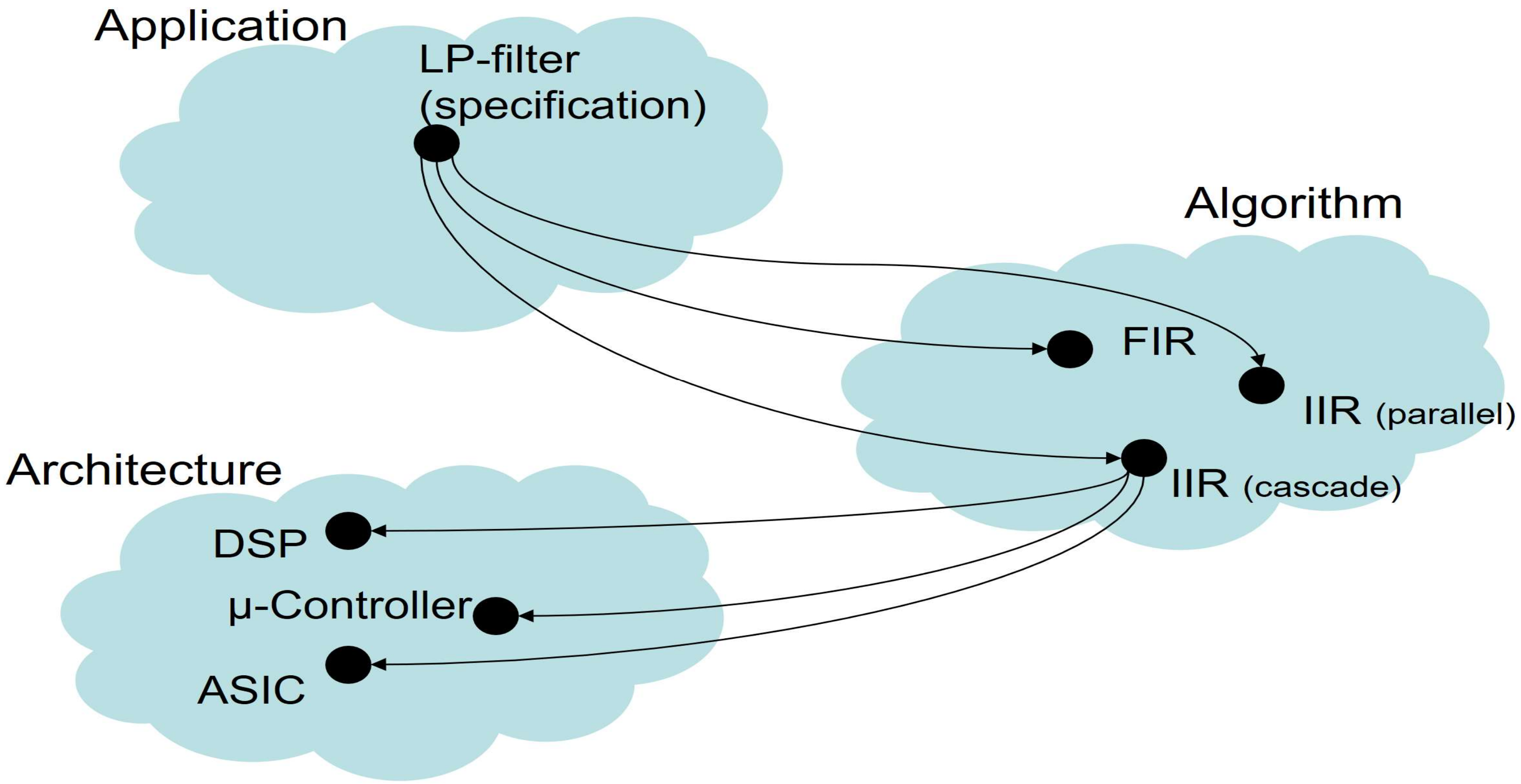
IEE 1711: Applied Signal Processing

Professor Muhammad Mahtab Alam (muhammad.alam@taltech.ee)

Lab Instructor: Julia Berdnikova



Outline

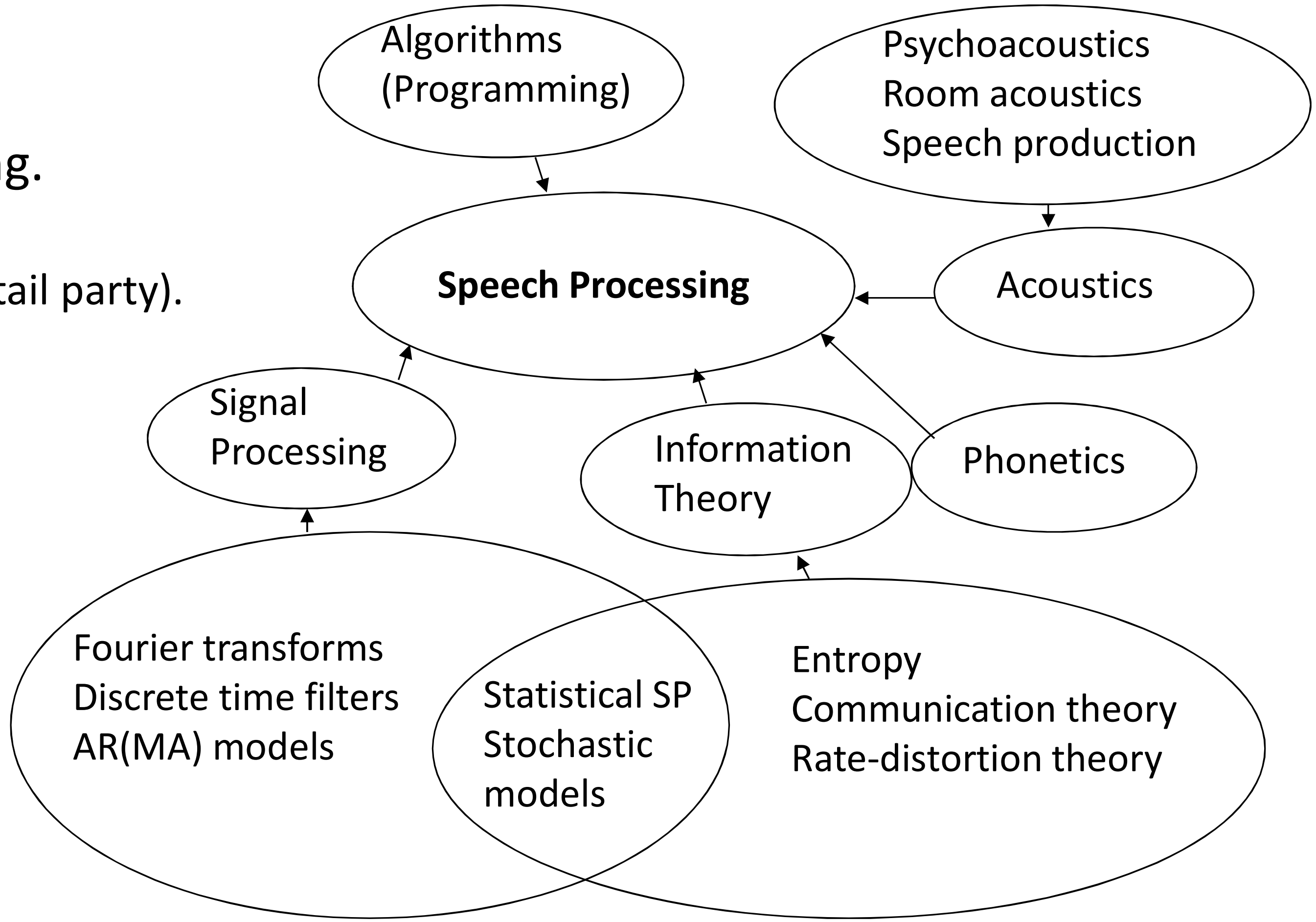
- Lecture 1: Applications of Digital Signal Processing
  - Followup
- Lecture 2: Speech Signal Processing
  - Basic Construct of the Speech Signal
  - Speech Waveforms Classifications
  - Time-Domain Methods
  - Zero Crossing
  - Speech Detection
  - Pitch of the Speech Signal
- Summary

Speech Signal Processing (SSP)

- Speech technology plays a critical role in various markets, including call centers, and mobile and consumer communication.
- The speech chain consists of several technologies that include speech recognition, speech synthesis, language understanding, dialog management, and language generation.
- Speech coding and synthesis are essential for lowering transmission bandwidth and reducing storage requirements.

Applications of Speech Processing

- Speech enhancement:
  - Microphone array processing.
    - Beamforming.
    - Blind signal separation (cocktail party).
  - Echo cancellation.
    - The LMS algorithm.
  - Noise suppression.
    - Spectral subtraction.
    - The Wiener filter.



Algorithmic Domains of Speech Signal Processing

- Analysis of speech signals:
  - Fourier analysis; spectrogram
  - Autocorrelation; pitch estimation
  - Linear prediction; compression, recognition
  - Cepstral analysis; pitch estimation, enhancement
- Speech compression.
  - Scalar quantization (PCM, DPCM).
  - (Transform Coding.)
  - Vector quantization.
  - Speech coders: Linear Predictive Vocoder, Mixed/Code Eecitation Linear Predictive etc.
- Statistical modeling of speech.
  - Gaussian mixtures; speaker identification.
  - Hidden Markov models; speech recognition.

Phoneme Hierarchybv bgp+äüüüü::::...*L'Ä_O
L.

Speech Production

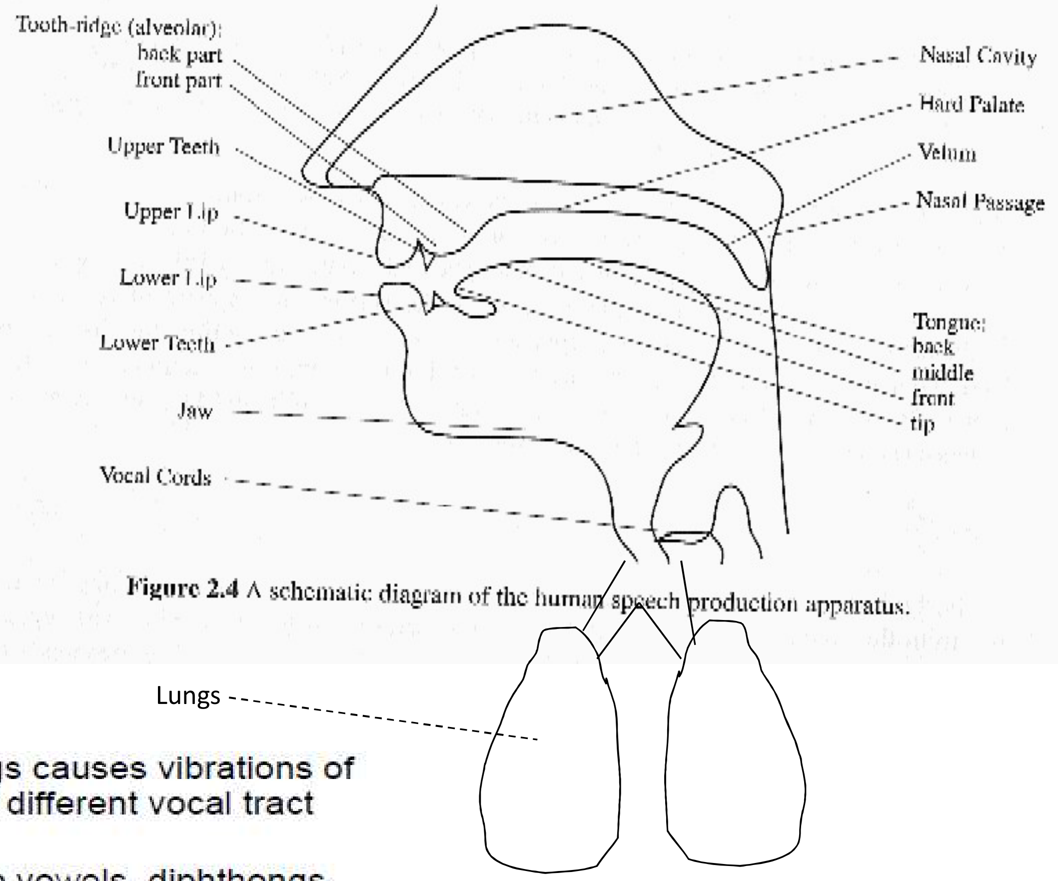
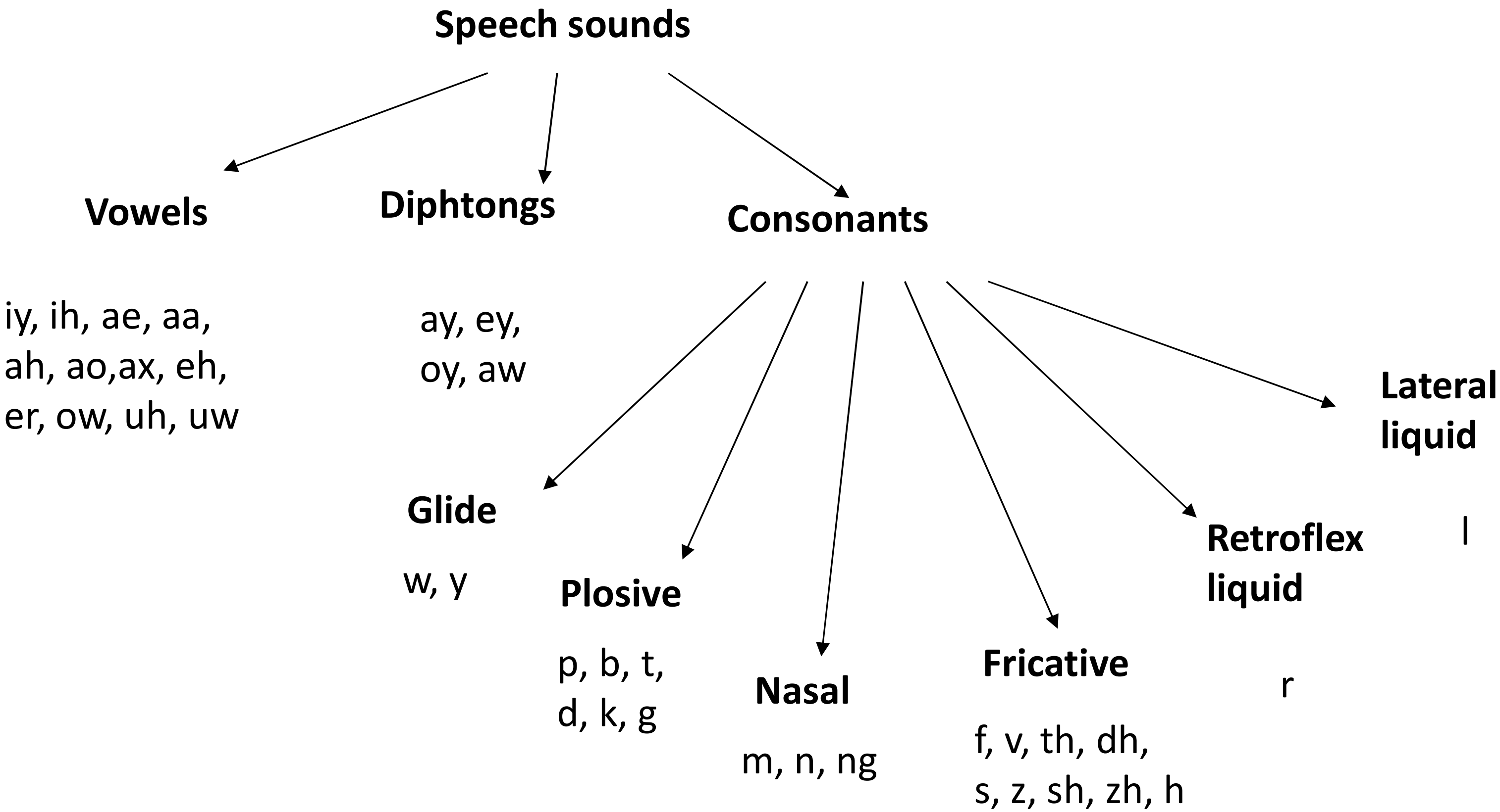


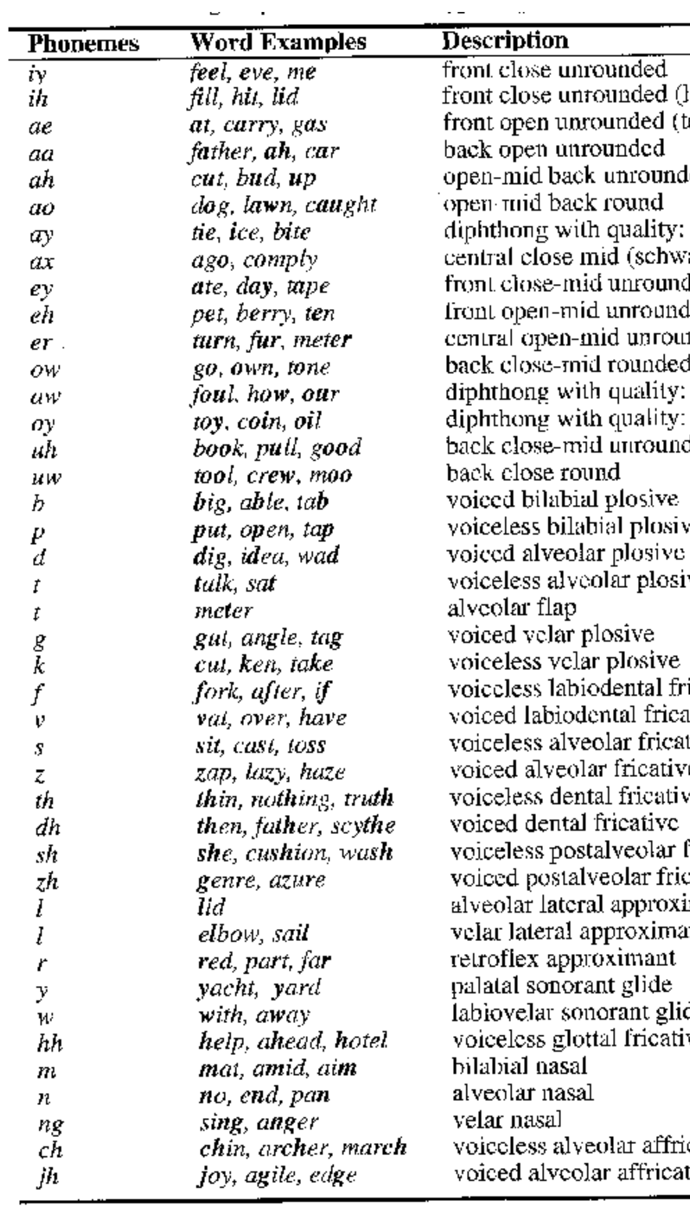
Figure 2.4 A schematic diagram of the human speech production apparatus.

- Speech is generated when air flowing from the lungs causes vibrations of the vocal cords. Different sounds are generated for different vocal tract shapes.
- There are 40-50 English phonemes, partitioned into vowels, diphthongs, nasals, stops, affricates, fricatives, and approximates.
- Voiced sounds are more periodic in nature and have higher energy than unvoiced sounds. All vowels, diphthongs, nasals and approximates are voiced.

Speech Sounds - Phoneme Hierarchy

- Coarse classification with *phonemes*. Language dependent. About 50 in English.
- A *phone* is the acoustic realization of a phoneme.
- *Allophones* are context dependent phonemes.



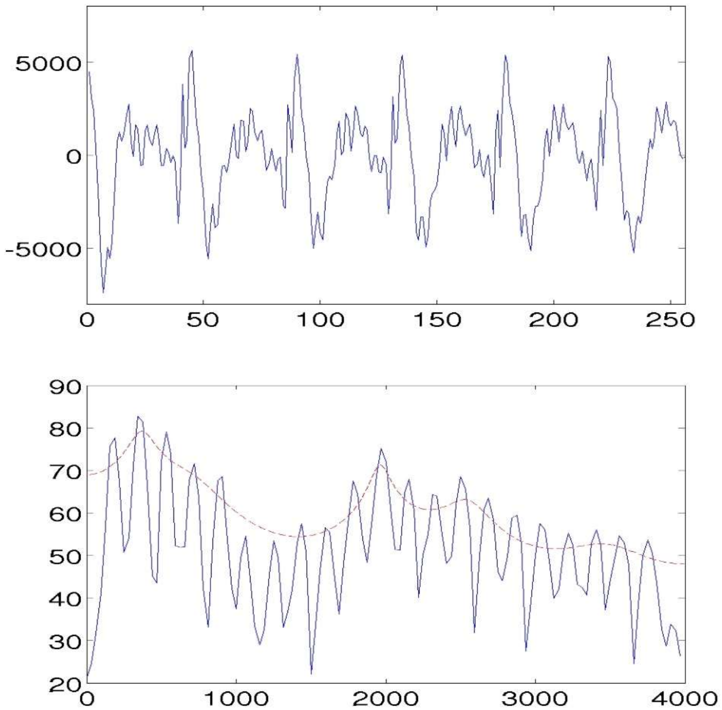


Phonemes	Word Examples	Description
iy	*feel, eve, me*	front close unrounded
ih	*fill, hit, lid*	front close unrounded (l
ae	*at, carry, gas*	front open unrounded (t
aa	*father, ah, car*	back open unrounded
ah	*cut, bud, up*	open-mid back unround
ao	*dog, lawn, caught*	open-mid back round
ay	*tie, ice, bite*	diphthong with quality:
ax	*ago, comply*	central close mid (schwa
ey	*ate, day, tape*	front close-mid unround
eh	*pet, berry, ten*	front open-mid unround
er	*turn, fur, meter*	central open-mid unrou
ow	*go, own, tone*	back close-mid rounded
aw	*foul, how, our*	diphthong with quality:
oy	*toy, coin, oil*	diphthong with quality:
uh	*book, pull, good*	back close-mid unround
uw	*tool, crew, moo*	back close round
b	*big, able, tab*	voiced bilabial plosive
p	*put, open, tap*	voiceless bilabial plosiv
d	*dig, idea, wad*	voiced alveolar plosive
t	*talk, sat*	voiceless alveolar plosi
t	*meter*	alveolar flap
g	*gut, angle, tag*	voiced velar plosive
k	*cut, ken, take*	voiceless velar plosive
f	*fork, after, if*	voiceless labiodental fri
v	*vat, over, have*	voiced labiodental frica
s	*sit, cast, toss*	voiceless alveolar fricat
z	*zap, lazy, haze*	voiced alveolar fricativ
th	*thin, nothing, truth*	voiceless dental fricativ
dh	*then, father, scythe*	voiced dental fricative
sh	*she, cushion, wash*	voiceless postalveolar f
zh	*genre, azure*	voiced postalveolar fric
l	*lid*	alveolar lateral approxi
l	*elbow, sail*	velar lateral approxima
r	*red, part, far*	retroflex approximant
y	*yacht, yard*	palatal sonorant glide
w	*with, away*	labiovelar sonorant glic
hh	*help, ahead, hotel*	voiceless glottal fricativ
m	*mat, amid, aim*	bilabial nasal
n	*no, end, pan*	alveolar nasal
ng	*sing, anger*	velar nasal
ch	*chin, archer, march*	voiceless alveolar affric
jh	*joy, agile, edge*	voiced alveolar affricat

Speech Waveform Characteristics

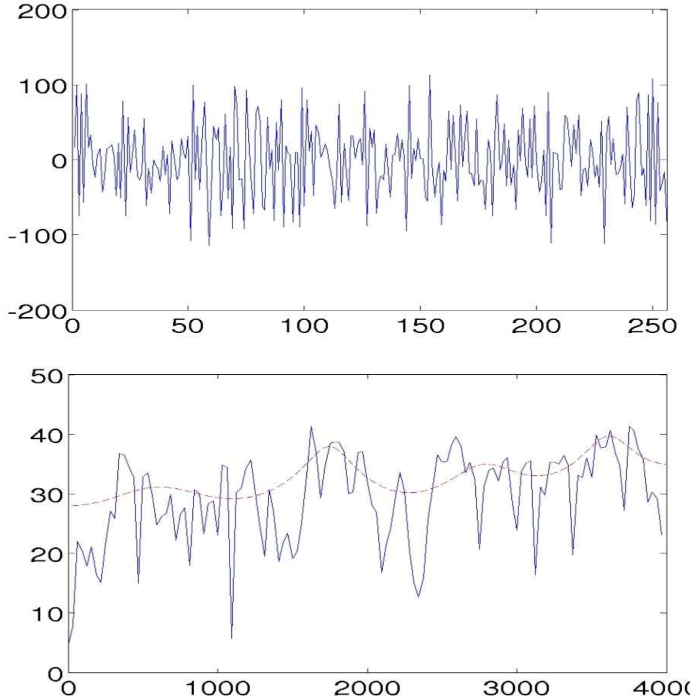
- Loudness
- Voiced/Unvoiced.
- Pitch.
  - Fundamental frequency.
- Spectral envelope.
  - Formants.

Voiced Speech



/ih/

Unvoiced Speech



/s/

Methods for Speech Processing in Time-Domain

- There are several operations that can be applied directly to the speech signal

  For example,

  Autocorrelation

  Energy analysis

  Pitch analysis

- Most operations require short-time analysis. For energy,

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

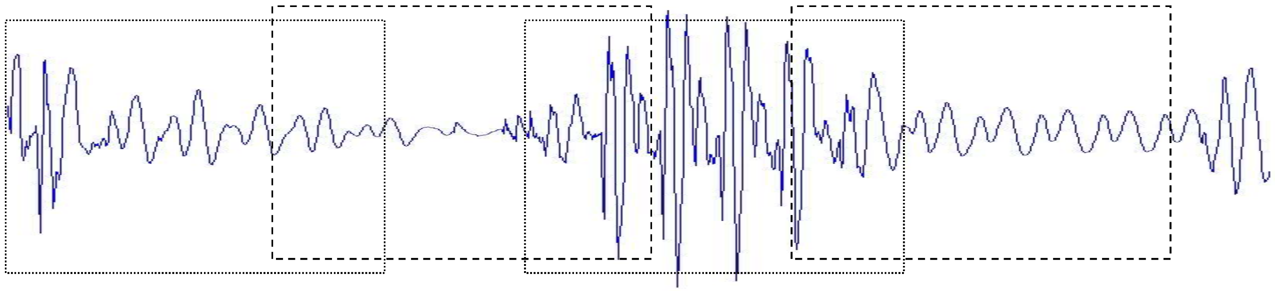
- In practice, short-time analysis is computed as

$$E_n = \sum_{m=n-N+1}^{n} x^2(m)$$

for frame n of N-samples.

Short-Time Speech Analysis

- Segments (or frames, or vectors) are typically of length 20 ms.
  - Speech characteristics are constant.
  - Allows for relatively simple modeling.
- Often overlapping segments are extracted.

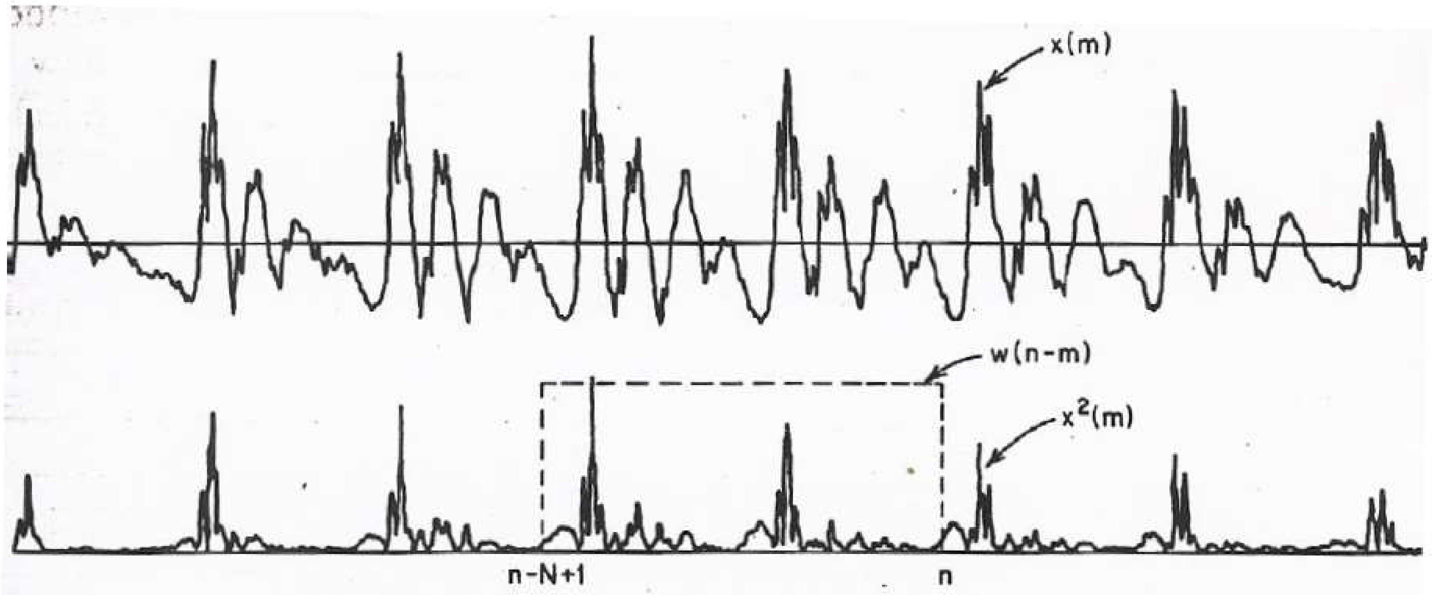


Methods for Speech Processing in Time-Domain

- This is equivalent to applying a square window

$$w(n) = 1 \qquad 0 \le n \le N-1$$
$$\qquad 0 \qquad \mathit{Otherwise}$$

which is a linear operation, and summing the square of the signal.



Windowing

Windowing speech is equivalent to applying a filtering effect (low-pass filtering) that help for (a) smoothing, and (b) segmenting the signal into smaller windows.

There are several different types of windows that can be applied on a speech signal

- Rectangular

$$h(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases}$$

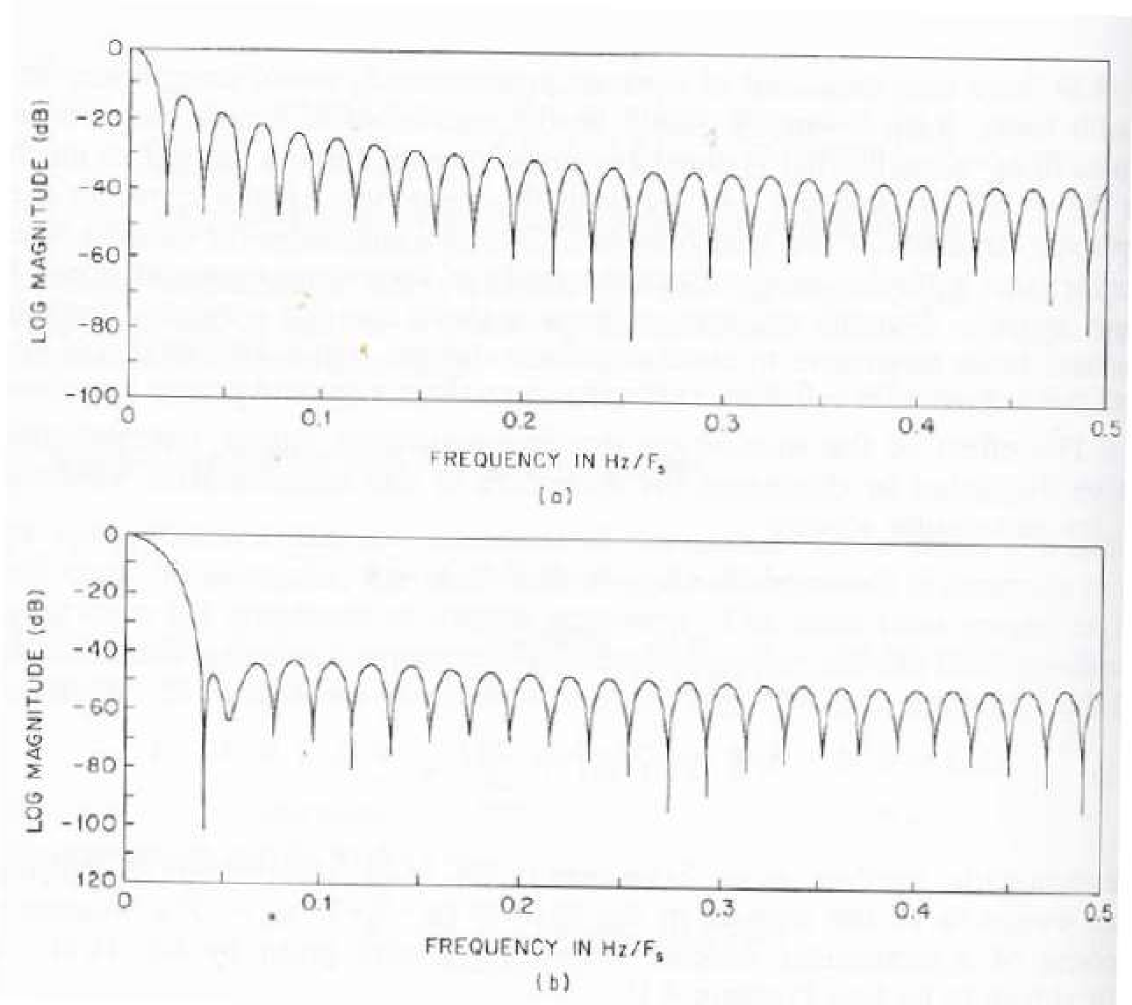
- Frequency Response

$$H(e^{jwT}) = \frac{\sin(wNT/2)}{\sin(wT/2)} e^{-jwT(N-1)/2}$$

- Sampling Frequency

$$F_s = 1/T$$

Windows Comparison



Rectangular

Hamming

Hamming Window

- Hamming

$$h(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases}$$

- Has twice through-bandwidth and much better attenuation than rectangular window
- Attenuation is independent of window duration, but duration is inversely proportional to the bandwidth.

Window Size

- A suitable selection of N is necessary to capture sufficient characteristics from the signal (e.g., pitch information).
- Window size is typically larger than one pitch period (e.g., 20-30 msec) to avoid large fluctuations in the short-time energy.
- Energy is a good indicator of voicing
    - High energy → Voiced sounds
    - Low to medium energy → Unvoiced sounds
    - Very low energy → Silence and background noise.
- Energy is a good voicing indicator for high SNR

Zero Crossing Rate

- Speech samples have both negative and positive signs.
- Crossing rate is a simple measure of frequency contents in the signal
- To reliably compute the zero crossing rate, a signal needs to pass through a Band-pass filter to eliminate dc offset, and low-frequency noise.

- For a sinusoidal signal,

$$ZC = 2f_0 / F_s$$

„Zeros Crossing is said to occur of successive samples have diferent algebraic signs.“

- For speech

$$ZC_n = \sum_{m=-\infty}^{\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|\, w(n-m)$$

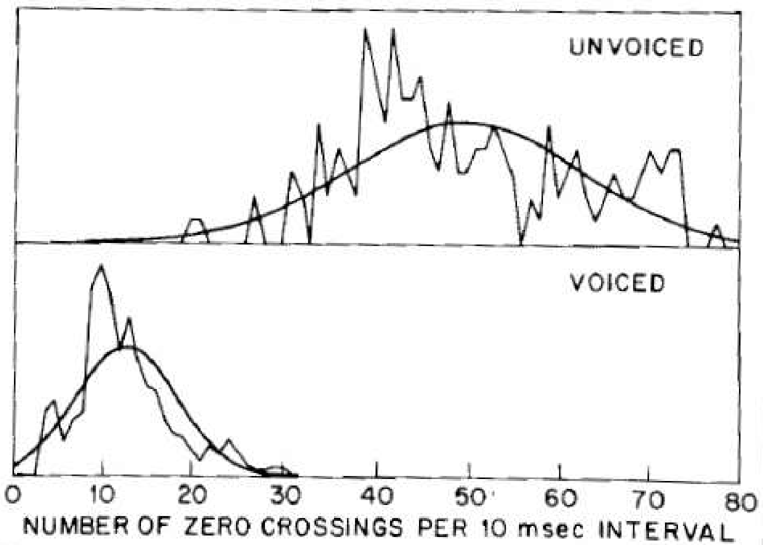
$$\operatorname{sgn}(x) = 1 \qquad x \geq 0$$
$$= -1 \qquad x < 0$$

$$w(n) = \frac{1}{2N} \qquad 0 \leq n \leq N-1$$
$$= 0 \qquad Otherwise$$

Zero Crossing for Detecting Voiced/Unvoiced

- Most speech energy < 4KHz
- For voiced speech, energy is typically at low frequency
- For unvoiced speech, energy is at high frequency

High frequency → high ZC → Unvoiced
Low frequency → Low ZC → Voiced



ZC~40-50

ZC~10-20

Speech/Non-Speech Classification

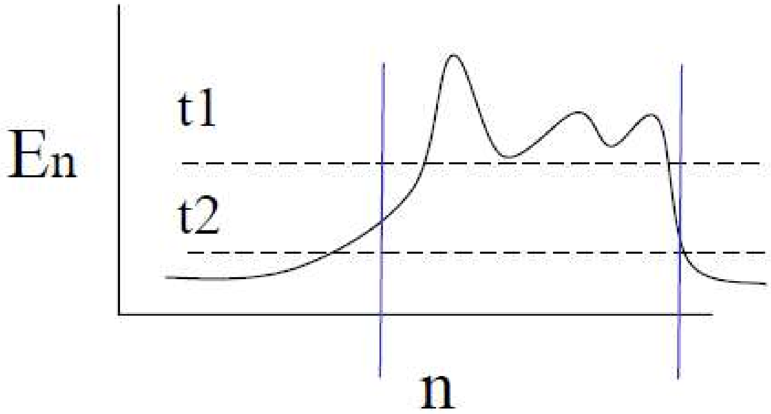
- E_n and ZC_n can be applied for speech/non-speech separation.
- These measurements are sensitive to background noise and weak fricatives.
- In many cases, even for high SNR, it is difficult to locate weak fricatives, nasals, weak plosives especially if they occur at the beginning or end of a sentence.

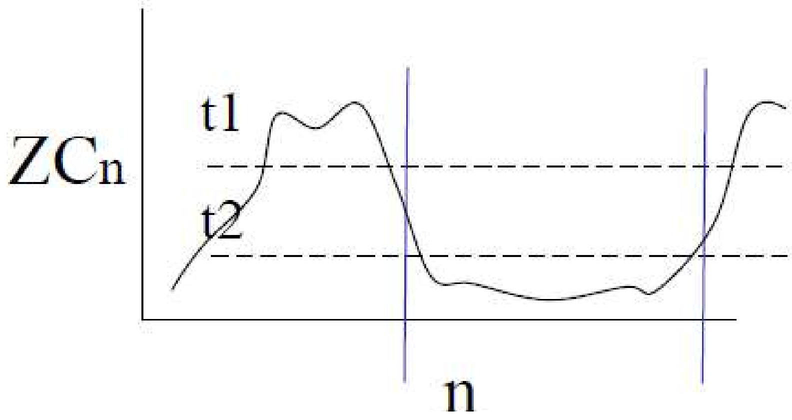
How would you create a speech/non-speech classifier using E_n and ZC_n?

Speech Detection – Example (1/2)

- Initial 100-200 msec are assumed background, and estimates for the mean and standard deviation of E_0 and ZC_0 are computed.
- Based on these estimates, appropriate thresholds are selected for determining speech regions.

  - Two or more thresholds can also be instituted for more detailed modeling. The signal can be classified into speech present/speech may be present/speech is not present.





Speech Detection – Example (2/2)

- “Segment duration” can also be used as a feature to establish smoother and more reliable speech regions.
- At the marked regions of speech, forward and backward extension is typically allowed to embrace weak fricatives and plosives.

- A statistical model can also be used for speech detection. A classifier can be trained on a subset of manually-labeled data, and then be presented with various features to discriminate speech from non-speech.

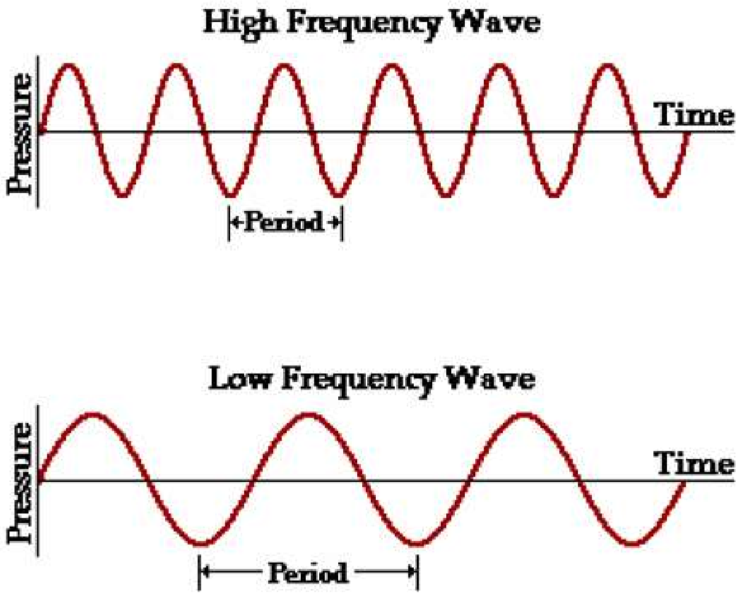
Pitch of the Speech Signal

- Pitch (fundamental frequency, or periodicity) is the period between two consecutive openings of the vocal cords.

- The primary difference between adult male and female/prepubescent speech is pitch. Before puberty, pitch frequency for normal speech ranges between 150-400 Hz for both males and females. After puberty, the vocal cords of males undergo a physical change, which has the effect of lowering their pitch frequency to the range 80-160 Hz.

- Pitch estimate is essential for many aspects of speech processing including speech coding, synthesis, recognition, and speaker verification.

- Most successful approaches to estimating pitch are based on autocorrelation.



Short-term Auto-Correlation (1/4)

- Long-term autocorrelation function:

$$AC_k = \sum_{m=-\infty}^{\infty} x(m).x(m+k)$$

- If the signal is periodic with pitch P,

$$AC_k = AC_{k+P}$$

Properties of Autocorrelation Function:

- Symmetric $AC_k = AC_{-k}$
- Maximum at k=0, +-P,+-2P,
- AC_0 is the signal energy
- Pitch is computed by finding the location of the first maximum in the autocorrelation function

Short-term Auto-Correlation (2/4)

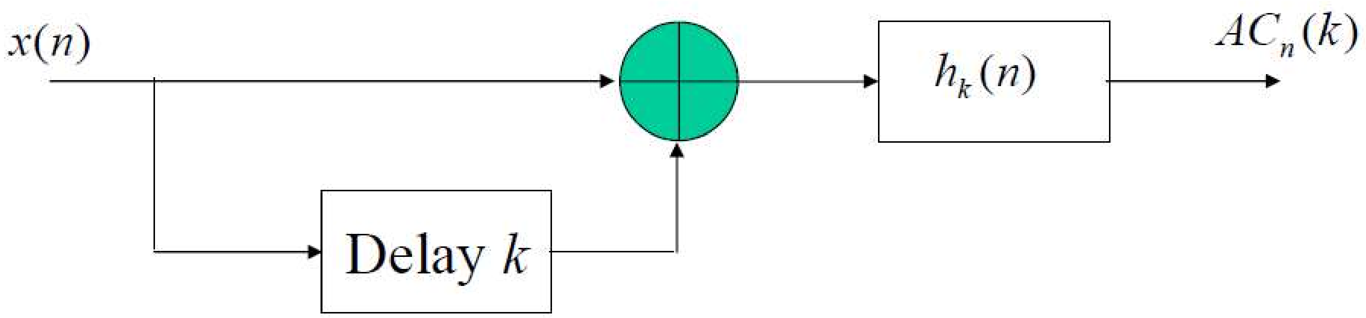
- Short-term autocorrelation function:

$$AC_n(k) = \sum_{m=0}^{N-1-k} x(n+m).x(n+m+k).h_k(n-m)$$

- This is similar to applying a filter on the sequence *x(m).x(m+k)*, where

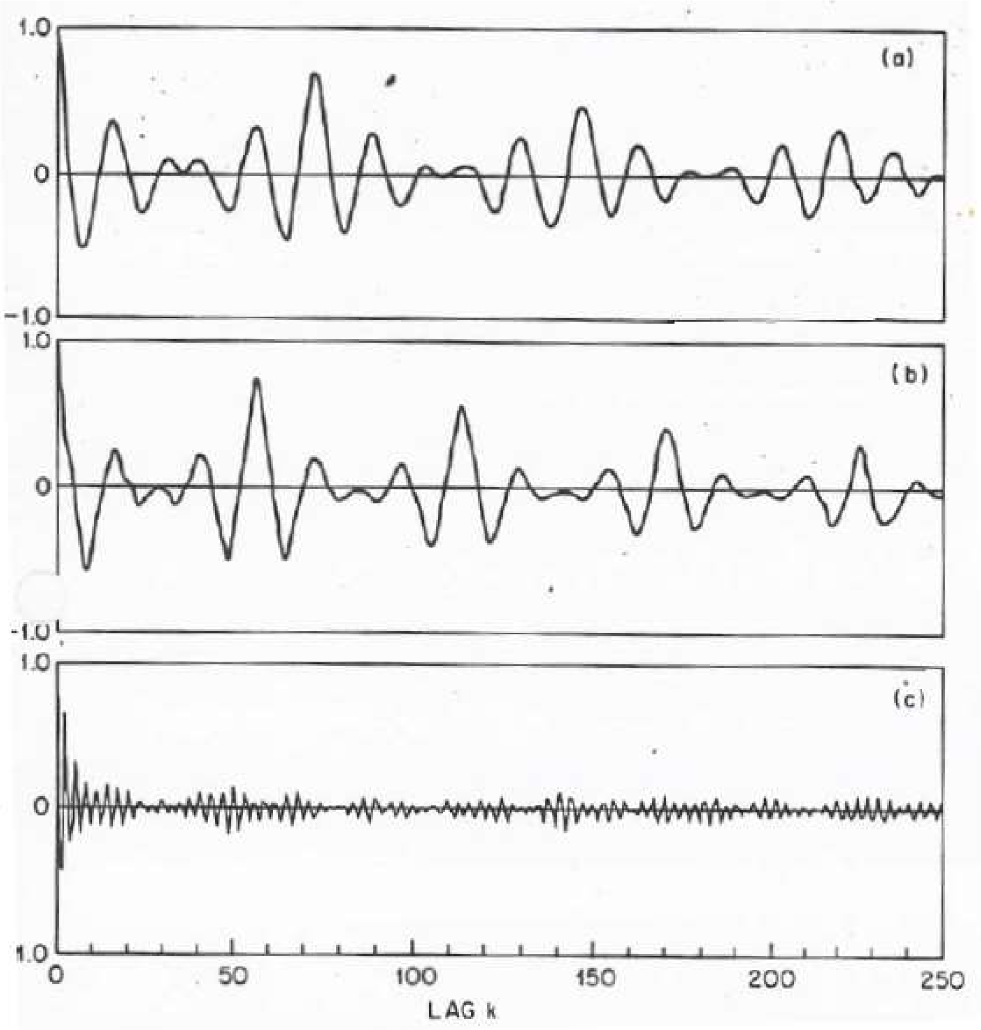
$$h_k(n) = w(n).w(n+k)$$

- The window w(.) is typically symmetric at m=0. It can be a rectangular or Hamming.



Short-term Auto-Correlation (3/4)

• Peaks occur at regular intervals



Voiced

Voiced

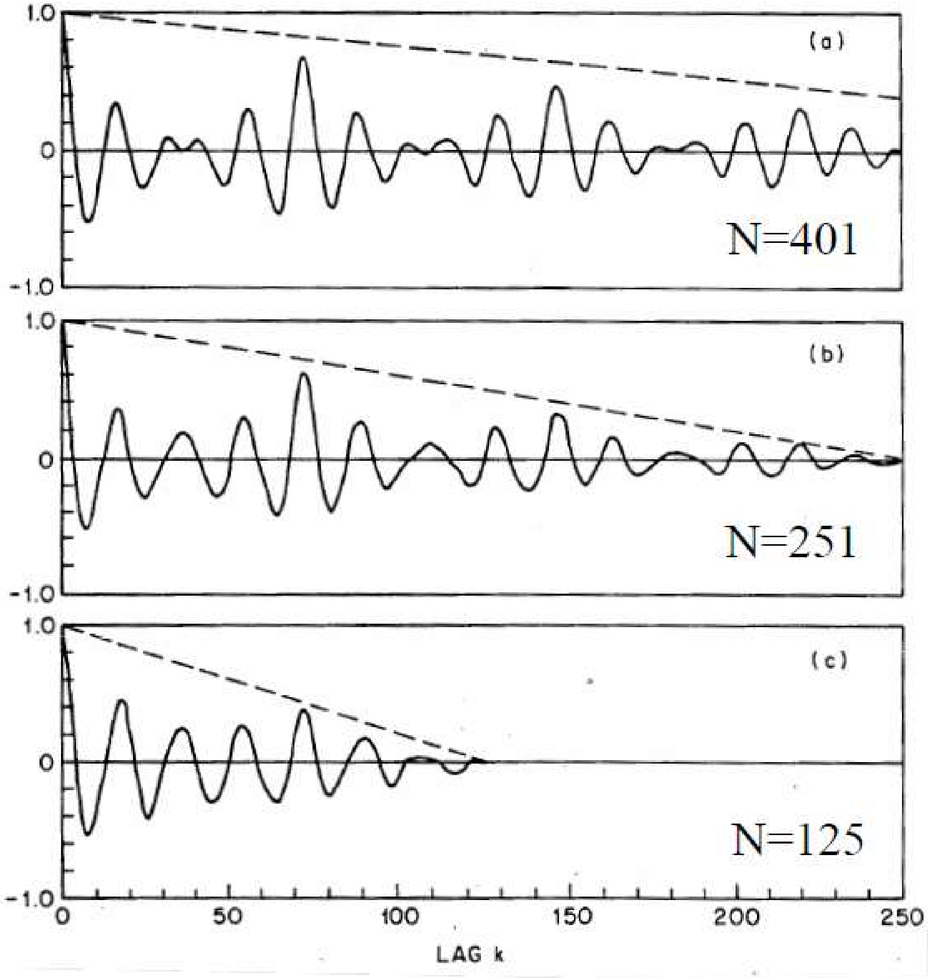
• Lack of periodicity

Unvoiced

Short-term Auto-Correlation (4/4)

Selection of Window Size:

- Analysis window must include at least two pitch periods (20-40msec)
- Should be small enough to capture details of the signal (less computation)



Tapering Effect:

- Using a rectangular window, an autocorrelation function will be tapered by

$$AC_k \qquad = 1 - k/N \qquad |k| < N$$

- Can be avoided by normalization, or by extending the autocorrelation window.

Average Magnitude Difference Function

- For a true periodic signal,

$$d(n) = x(n) - x(n-k)$$

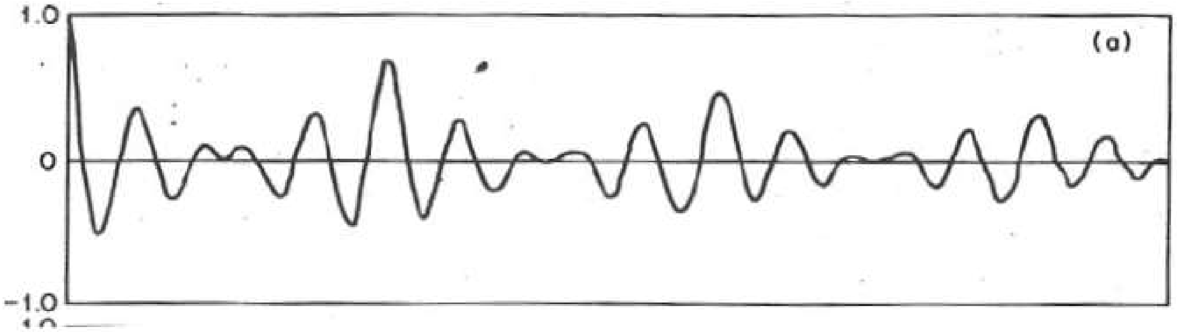
Where, $$d(n) = 0 \qquad k = 0, \pm P, \pm 2P, ..$$

- AMDF function

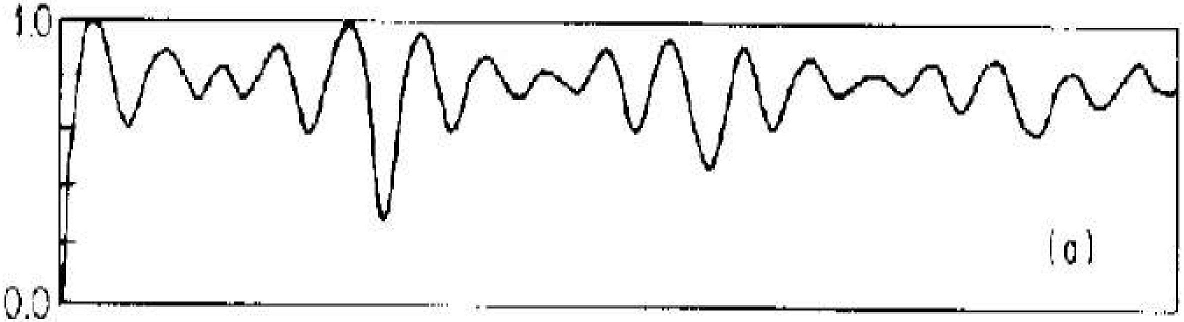
$$AM_n(k) = \sum_{m=-\infty}^{\infty} | x(n+m)w_1(m) - x(n+m-k)w_2(m-k) |$$

•AMDF function can be implemented with subtraction, addition and absolute value operation – more efficient than computing an autocorrelation function.

AMDF Vs AC



AC



AMDF

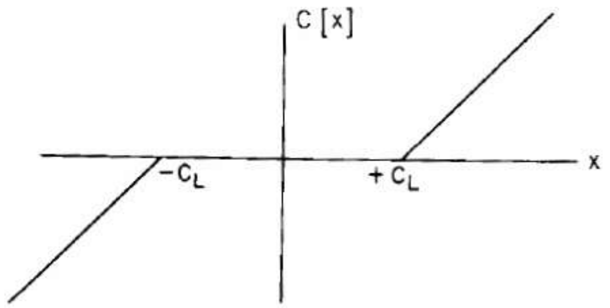
Pitch Period

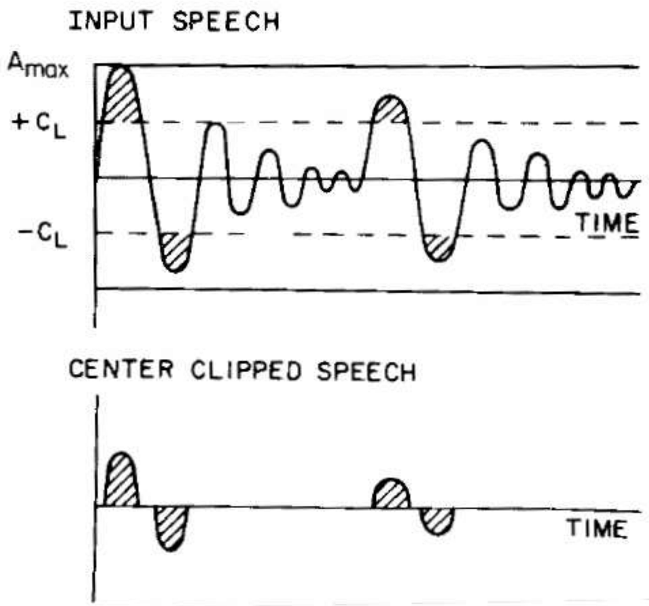
- Autocorrelation (or AMDF) can be used to estimate the pitch period of speech
- To enhance the pitch estimate, details of the autocorrelation function, corresponding to high frequency variations, are minimized? Why?
    - Spectral flattening is necessary to minimize vocal tract modeling effects and hence enhance pitch harmonics.

- Several methods are used for pitch enhancements. For example,
    - Center clipping
    - 3-level center clipping

Center Clipping

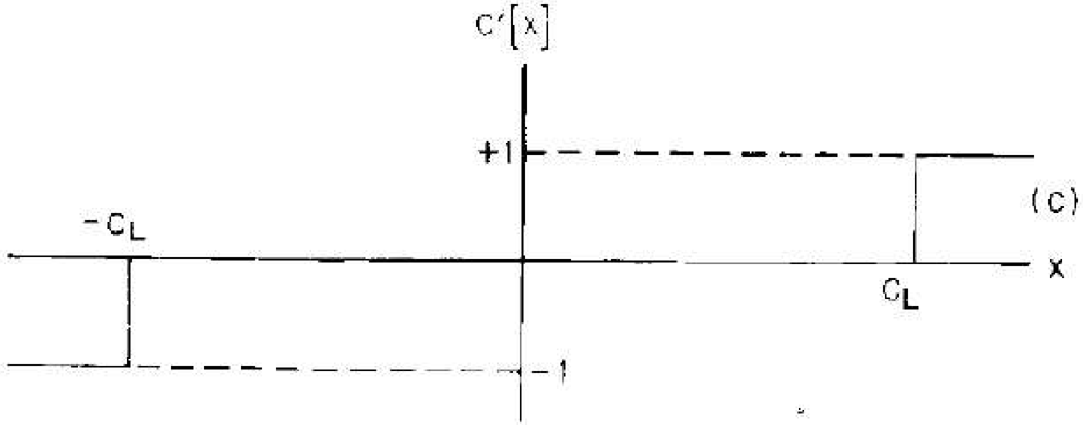
$$y(n) = C[x(n)]$$





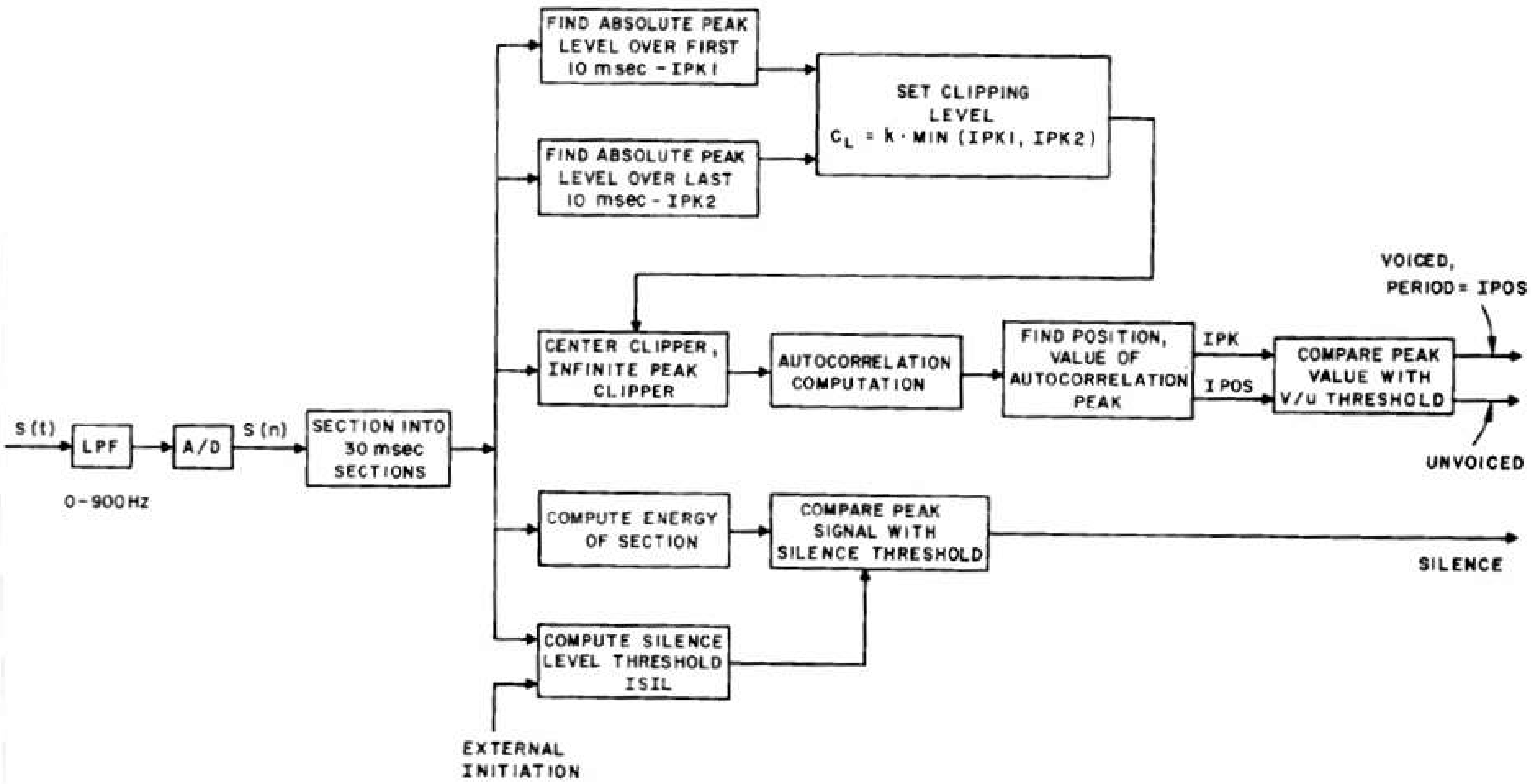
C_L is selected based on $x(n)_{max}$

3-Level Center Clipping



• Convert speech signal above/below CL into +1/-1, otherwise 0.
• This non-linear filtering avoids magnitude effects.

Methods for Pitch Estimation



- May want to do speech detection first using energy and zero crossing
- Hamming window may be applied to impose low-magnitude smoothing