Lecture 07:
[Rabiner] Speech/Non-speech detection and end-point detection (EPD)

DEEE725 음성신호처리실습

Instructor: 장길진

Original slides from Lawrence Rabiner

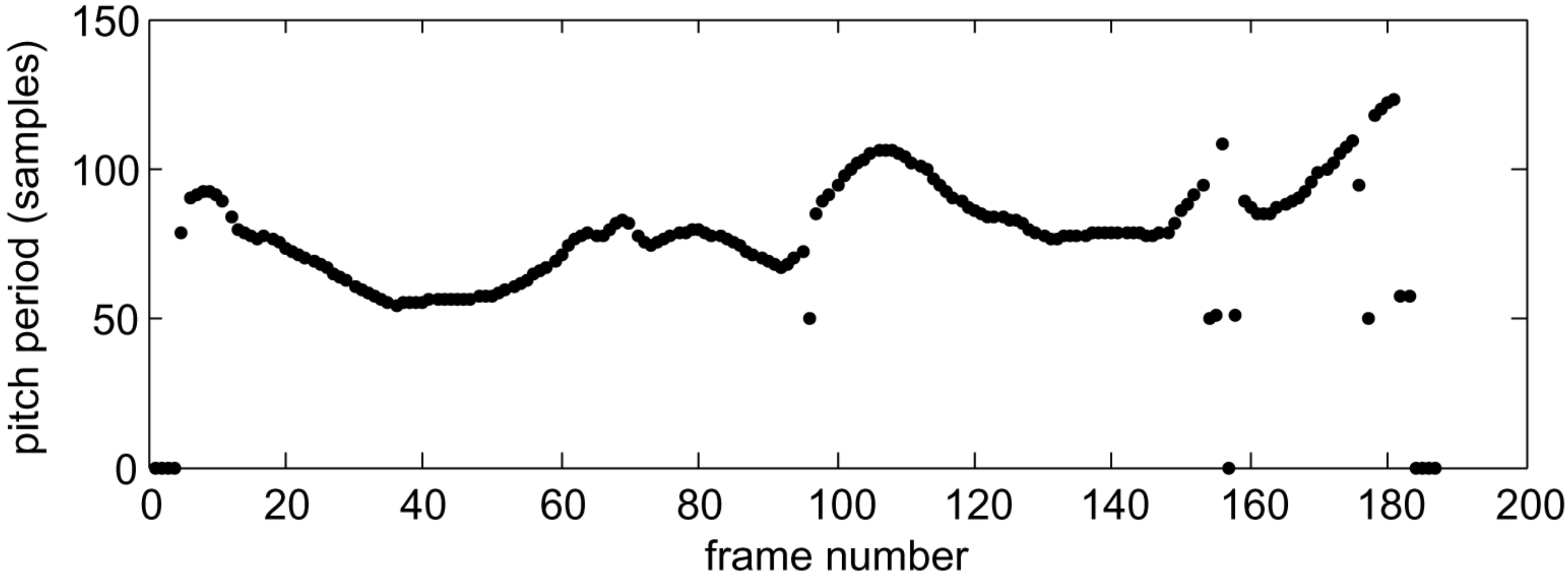
Speech Processing Algorithms

- Speech/Non-speech detection
  – Rule-based method using (log) energy and zero crossing rate
  – Single speech interval in background noise (end-point detection, EPD)

- Voiced/Unvoiced/Background classification
  – Bayesian approach using 5 speech parameters
  – Needs to be trained (mainly to establish statistics for background signals)

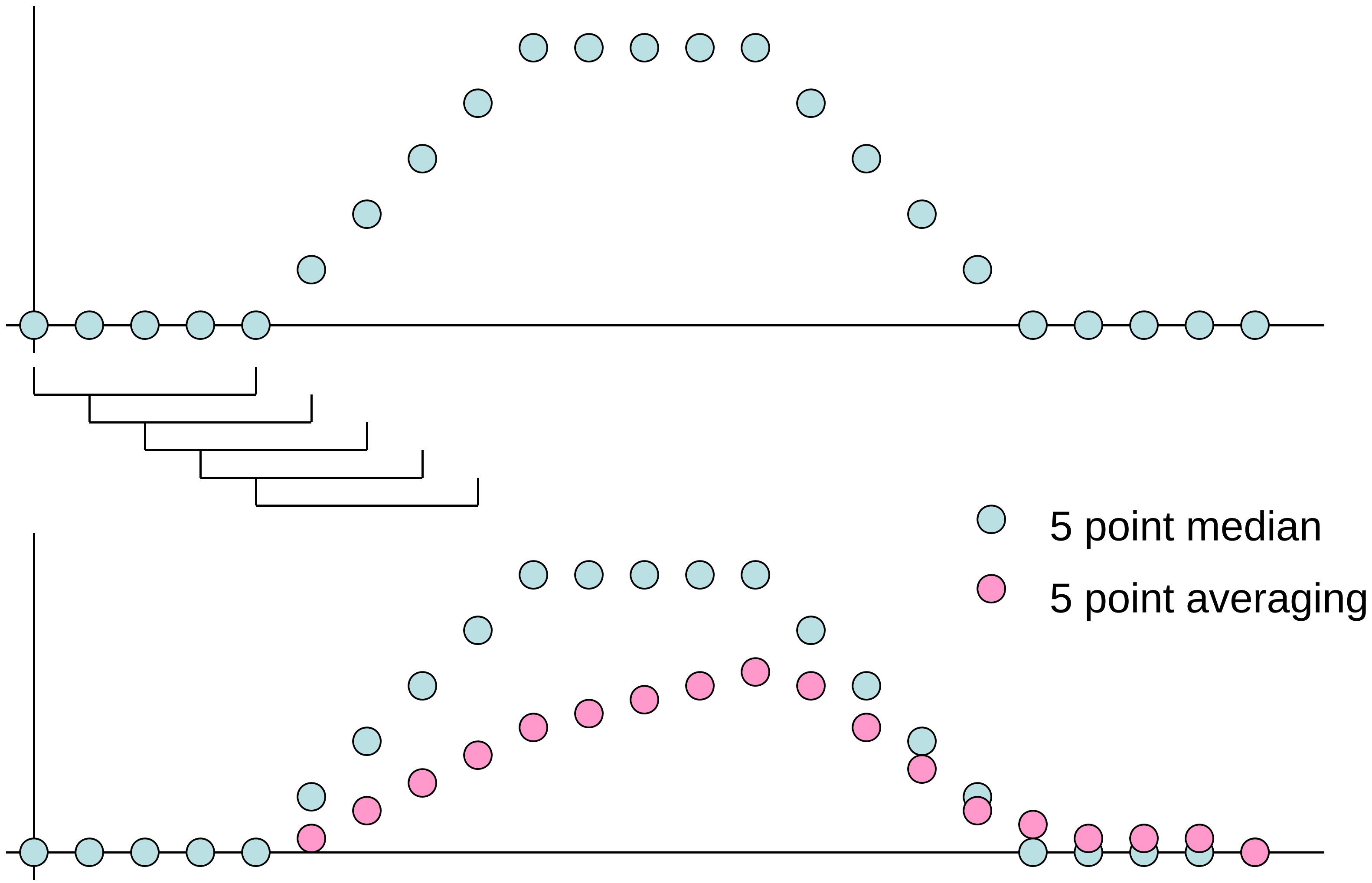
Median Smoothing and Speech Processing

Why Median Smoothing



Obvious pitch period discontinuities that need to be smoothed in a manner that preserves the character of the surrounding regions – using a median (rather than a linear filter) smoother.

Running Medians



Non-Linear Smoothing

- linear smoothers (filters) are not always appropriate for smoothing parameter estimates because of smearing and blurring discontinuities
- pitch period smoothing would emphasize errors and distort the contour
- use combination of non-linear smoother of running medians and linear smoothing
- linear smoothing => separation of signals based on non-overlapping frequency content
- non-linear smoothing => separating signals based on their character (smooth or noise-like)

$$x[n] = S(x[n]) + R(x[n]) \quad \text{- smooth} + \text{rough components}$$

$$y(x[n]) = \text{median}(x[n]) = M_L(x[n])$$

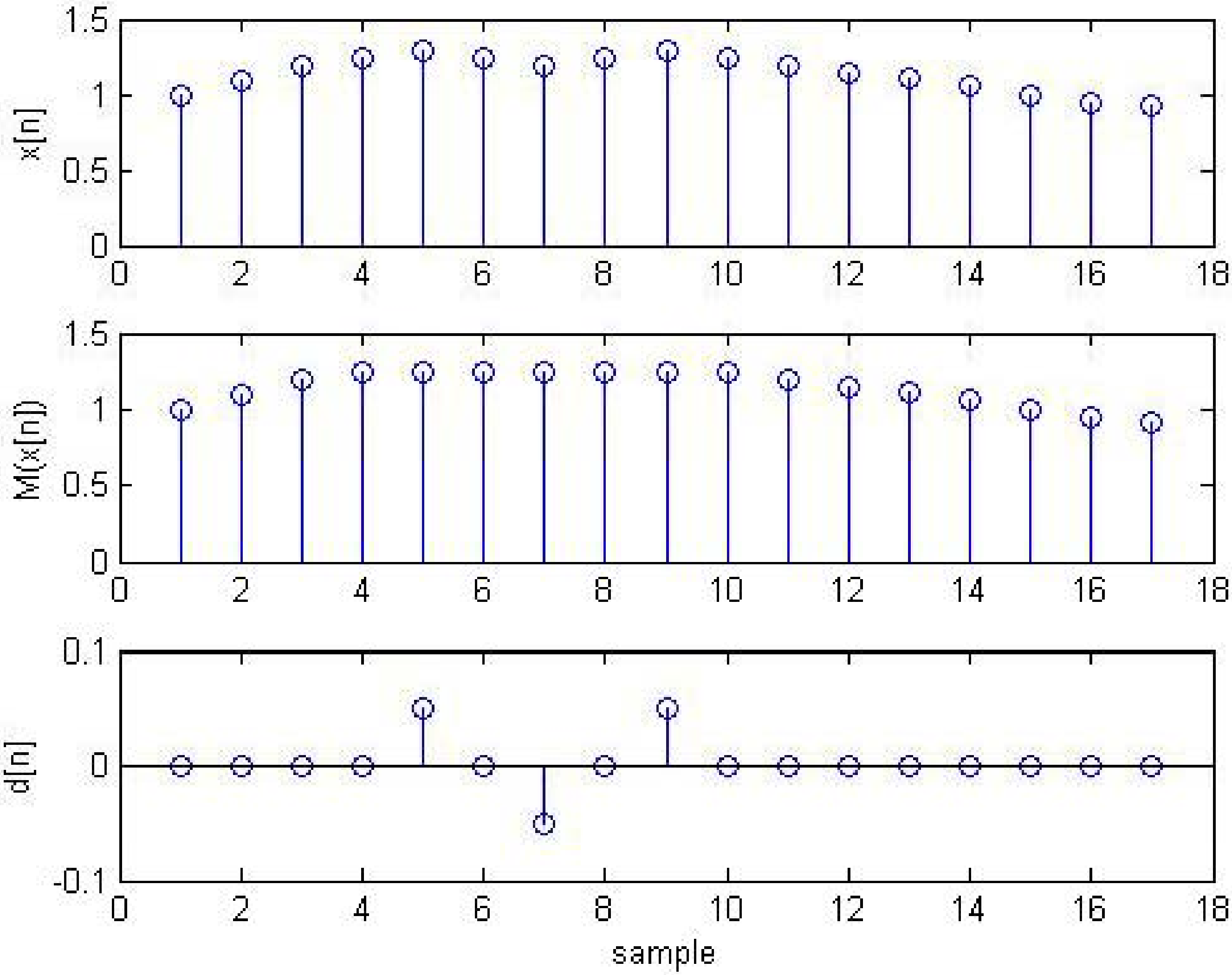
$$M_L(x[n]) = \text{median of } x[n] \ldots x[n-L+1]$$

Properties of Running Medians

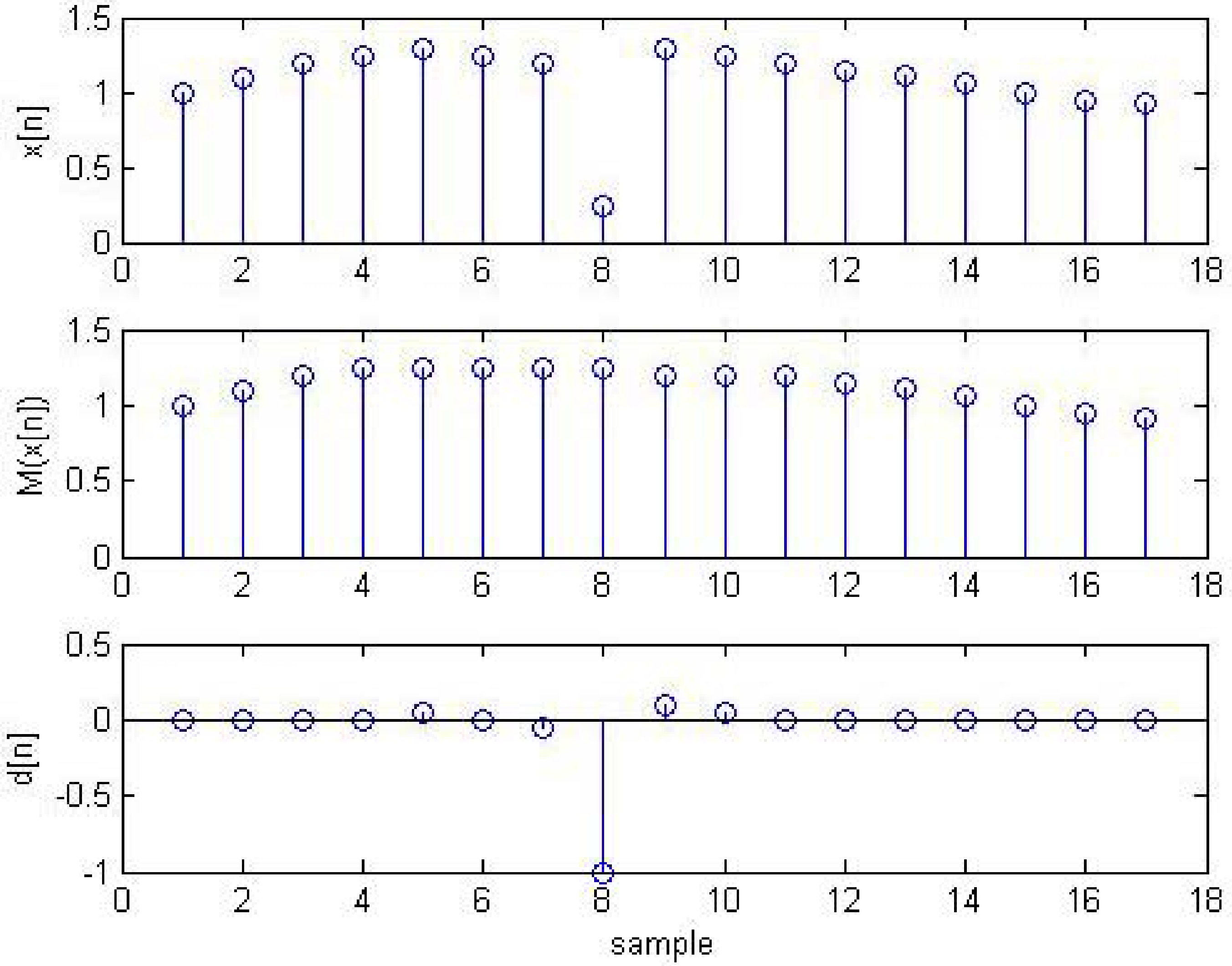
Running medians of length L:

1. $M_L(\alpha\, x[n]) = \alpha\, M_L(x[n])$
2. Medians will <u>not</u> smear out discontinuities (jumps) in the signal if there are no discontinuities within $L/2$ samples
3. $M_L(\alpha\, x_1[n]+\beta\, x_2[n]) \neq \alpha\, M_L(x_1[n]) + \beta\, M_L(x_2[n])$
4. Median smoothers generally preserve sharp discontinuities in signal, but fail to adequately smooth noise-like components

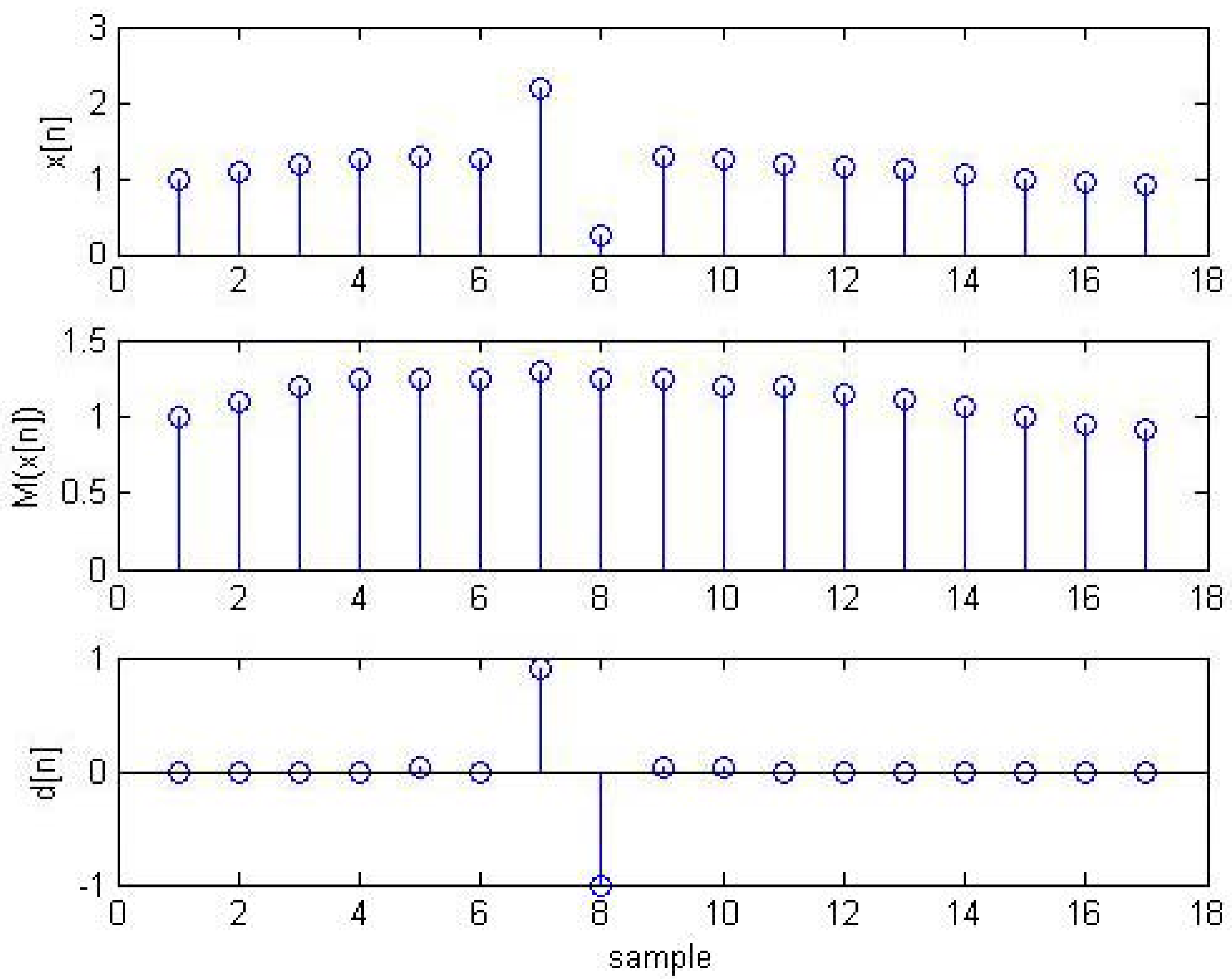
Median Smoothing



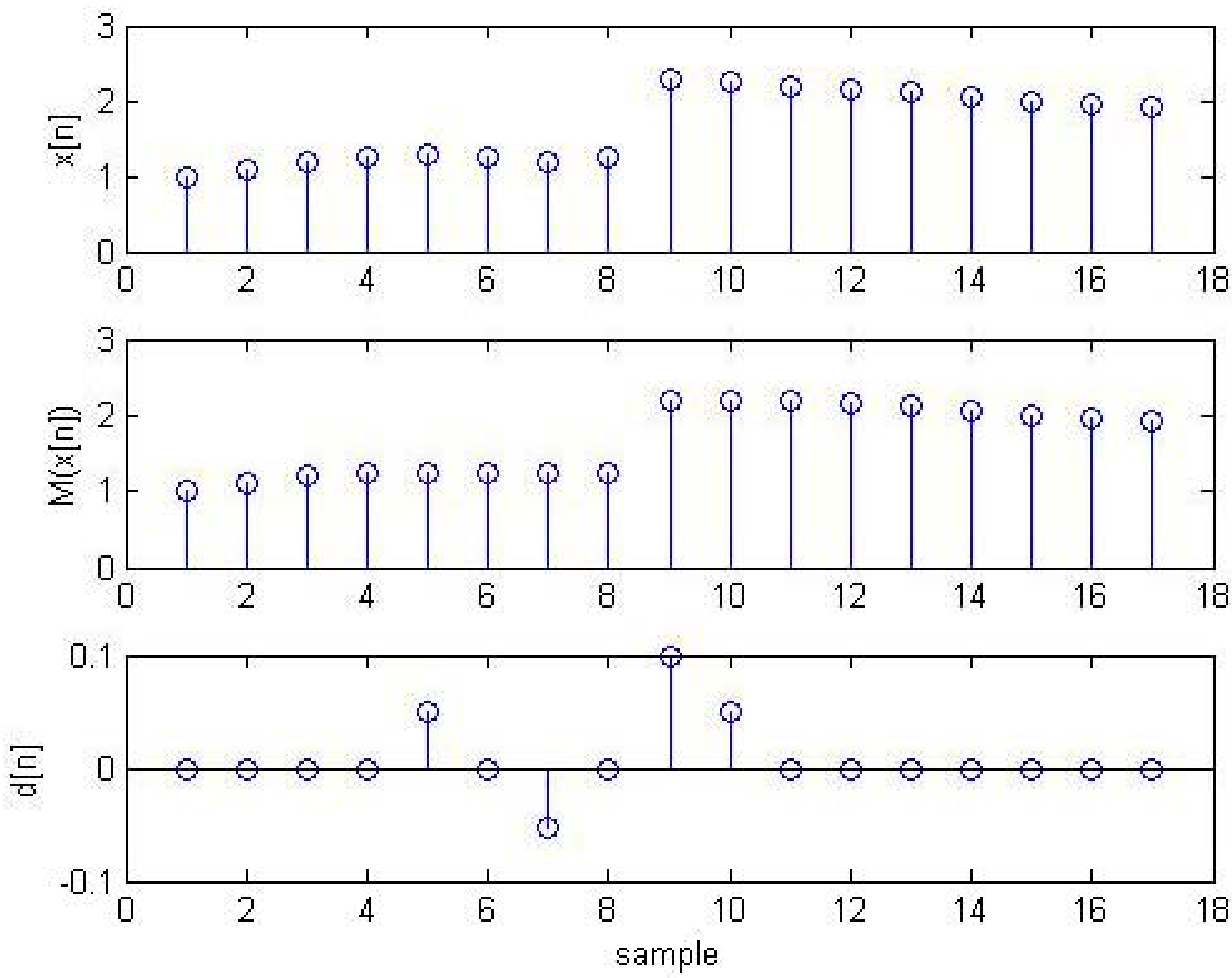
Median Smoothing



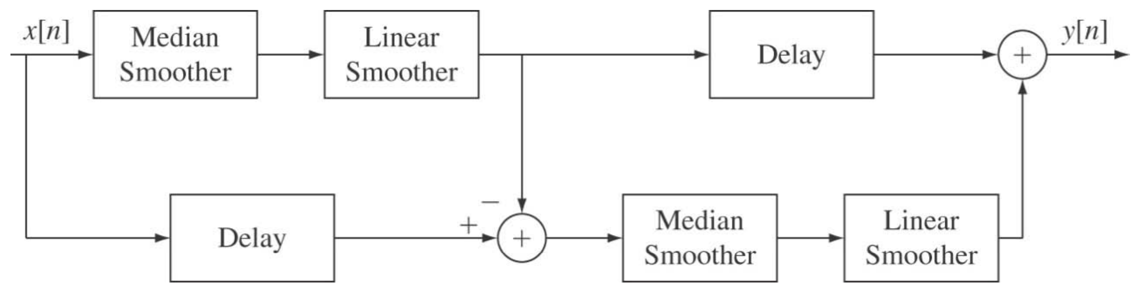
Median Smoothing

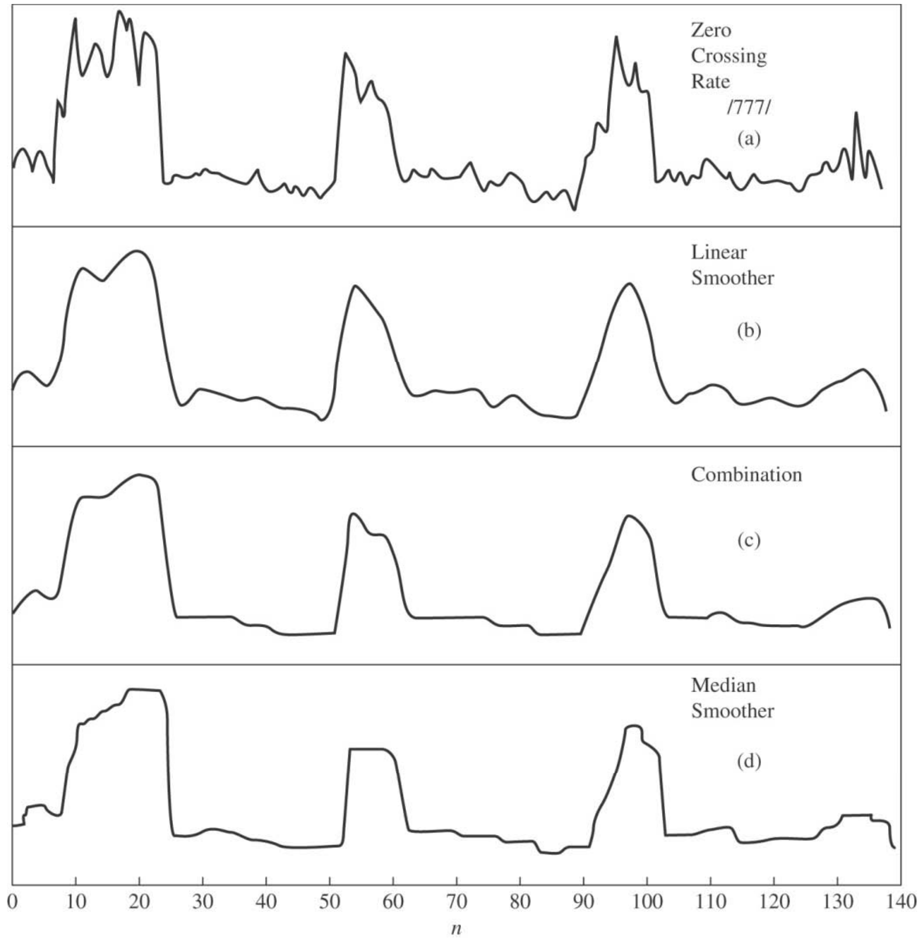


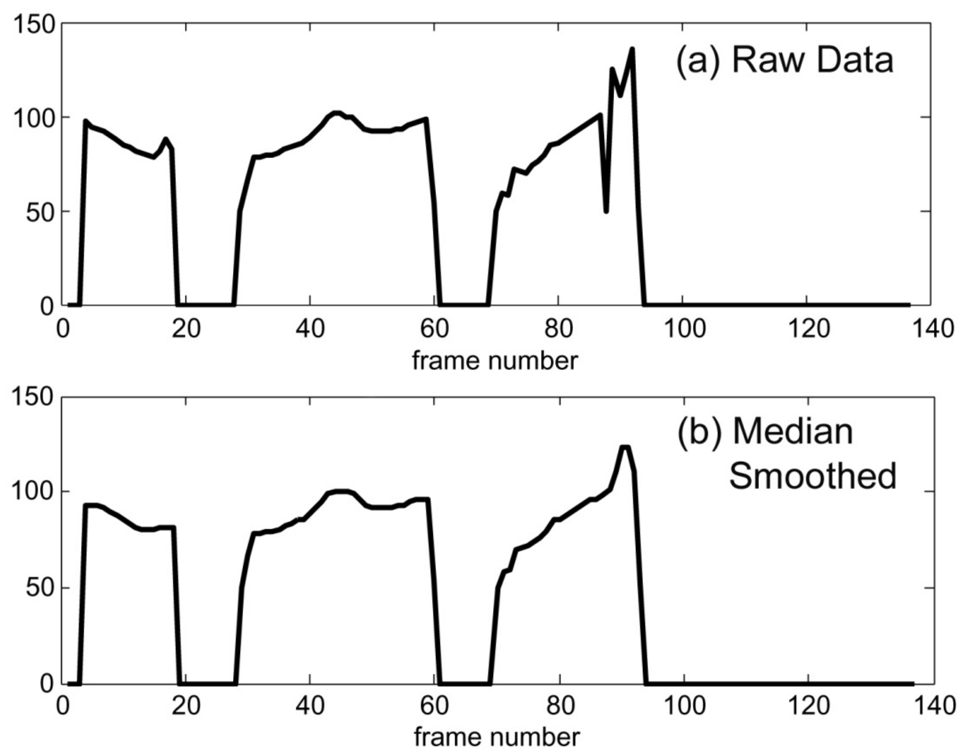
Median Smoothing



Nonlinear Smoother with Delay Compensation







Algorithm #1

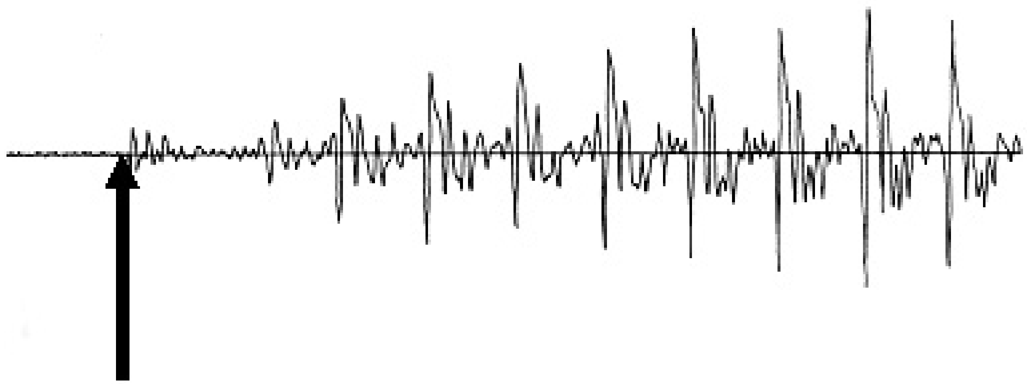
Speech/Non-Speech Detection Using Simple Rules

Speech Detection Issues

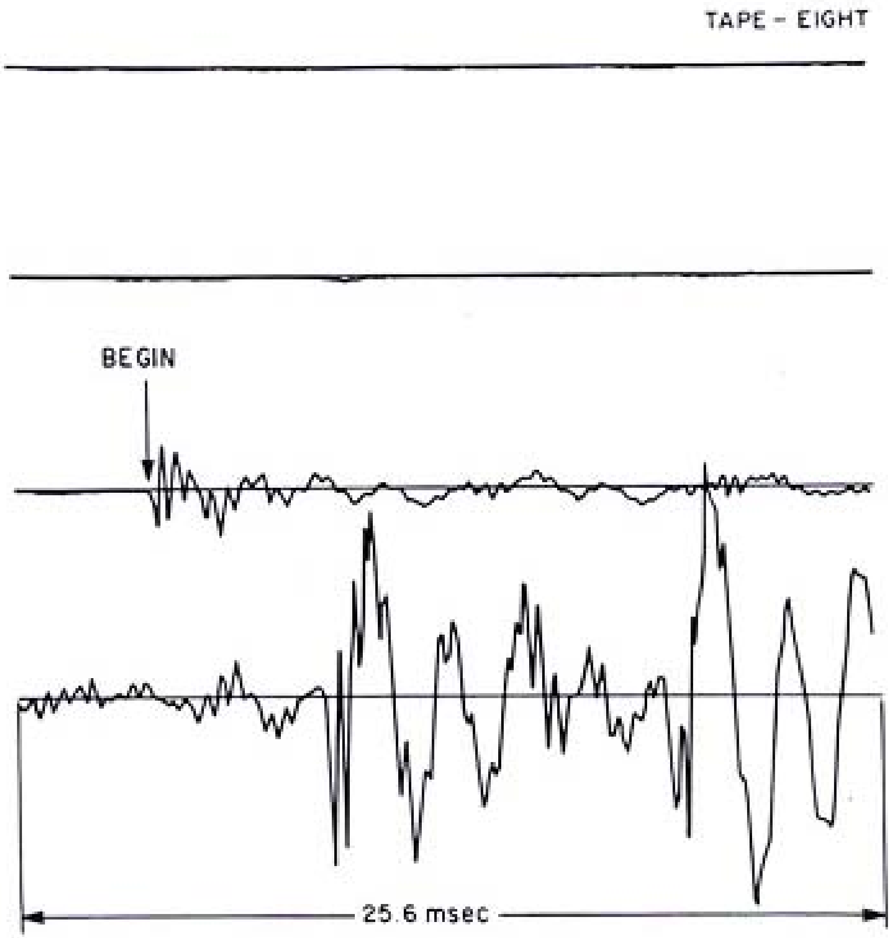
- key problem in speech processing is locating accurately the beginning and end of a speech utterance in noise/background signal

- need endpoint detection to enable:
    - computation reduction (don't have to process background signal)
    - better recognition performance (can't mistake background for speech)
- non-trivial problem except for high SNR recordings

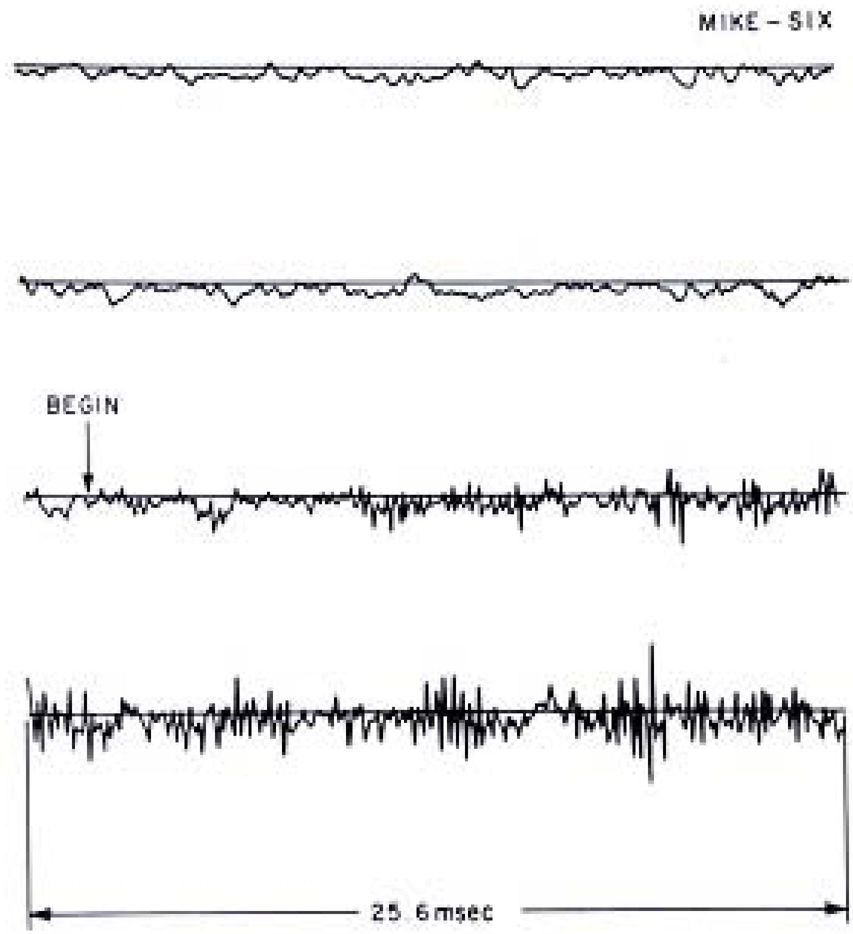
Ideal Speech/Non-Speech Detection



Speech Detection Examples

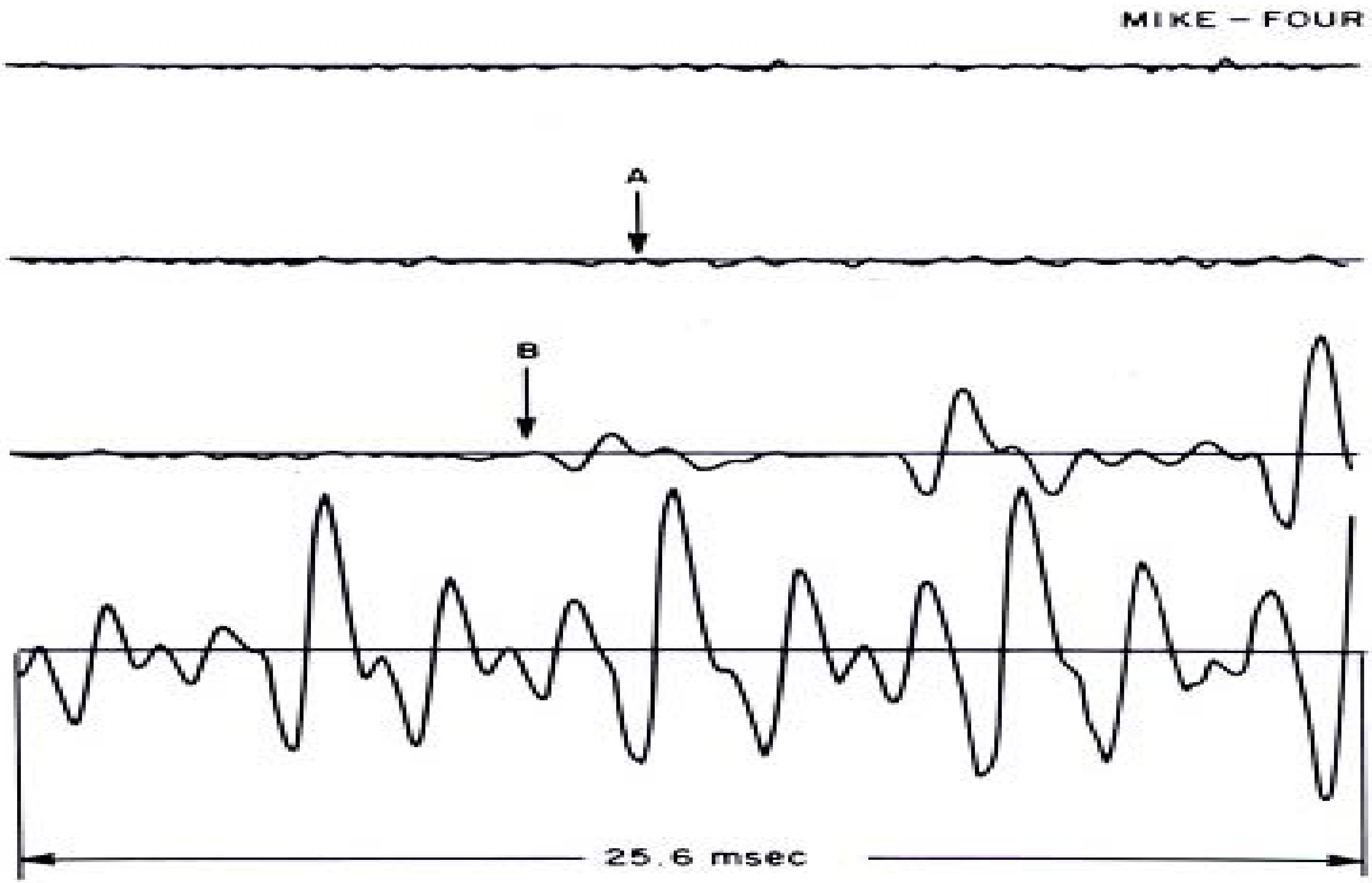


case of low background noise =>
simple case



can find beginning of speech based on knowledge of sounds (/S/ in six)

Speech Detection Examples



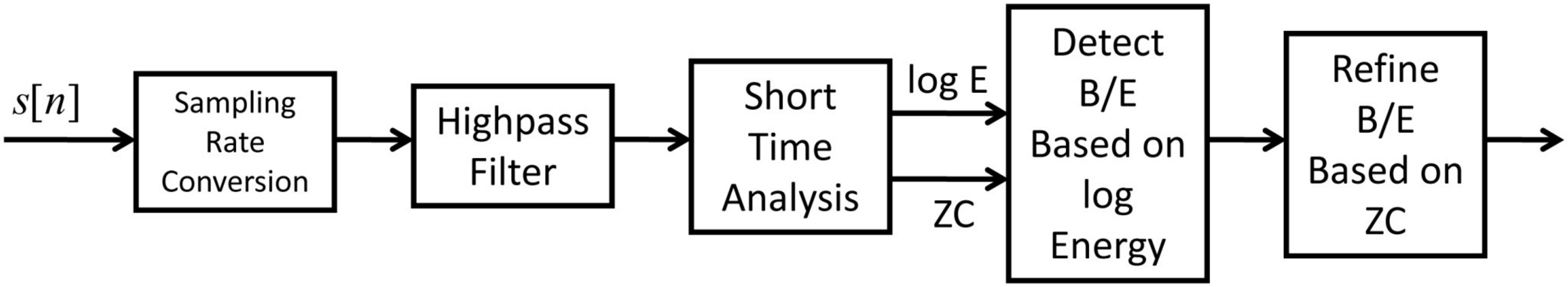
difficult case because of weak fricative sound,
/f/, at beginning of speech

Problems for Reliable Speech Detection

- weak fricatives (/f/, /th/, /h/) at beginning or end of utterance
- weak plosive bursts for /p/, /t/, or /k/
- nasals at end of utterance (often devoiced and reduced levels)
- voiced fricatives which become devoiced at end of utterance
- trailing off of vowel sounds at end of utterance

the good news is that highly reliable endpoint detection is not required for most practical applications; also we will see how some applications can process background signal/silence in the same way that speech is processed, so endpoint detection becomes a moot issue

Speech/Non-Speech Detection

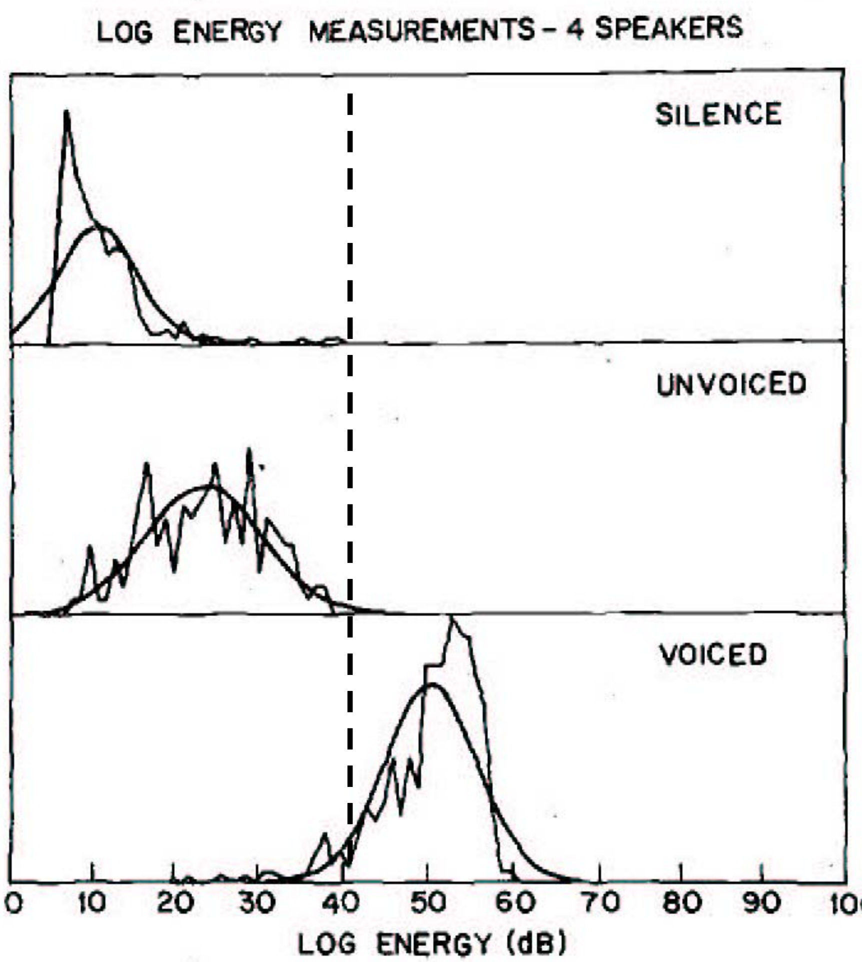


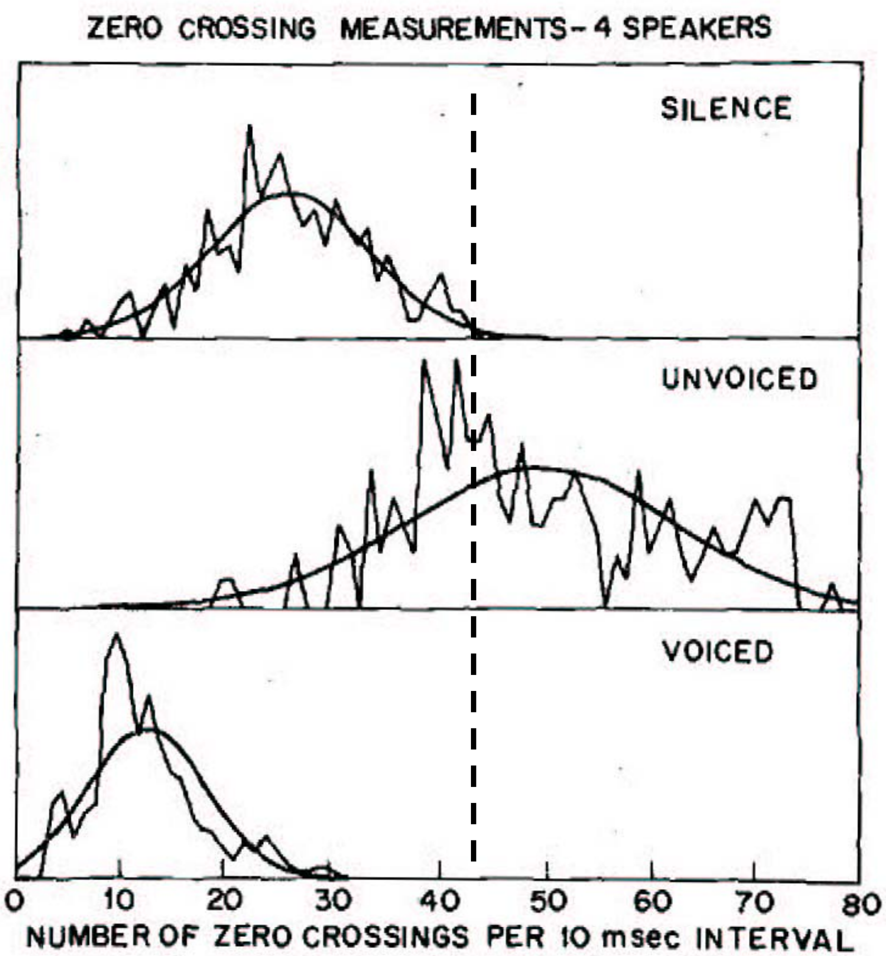
- sampling rate conversion to standard rate (10 kHz)
- highpass filtering to eliminate DC offset and hum, using a length 101 FIR equiripple highpass filter
- short-time analysis using frame size of 40 msec, with a frame shift of 10 msec; compute short-time log energy and short-time zero crossing rate
- detect putative beginning and ending frames based entirely on short-time log energy concentrations
- detect improved beginning and ending frames based on extensions to putative endpoints using short-time zero crossing concentrations

Speech/Non-Speech Detection – Algorithm #1

1. Detect **beginning** and **ending** of speech intervals using short-time energy and short-time zero crossings
2. Find **major concentration of signal** (guaranteed to be speech) using region of signal energy around maximum value of short-time energy => energy normalization
3. **Refine region of concentration** of speech using reasonably tight short-time energy thresholds that separate speech from backgrounds—but may fail to find weak fricatives, low level nasals, etc
4. **Refine endpoint estimates** using zero crossing information outside intervals identified from energy concentrations—based on zero crossing rates commensurate with unvoiced speech

Speech/Non-Speech Detection





Log energy separates Voiced from Unvoiced and Silence

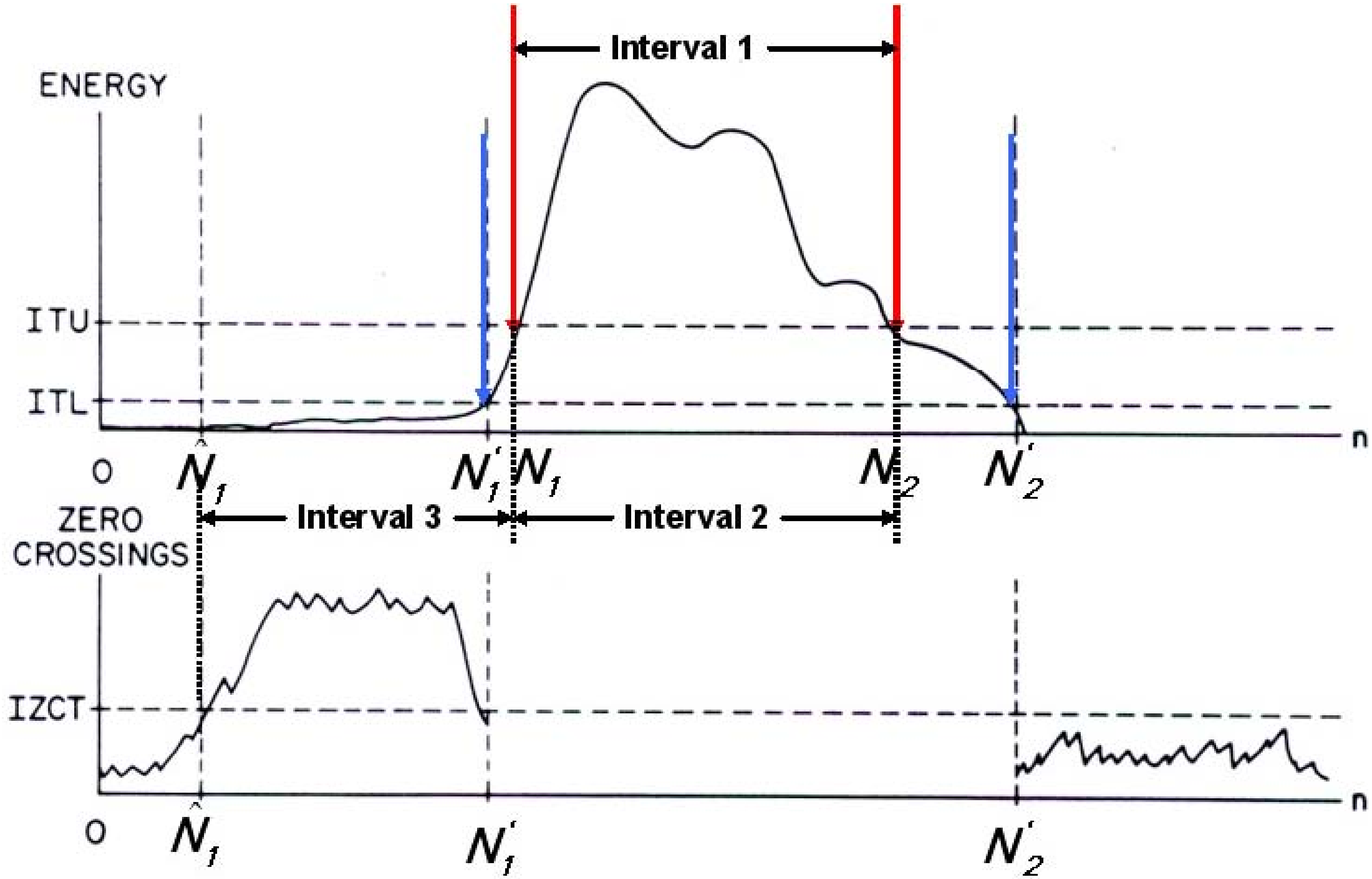
Zero crossings separate Unvoiced from Silence and Voiced

Rule-Based Short-Time Measurements of Speech

Algorithm for endpoint detection:

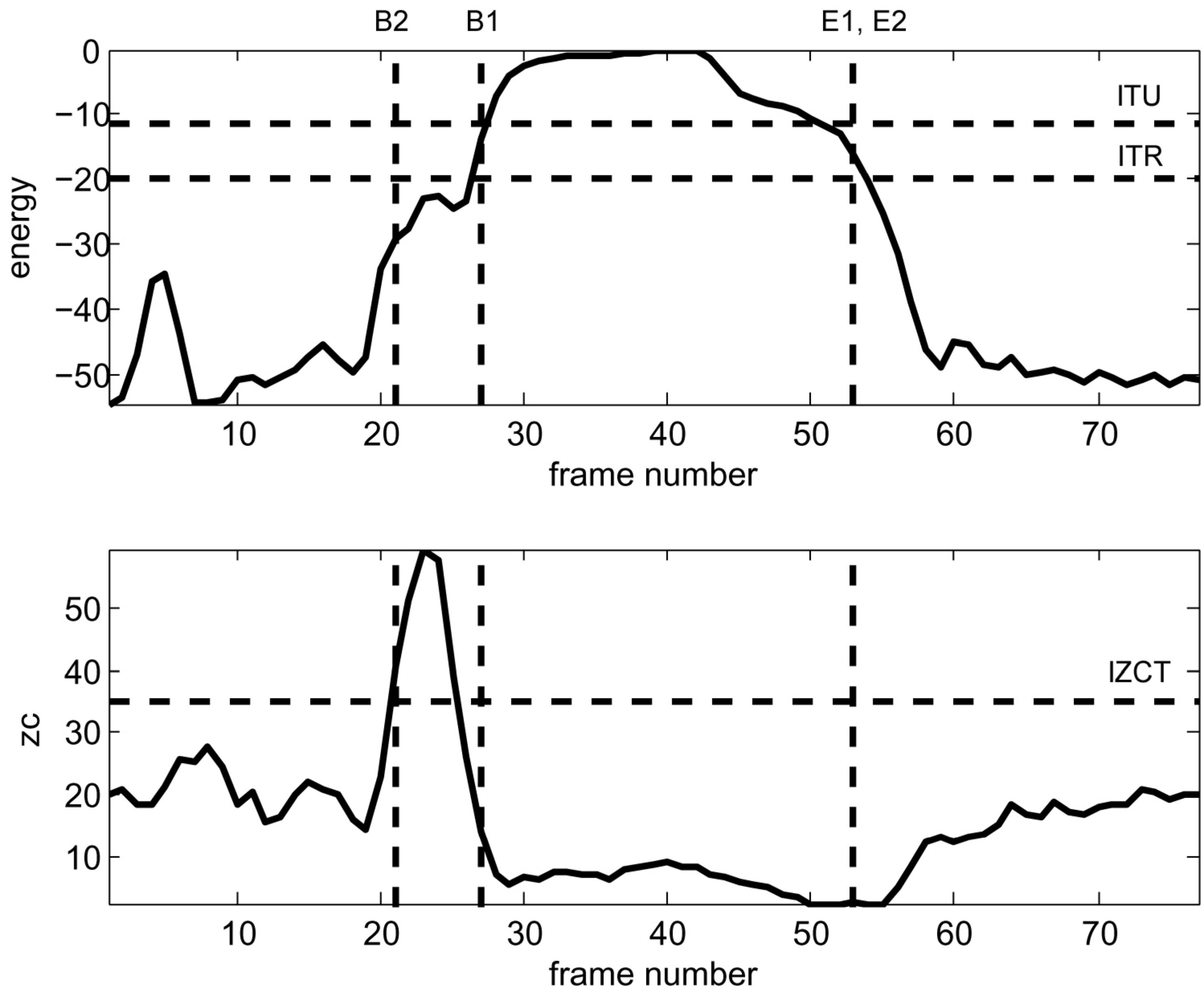
1. compute mean and σ of log E_n and Z_{100} for first 100 msec of signal (assuming no speech in this interval and assuming $F_S = 10{,}000\ Hz$).
2. determine maximum value of log E_n for entire recording => normalization.
3. compute log E_n thresholds based on results of steps 1 and 2—e.g., take some percentage of the peaks over the entire interval. Use threshold for zero crossings based on ZC distribution for unvoiced speech.
4. find an interval of log E_n that exceeds a high threshold ITU.
5. find a putative starting point (N_1) where log E_n crosses ITL from above; find a putative ending point (N_2) where log E_n crosses ITL from above.
6. move backwards from N_1 by comparing Z_{100} to IZCT, and find the first point where Z_{100} exceeds IZCT; similarly move forward from N_2 by comparing Z_{100} to IZCT and finding last point where Z_{100} exceeds IZCT.

Endpoint Detection Algorithm

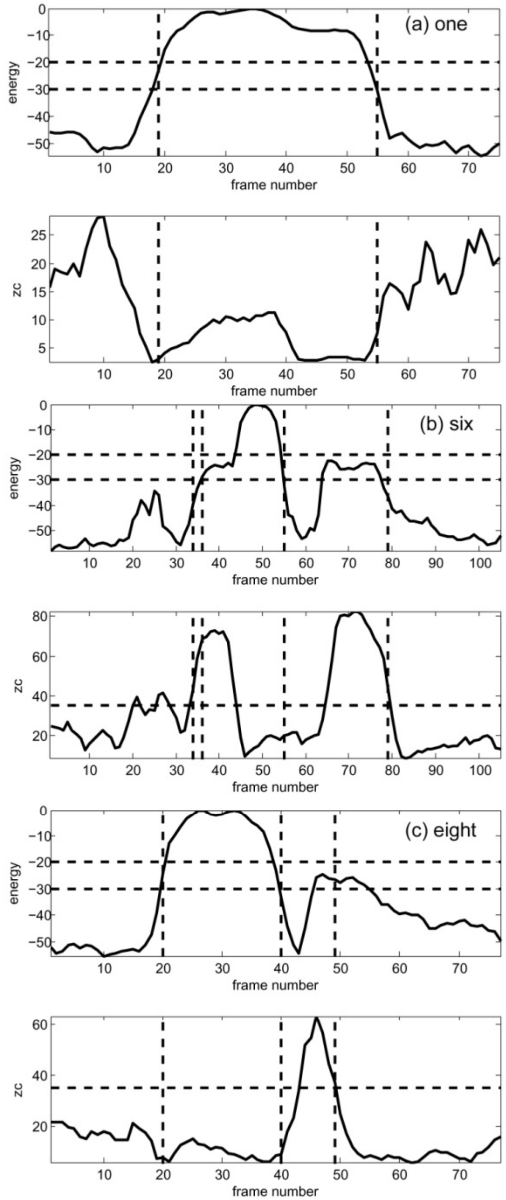


1. find heart of signal via conservative energy threshold => Interval 1
2. refine beginning and ending points using tighter threshold on energy => Interval 2
3. check outside the regions using zero crossing and unvoiced threshold => Interval 3

Endpoint Detection Algorithm



Isolated Digit Detection



Panels 1 and 2: digit /one/
- both initial and final endpoint frames determined from short-time log energy

Panels 3 and 4: digit /six/
- both initial and final endpoints determined from both short-time log energy and short-time zero crossings

Panels 5 and 6: digit /eight/
- initial endpoint determined from short-time log energy; final endpoint determined from both short-time log energy and short-time zero crossings

Algorithm #2

Voiced/Unvoiced/Background (Silence) Classification

Voiced/Unvoiced/Background Classification—Algorithm #2

- Utilize a Bayesian statistical approach to classification of frames as voiced speech, unvoiced speech or background signal (i.e., 3-class recognition/classification problem)
- Use 5 short-time speech parameters as the basic feature set
- Utilize a (hand) labeled training set to learn the statistics (means and variances for Gaussian model) of each of the 5 short-time speech parameters for each of the classes

Speech Parameters

$X = [x_1, x_2, x_3, x_4, x_5]$

$x_1 = \log E_S$ -- short-time log energy of the signal

$x_2 = Z_{100}$ -- short-time zero crossing rate of the signal for a 100-sample frame

$x_3 = C_1$ -- short-time autocorrelation coefficient at unit sample delay

$x_4 = \alpha_1$ -- first predictor coefficient of a p^{th} order linear predictor

$x_5 = E_p$ -- normalized energy of the prediction error of a p^{th} order linear predictor

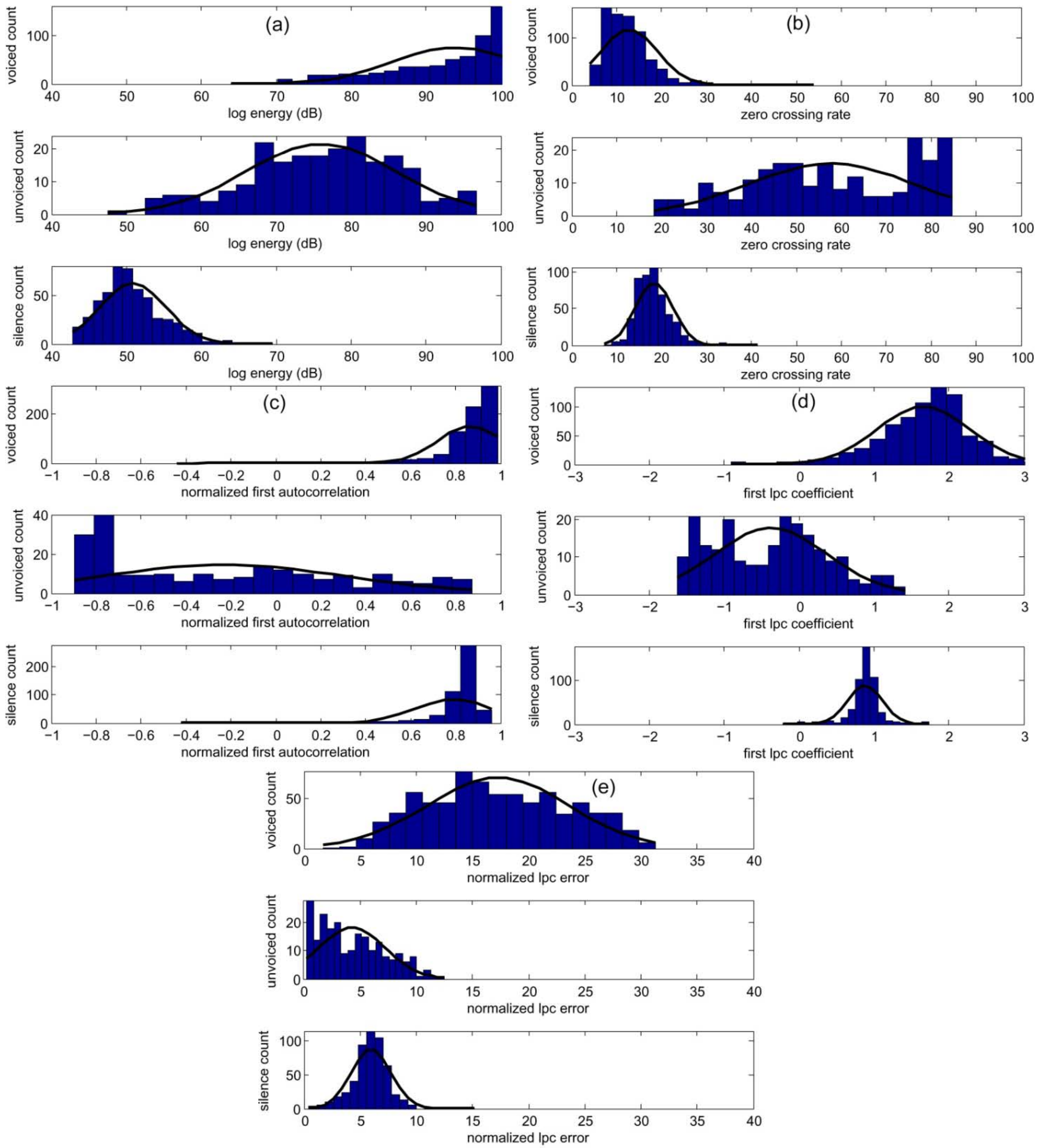
http://www.clear.rice.edu/elec532/PROJECTS00/vocode/uv/uvdet.html
The detailed calculation of Es Ep can be found; note - log Es here = Es at the above link

Speech Parameter Signal Processing

- Frame-based measurements
- Frame size of 10 msec
- Frame shift of 10 msec
- 200 Hz highpass filter used to eliminate any residual low frequency hum or dc offset in signal

Manual Training

- Using a designated training set of sentences, each 10 msec interval is classified manually (based on waveform displays and plots of parameter values) as either:
  - Voiced speech – clear periodicity seen in waveform
  - Unvoiced speech – clear indication of frication or whisper
  - Background signal – lack of voicing or unvoicing traits
  - Unclassified – unclear as to whether low level voiced, low level unvoiced, or background signal (usually at speech beginnings and endings); not used as part of the training set
- Each classified frame is used to train a single Gaussian model, for each speech parameter and for each pattern class; i.e., the mean and variance of each speech parameter is measured for each of the 3 classes



Gaussian Fits to Training Data

Bayesian Classifier

Class 1, $\omega_i, i = 1,$ representing the background signal class

Class 2, $\omega_i, i = 2,$ representing the unvoiced class

Class 3, $\omega_i, i = 3,$ representing the voiced class

$\mathbf{m}_i = E[x]$ for all x in class ω_i

$W_i = E[(x - \mathbf{m}_i)(x - \mathbf{m}_i)^T]$ for all x in class ω_i

Bayesian Classifier

Maximize the probability:

$$p(\omega_i \mid x) = \frac{p(x \mid \omega_i) \cdot P(\omega_i)}{p(x)}$$

where

$$p(x) = \sum_{i=1}^{3} p(x \mid \omega_i) \cdot P(\omega_i)$$

$$p(x \mid \omega_i) = \frac{1}{(2\pi)^{5/2} \mid W_i \mid^{1/2}} e^{-(1/2)(x-\mathbf{m}_i)^T W_i^{-1} (x-\mathbf{m}_i)}$$

Bayesian Classifier

Maximize $p(\omega_i \mid x)$ using the monotonic discriminant function

$$\begin{aligned} g_i(x) &= \ln p(\omega_i \mid x) \\ &= \ln[p(x \mid \omega_i) \cdot P(\omega_i)] - \ln p(x) \\ &= \ln p(x \mid \omega_i) + \ln P\left(\omega_i\right) - \ln p(x) \end{aligned}$$

Disregard term $\ln p(x)$ since it is independent of class, ω_i, giving

$$g_i(x) = -\frac{1}{2}(x - \mathbf{m}_i)^T W_i^{-1}(x - \mathbf{m}_i) + \ln P(\omega_i) + c_i$$

$$c_i = -\frac{5}{2}\ln(2\pi) - \frac{1}{2}\ln |W_i|$$

Bayesian Classifier

- Ignore bias term, c_i, and apriori class probability, $\ln P_i$. Then we can convert maximization to a minimization by reversing the sign, giving the decision rule:

Decide class ω_i if and only if

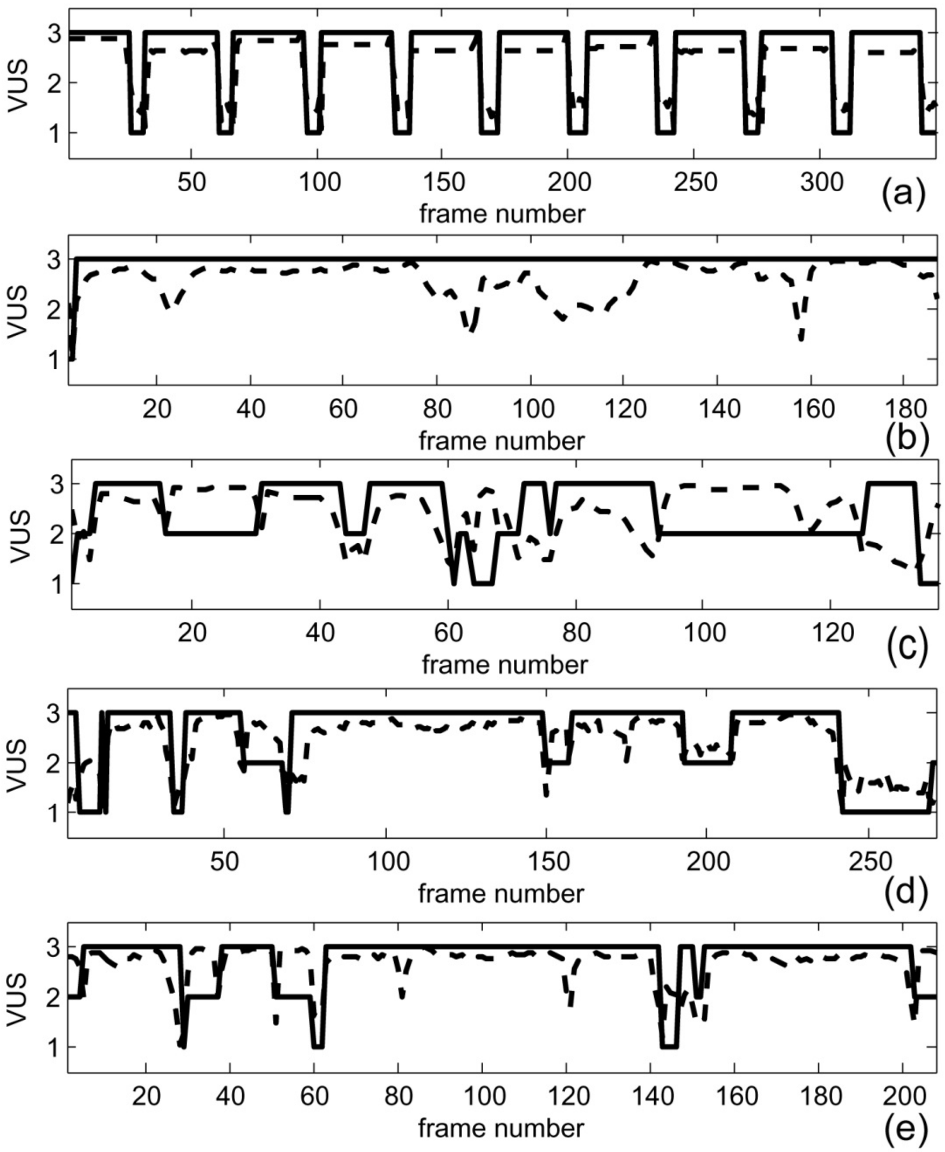
$$d_i(x) = (x - m_i)^T W_i^{-1} (x - m_i) \leq d_j(x) \;\forall\; j \neq i$$

- Utilize confidence measure, based on relative decision scores, to enable a no-decision output when no reliable class information is obtained.

Classification Performance

	Training Set	Count	Testing Set	Count
Background-Class 1	85.5%	76	96.8%	94
Unvoiced – Class 2	98.2%	57	85.4%	82
Voiced – Class 3	99%	313	98.9%	375

VUS Classifications



Panel (a): synthetic vowel sequence

Panel (b): all voiced utterance

Panels (c-e): speech utterances with a mixture of regions of voiced speech, unvoiced speech and background signal (silence)

DEEE725 Speech Signal Processing Lab

Gil-Jin Jang

END OF LECTURE 07 SPEECH/NON-SPEECH DETECTION AND END-POINT DETECTION (EPD)