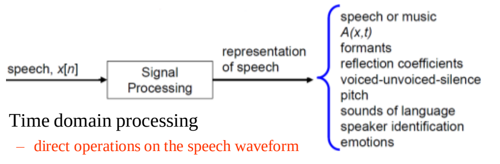
- Log Areas, Reflection Coefficients, Formants, Vocal Tract Polynomial, Articulatory Parameters, …
- Pitch Detection, Voiced/Unvoiced/Silence Detection, Gain Estimation, Vocal Tract Parameter Estimation, Glottal Pulse Shape, Radiation Model

General Analysis Model

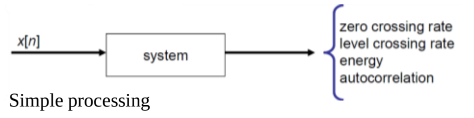


- All analysis parameters are time-varying at rates commensurate with information in the parameters;
- We need algorithms for estimating the analysis parameters and their variations over time

Overview

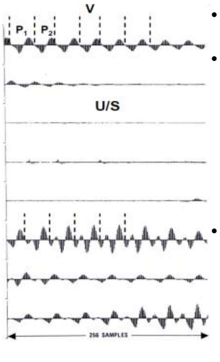


- Time domain processing
  - direct operations on the speech waveform
- Frequency domain processing
  - direct operations on a spectral representation of the signal



- 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

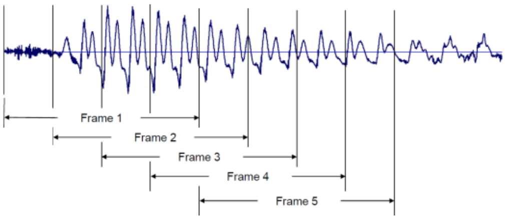
Fundamental Assumptions

- 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

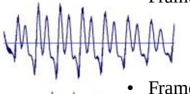
Frame-by-Frame Processing in Successive Windows



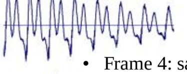
- 75% frame overlap, frame length=L, frame shift=R=L/4
  - Frame1={x[0],x[1],…,x[L-1]}
  - Frame2={x[R],x[R+1],…,x[R+L-1]}
  - Frame3={x[2R],x[2R+1],…,x[2R+L-1]}
  - …

Frame-by-Frame Processing in Successive Windows

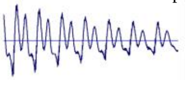
- Frame 1: samples $0,1,..., L-1$



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

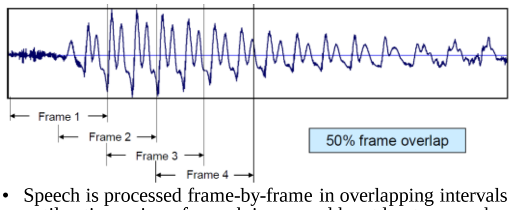


- Frame 3: samples $2R,2R+1,...,2R+L-1$



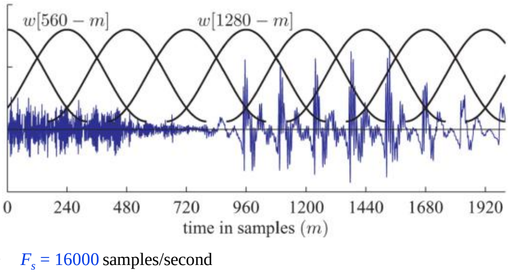
- Frame 4: samples $3R,3R+1,...,3R+L-1$

Frame-by-Frame Processing in Successive Windows



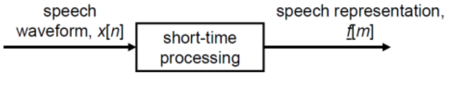
- Speech is processed frame-by-frame in overlapping intervals until entire region of speech is covered by at least one such frame
  - Results of analysis of individual frames used to derive model parameters invsome manner
  - Representation goes from time sample $x[n],\ n=\cdots,0,1,2,\cdots$ to parameter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

Frames and Windows



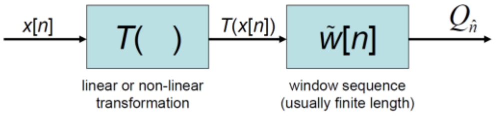
- F_s = 16000 samples/second
- L = 641 samples (equivalent to 40 msec frame (window) length)
- R = 240 samples (equivalent to 15 msec frame (window) shift)
- Frame rate of 66.7 frames/second

Short-Time Processing



- $x[n]$ = samples at 8000/sec rate
  - e.g., 2 seconds of 4 kHz bandlimited speech,
    - $x[n],\ 0 \le n \le 16000$
- $\vec{f}[m] = \{f_1[m], f_2[m], \cdots, f_L[m]\}$ = 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

- 

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

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

Short-Time Energy

- The long term definition of signal energy

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

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

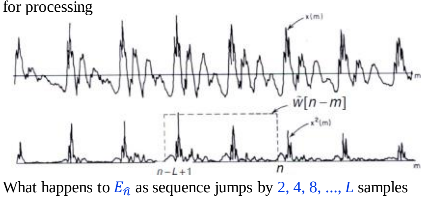
$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} x^2[m] = x^2[\hat{n} - L + 1] + \cdots + x^2[\hat{n}]$

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

$$T(x) = x^2$$
$$\tilde{w}[n] = 1 \quad 0 \le n \le L-1$$
$$= 0 \quad \text{otherwise}$$

Computation of Short-Time Energy

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

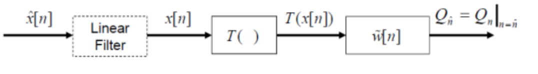


- What happens to $E_{\hat{n}}$ as sequence jumps by 2, 4, 8, ..., L samples
  - $E_{\hat{n}}$ is a lowpass function
    - so it can be decimated without loss of information;
      - why is $E_{\hat{n}}$ lowpass?
- Effects of decimation depend on L;
  - if L is small, then $E_{\hat{n}}$ is a lot more variable than if L is large
    - window bandwidth changes with L

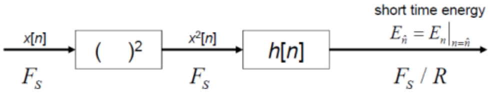
Effects of Window

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

- $\tilde{w}[n]$ 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:
  - sum or squares of portion of signal

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

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

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

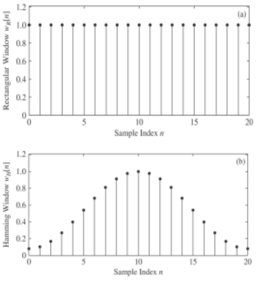
$$h[n] = \tilde{w}^2[n]$$

Short-Time Energy Properties

- Depends on choice of $h[n]$, or equivalently, window $\tilde{w}[n]$
  - If $\tilde{w}[n]$ duration is very long and constant amplitude ($\tilde{w}[n]$=1, n=0,1,...,L-1), $E_{\hat{n}}$ 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
  - We want $E_{\hat{n}}$ to change at a rate comparable to the changing sounds of the speech
    - 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 (RW):
- $h[n] = 1$, $0 \leq n \leq L-1$ and 0 otherwise
- gives equal weight to all L samples in the window $(n,...,n-L+1)$

– Hamming window (HW, raised cosine window):
- $h[n] = 0.54 - 0.46\cos(2\pi n/(L-1))$, $0 \leq n \leq L-1$ and 0 otherwise
- gives most weight to middle samples and tapers off strongly at the beginning and the end of the window

L = 21 samples

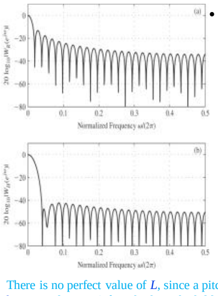
19

Window Frequency Responses

- Rectangular window
  - $H(e^{j\omega T}) = \frac{\sin(\frac{\omega LT}{2})}{\sin(\frac{\omega T}{2})} e^{-j\omega T \frac{L-1}{2}}$
  - 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
  - $\tilde{w}_H[n] = 0.54\tilde{w}_R[n] - 0.46 * \cos(\frac{2\pi n}{L-1})\, \tilde{w}_R[n]$
  - can decompose Hamming Window FR into combination of three terms

20

Frequency Responses of RW and HW



- 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
      - otherwise severe fluctuations will occur in E_n,
    - but smaller than a sound duration
      - otherwise E_n will not adequately reflect the changes in the speech signal

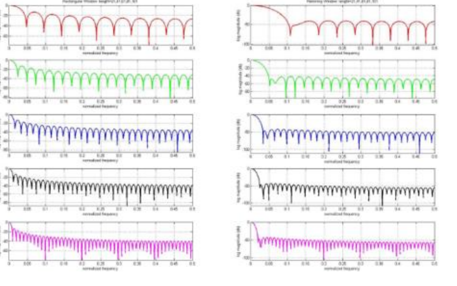
There is no perfect value of L, since a pitch period can be as short as 20 samples (500 Hz at a 10 kHz sampling rate) for a high pitch child or female, and up to 250 samples (40 Hz pitch at a 10 kHz sampling rate) for a low pitch male; a compromise value of L on the order of 100-200 samples for a 10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

21

Window Frequency Responses

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

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



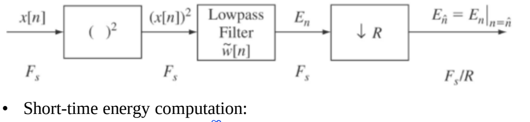
22

Voiced/unvoiced detection

- Methods to distinguish between voiced and unvoiced segments
  - Short-time energy
  - Short-time magnitude
  - Short-time zero crossing

23

Short-Time Energy



- Short-time energy computation:

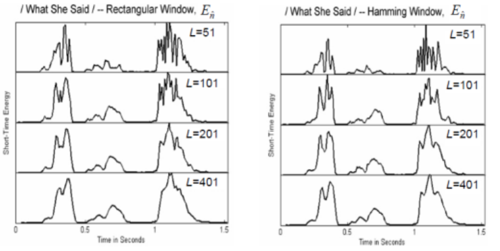
$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} [x[m]\tilde{w}[\hat{n}-m]]^2$$

- For L-point rectangular window,
  $\tilde{w}[m] = 1, \qquad m = 0, 1, \cdots, L-1$
- Giving

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

24

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

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

  - 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

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

- For the index $n-1$ we have

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

- Thus giving the relationship

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

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

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

Recursive Short-Time Energy

$$\sigma^2[n] = x^2[n] * h[n]$$
$$h[n] = (1-\alpha)\alpha^{n-1}u[n-1]$$
$$\sigma^2[z] = X^2[z] \times H[z]$$

$$H(z) = \sum_{n=-\infty}^{\infty} h[n]z^{-n} = \sum_{n=-\infty}^{\infty} (1-\alpha)\alpha^{n-1}u[n-1]z^{-n}$$

$$= \sum_{n=1}^{\infty} (1-\alpha)\alpha^{n-1}z^{-n}$$

$$m = n-1$$

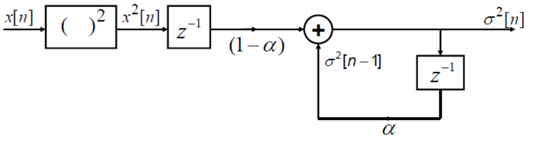
$$H(z) = \sum_{m=0}^{\infty} (1-\alpha)\alpha^m z^{-(m+1)} = \sum_{m=0}^{\infty} (1-\alpha)z^{-1}\alpha^m z^{-m}$$

$$= (1-\alpha)z^{-1}\sum_{m=0}^{\infty} \alpha^m z^{-m} = (1-\alpha)z^{-1}\frac{1}{1-\alpha z^{-1}} = \frac{\sigma^2[z]}{X^2[z]}$$

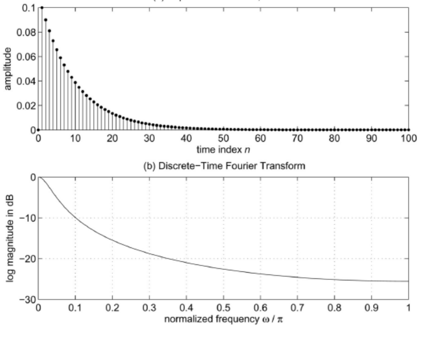
$$\sigma^2[n] = \alpha\sigma^2[n-1](1-\alpha)x^2[n-1]$$

Recursive Short-Time Energy

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

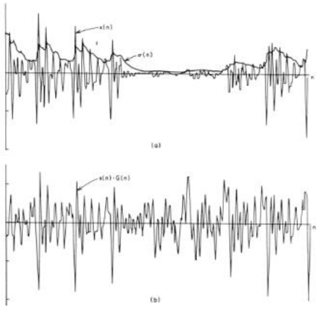


Recursive Short-Time Energy



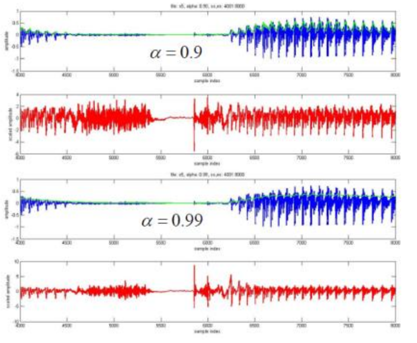
Use of Short-Time Energy for AGC

- Variance estimate, $\alpha = 0.9$



31

Use of Short-Time Energy for AGC



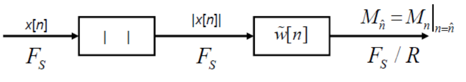
32

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)

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

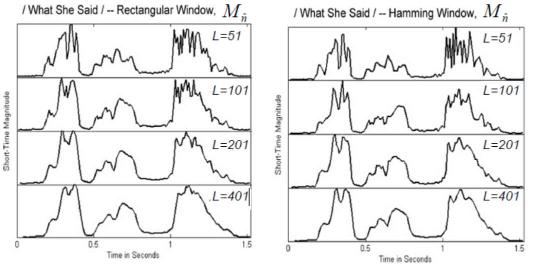
  - Weighted sum of magnitudes, rather than weighted sum of squares



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

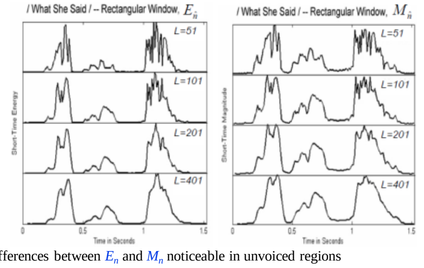
33

Short-Time Magnitudes



34

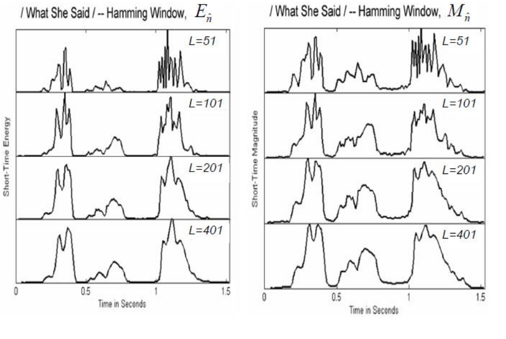
Short Time Energy and Magnitude-Rectangular Window



- 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

35

Short Time Energy and Magnitude-Hamming Window



36

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 R samples for efficiency with no loss of information (what should R be?)
- Can even use an infinite duration window if its z-transform is a rational function, i.e.,

$$h[n] = a^n, \; n \geq 0, \; 0 < a < 1$$
$$h[n] = 0, \qquad n < 0$$
$$H(z) = \frac{1}{1 - az^{-1}} \qquad |z| > |a|$$

Other Lowpass Windows

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

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

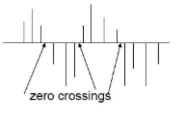
$$M_n = aM_{n-1} + (1-a)x[n] \quad \text{(short-time magnitude)}$$

- 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

- Energy for voiced speech tends to concentrate below 3 KHz, whereas for unvoiced speech energy is found at higher frequencies
- Since high frequencies imply high zero-crossing rates, one can discriminate both types of segments from their zero-crossing rate
  - As before, split the speech signal $x[n]$ into short blocks (i.e., 10-20 ms)
  - Calculate the zero-crossing rate within each block
  - Determine a maximum likelihood threshold

Short-Time Average ZC Rate



- zero crossing
  - successive samples have different algebraic signs
- The rate at which zero crossings occur is a simple measure of the frequency content of a signal.
- This is particularly true of narrowband signals.
- For example, a sinusoidal signal of frequency F_o, sampled at a rate F_s, has F_s / F_o samples per cycle of the sine wave.
- Each cycle has two zero crossings so that the long-time average rate of zero-crossings is

  $Z = 2 F_s / F_o$, crossings/sample
- The average zero-crossing rate gives a reasonable way to estimate the frequencyof a sine wave.

Short-Time Average ZC Rate

- Speech signals are broadband signals and the interpretation of average zero-crossing rate is therefore much less precise.
  - However, rough estimates of spectral properties can be obtained using a representation based on the shorttime average zero-crossing rate.
- ZC Rate can be defined as

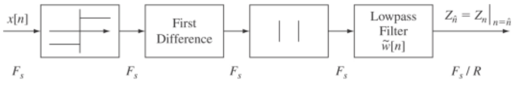
$$Z_{\hat{n}} = \frac{1}{2L_{\text{eff}}} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}\{x[m] - \text{sgn}\{x[m-1]\}| \tilde{w}[\hat{n}-m]$$

where $\text{sgn}\{x[n]\} = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$

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

Short-Time Average ZC Rate

- The short-time average zero-crossing rate has the same general properties as the short-time energy and the short time average magnitude.



- The computation of $Z_{\hat{n}}$ is done by checking samples in pairs to determine where the zero-crossings occur and then the average is computed over L consecutive samples.

Zero Crossing Normalization

- The formal definition of $Z_{\hat{n}}$ is:

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

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

$z_M = z_1 M =$ rate of zero crossings per M sample interval

43

Zero Crossing Normalization

- Thus, for an interval of τ sec., corresponding to M samples we get

$$z_M = z_1 M; \quad M = \tau F_s = \frac{\tau}{T_s}$$

- Zero crossings/10 msec interval as a function of sampling rate:
  - $F_s = 10000$ Hz; $T = 100\ \mu$sec; $\tau = 10$ msec; $M = 100$ samples
  - $F_s = 8000$ Hz; $T = 100\ \mu$sec; $\tau = 10$ msec; $M = 80$ samples
  - $F_s = 16000$ Hz; $T = 100\ \mu$sec; $\tau = 10$ msec; $M = 160$ samples

44

Zero Crossing Normalization

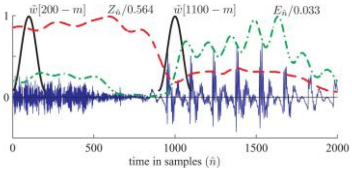
- For a 1000 Hz sinewave as input, using a 40 msec window length (L), with various values of sampling rate (F_s), we get the following:

F_s	L	z_s	M	z_M
8000	320	1/4	80	20
10000	400	1/5	100	20
16000	640	1/8	160	20

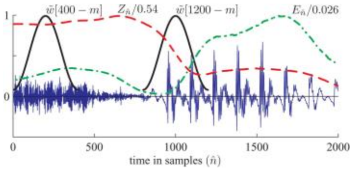
- Thus we see that the normalized (per interval) zero crossing rate, z_M, is independent of the sampling rate and can be used as a measure of the dominant energy in a band.

45

Zero Crossing and Energy Computation



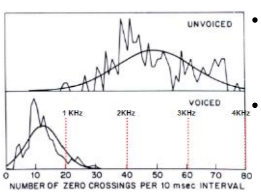
- Hamming window with duration L=201 samples (12.5 msec at F_s=16 kHz)



- Hamming window with duration L=401 samples (25 msec at F_s=16 kHz)

46

Zero Crossing Rate Distributions

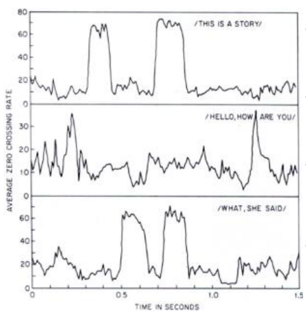


- Unvoiced Speech:
  - The dominant energy component is at about 2.5 kHz
- Voiced Speech:
  - The dominant energy component is at about 700 Hz
- For voiced speech, energy is mainly below 1.5 kHz
- For unvoiced speech, energy is mainly above 1.5 kHz
- Mean ZC rate for unvoiced speech is 49 per 10 msec interval
- Mean ZC rate for voiced speech is 14 per 10 msec interval

47

Zero Crossing Rates for Speech

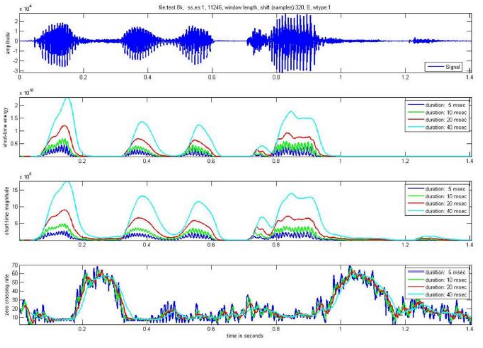
- Some examples of average ZC rate measurements:



- The duration of the averaging window is 15 msec
  - 150 samples at 10 kHz sampling rate
- The output is computed 100 times/sec
  - window moved in steps of 100 samples.
- Note that just as in the case of short-time energy and average magnitude, the short-time average ZC rate can be sampled at a very low rate.
- Although the ZC rate varies considerably, the voiced and unvoiced regions are quite prominent

48

Short-Time Energy, Magnitude, ZC



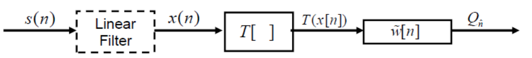
49

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

50

Summary of Simple Time Domain Measures



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

- 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
- Magnitude:

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

- Zero Crossing Rate:

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

51

Short-Time Autocorrelation

- The autocorrelation function of a discrete-time deterministic signal:

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

- For a random or periodic signal:

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

- If $x[n] = x[n+P]$, then $\emptyset[k] = \emptyset[k+P]$
  - the autocorrelation function preserves periodicity
- Properties of $\emptyset[k]$:
  - $\emptyset[k]$ is even, $\emptyset[k] = \emptyset[-k]$
  - $\emptyset[k]$ is maximum at $k=0$, $|\emptyset[k]| \le \emptyset[0], \forall k$
  - $\emptyset[0]$ is the signal energy or power (for random signals)

52

Periodic Signals

- For a periodic signal we have (at least in theory) $\emptyset[P] = \emptyset[0]$so the period of a periodic signal can be estimated as the first non-zero maximum of $\emptyset[k]$
  - 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 signa

53

Short-Time Autocorrelation

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

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

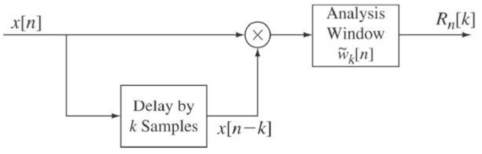
  - Select a segment of speech by windowing
  - Compute deterministic autocorrelation of the windowed speech

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

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

54

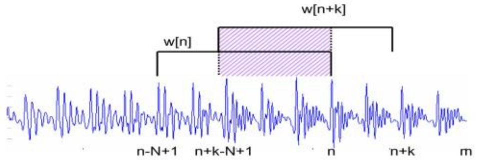
Short-Time Autocorrelation



- Define filter of the form :
  $\tilde{w}_k = \tilde{w}[\hat{n}]\tilde{w}[\hat{n}+k]$
- This enables us to write the short-time autocorrelation in the form:
  $$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} x[m]x[m-k]\tilde{w}[\hat{n}-m]$$
- 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 $\tilde{w}_k[\hat{n}]$

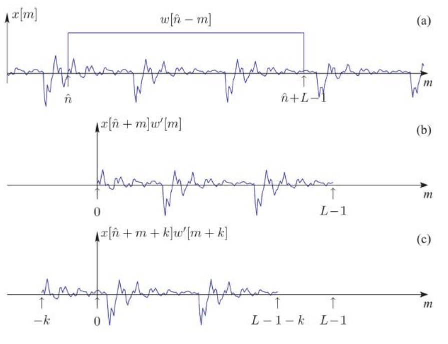
Short-Time Autocorrelation

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



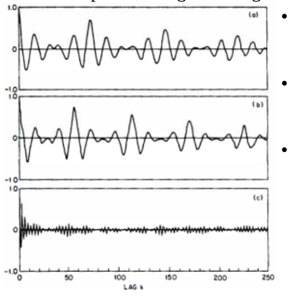
- L points used to compute $R_{\hat{n}}[0]$
- L-1 points used to compute $R_{\hat{n}}[k]$

Short-Time Autocorrelation



Examples of Autocorrelations

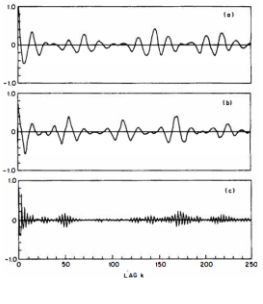
- Autocorrelation function for (a) and (b) voiced speech, and (c) unvoiced speech, using a rectangular window with L = 401



- 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

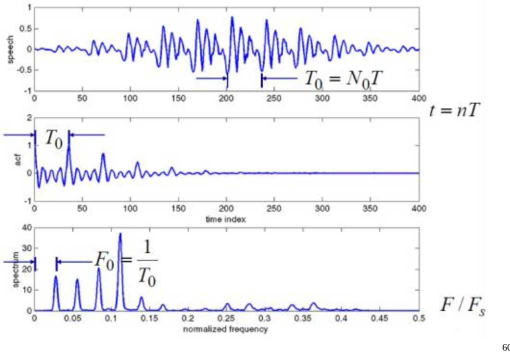
Examples of Autocorrelations

- Autocorrelation function for (a) and (b) voiced speech, and (c) unvoiced speech, using a Hanning window with L = 401

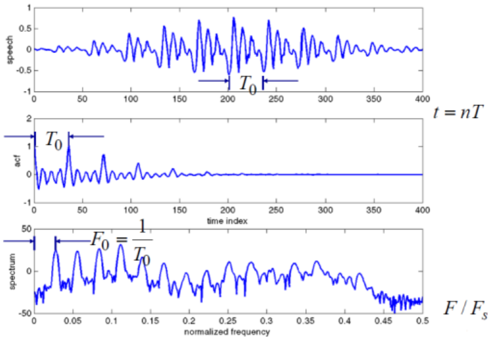


- Much less clear estimates of periodicity since HW tapers signal so strongly, making it look like a non-periodic signal
- No strong peak for unvoiced speech

Voiced (female) L=401 (magnitude)

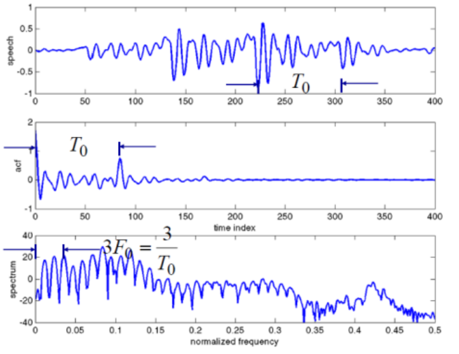


Voiced (female) L=401 (log mag)



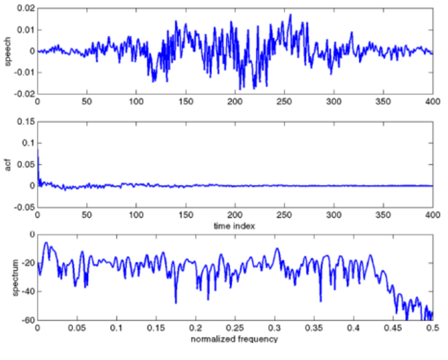
61

Voiced (male) L=401



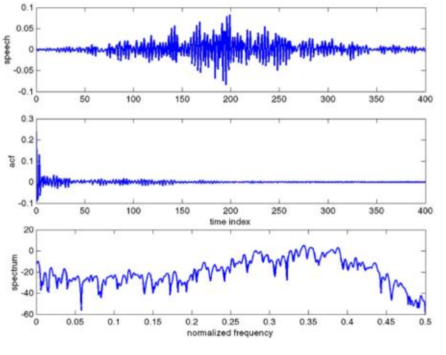
62

Unvoiced L=401



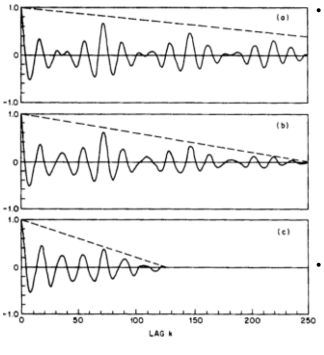
63

Unvoiced L=401



64

Effects of Window Size



- An important issue is how L should be chosen to give a good indication of periodicity.
  - small L
    - pitch period almost constant in window
  - large L
    - 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

65

Modified Autocorrelation

- Another approach is to allow the window length to adapt to match the expected pitch period.
- The modified short-time autocorrelation function is defined as

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

  - where $\tilde{w}_1$: standard L-point window, $\tilde{w}_2$: extended window of duration $L+K$ samples, where K is the largest lag of interest

$$\tilde{w}_1[m] = \tilde{w}_1[-m] \quad \text{and} \quad \tilde{w}_2[m] = \tilde{w}_2[-m]$$

  - For rectangular windows we choose the following:

$$\tilde{w}_1[m] = 1, \qquad 0 \le m \le L-1$$
$$\tilde{w}_2[m] = 1, \qquad 0 \le m \le L-1+K$$

  - Giving

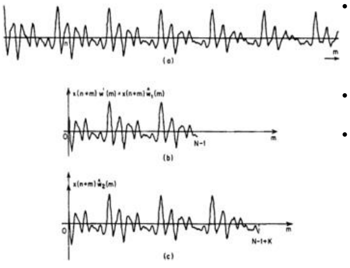
$$\hat{R}_{\hat{n}}[k] = \sum_{m=0}^{L-1} x[\hat{n}+m]\,x[\hat{n}+m+k], \qquad 0 \le k \le K$$

  - Always use L samples in computation of $\hat{R}_{\hat{n}}[k]\ \forall k$

66

Examples of Modified Autocorrelation

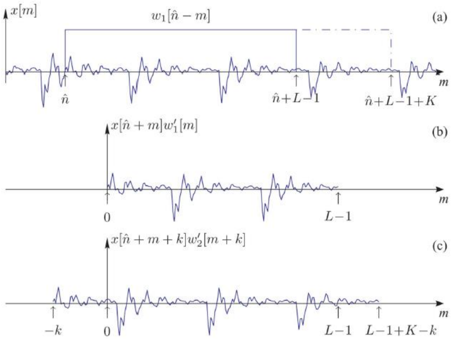
- The cross-correlation (not autocorrelation) function for the two different finite length segments of speech, $x[\hat{n}+m]\tilde{w}_1[m]$ and $x[\hat{n}+m]\tilde{w}_2[m]$.



- Thus $\hat{R}_{\hat{n}}[k]$ has the properties of a cross-correlation function, not an autocorrelation function.
- For example, $\hat{R}_{\hat{n}}[k] \neq \hat{R}_{\hat{n}}[k]$.
- Nevertheless, $\hat{R}_{\hat{n}}[k]$ will display peaks at multiples of the period of a periodic signal and it will not display a fall-off in amplitude at large values of k.

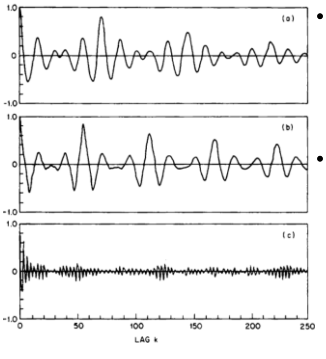
67

Examples of Modified Autocorrelation



68

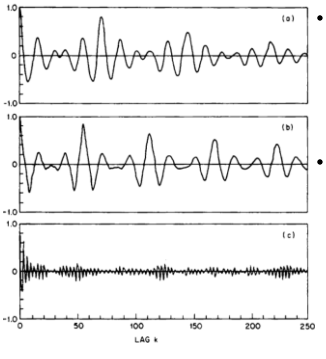
Examples of Modified Autocorrelation



- The modified autocorrelation functions corresponding to the examples of Figure in slide 58.
- Because for $L = 401$ the effects of waveform variation dominate the tapering effect in Figure in slide 58, the two figures look much alike.

69

Examples of Modified Autocorrelation



- A comparison with the Figure in the slide 59 shows that the difference is more apparent for smaller values of L.
- It is clear that the peaks are less than the $k = 0$ peak only because of deviations from periodicity over the interval n to $n+L-1+K$.

70

Short-Time Average Magnitude Difference Function (AMDF)

- Belief that for periodic signals of period P, the difference function
  $$d[n] = x[n] - x[n-k]$$
  will be approximately zero for $k = 0, \pm P, \pm 2P, \cdots$
  - For realistic speech signals, $d[n]$ will be small at $k = P$, but not zero.
- Based on this reasoning, the short-time 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 rectengular windows.

71

Short-Time Average Magnitude Difference Function (AMDF)

- If both windows are the same length, then $\gamma_{\hat{n}}[k]$ is similar to the short-time autocorrelation
- 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,
  - but does not change rapidly with k for a particular speech segment

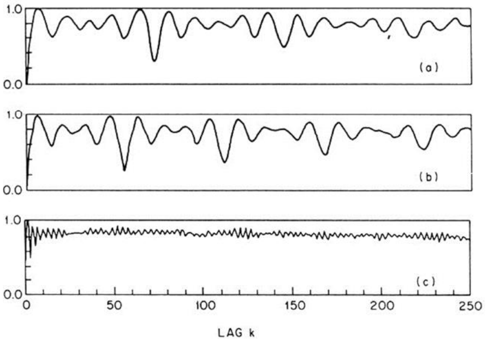
72

Short-Time Average Magnitude Difference Function (AMDF)

- Implemented with subtraction , addition, and absolute value operations,
  - in contrast to addition and multiplication operations for the autocorrelation function.
- With floating point arithmetic, where multiplies and adds take approximately the same time,
  - about the same time is required for either method with the same window length.
- However, for special purpose hardware, or with fixed point arithmetic, the AMDF appears to have the advantage.
  - In this case multiplies usually are more time consuming and furthermore either scaling or a double precision accumulator is required to hold the sum of lagged products.
- For this reason the AMDF function has been used in numerous real-time speech processing systems.

73

AMDF for Speech Segments



74

Summary

- Short-time parameters in the time domain:
  - Short-time energy

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

  - Short-time average magnitude

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

  - Short-time zero crossing rate

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

  - Short-time autocorrelation

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

  - Modified short-time autocorrelation

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

  - Short-time average magnitude difference function

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

75