Chapter 10

Algorithms for Estimating Speech Parameters

语音参数估计算法

Speech Processing Algorithms

- Speech/Non-speech detection
  - Rule-based method using log energy and zero crossing rate
  - Single speech interval in background noise
- Voiced/Unvoiced/Background classification
  - Bayesian approach using 5 speech parameters
  - Needs to be trained (mainly to establish statistics for background signals)
- F0 detection
  - Estimation of fundamental frequency (F0) during regions of voiced speech
  - Implicitly needs classification of signal as voiced speech
  - Algorithms in time domain, frequency domain, cepstral domain, or using LPC-based processing methods
- Formant estimation
  - Estimation of the frequencies of the major resonances during voiced speech regions
  - Implicitly needs classification of signal as voiced speech
  - Need to handle birth and death processes as formants appear and disappear depending on spectral intensity

Algorithm #1

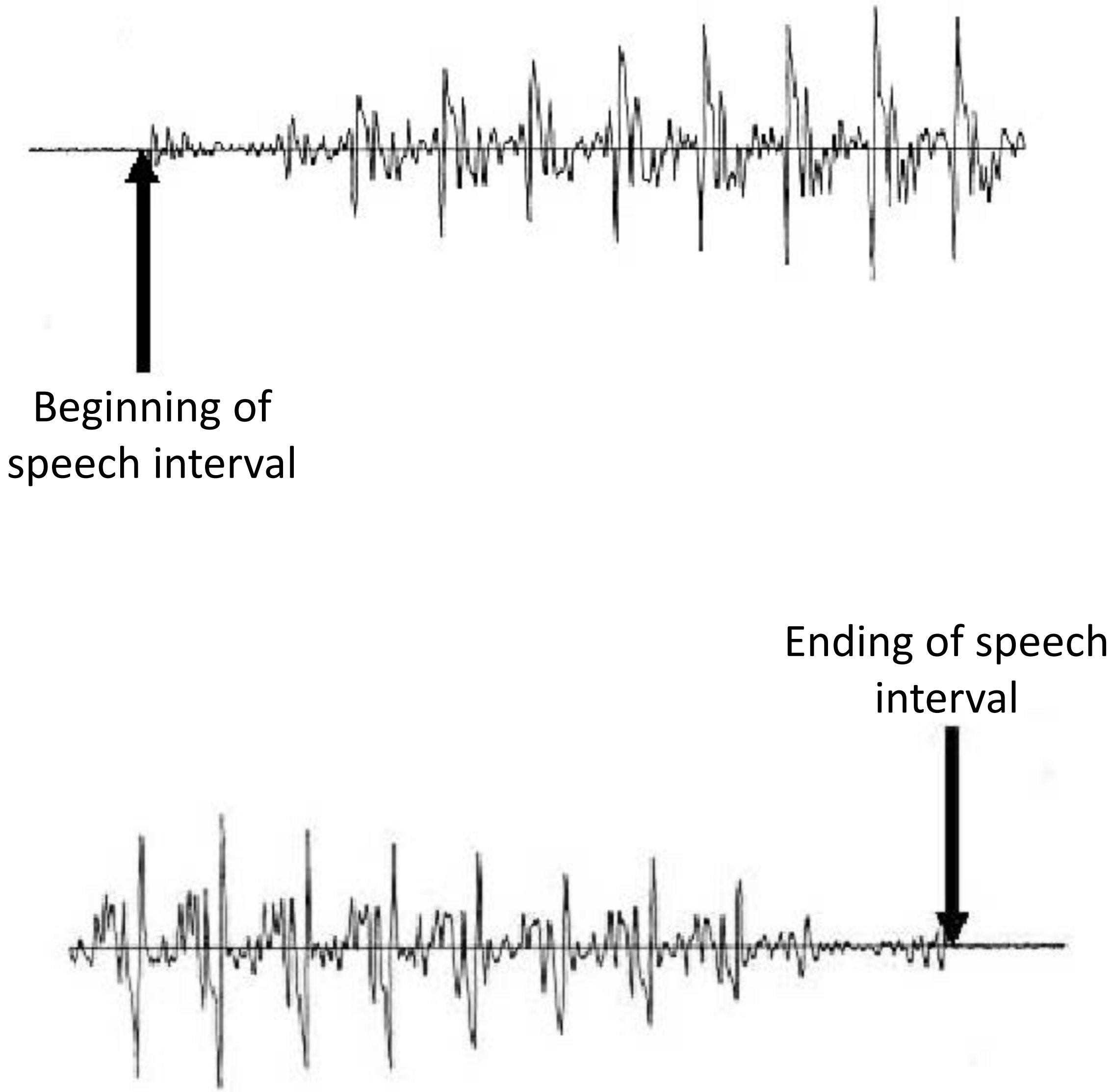
Speech/Non-Speech Detection
Using Simple Rules

Speech Detection Issues

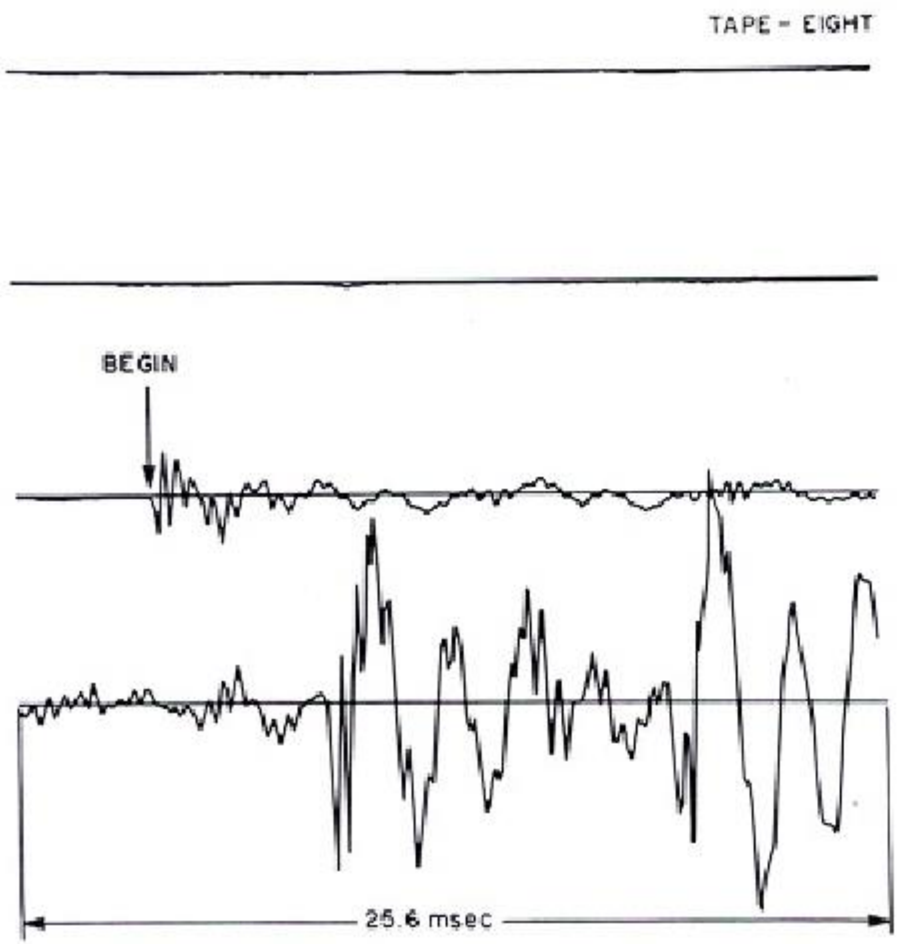
- key problem in speech processing is locating accurately the beginning and end of a speech utterance in noise/background signal

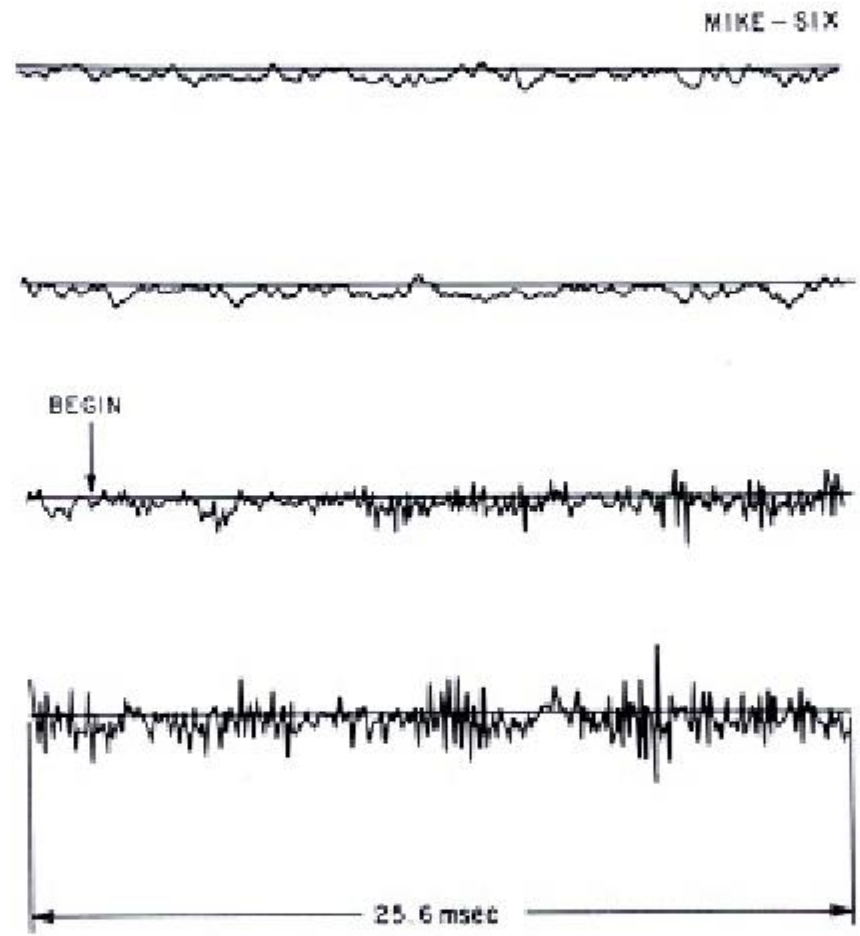
- need endpoint detection to enable:
  – computation reduction (don't have to process background signal)
  – better recognition performance (can't mistake background for speech)
  – non-trivial problem except for high SNR recordings

Ideal Speech/Non-Speech Detection



Speech Detection Examples

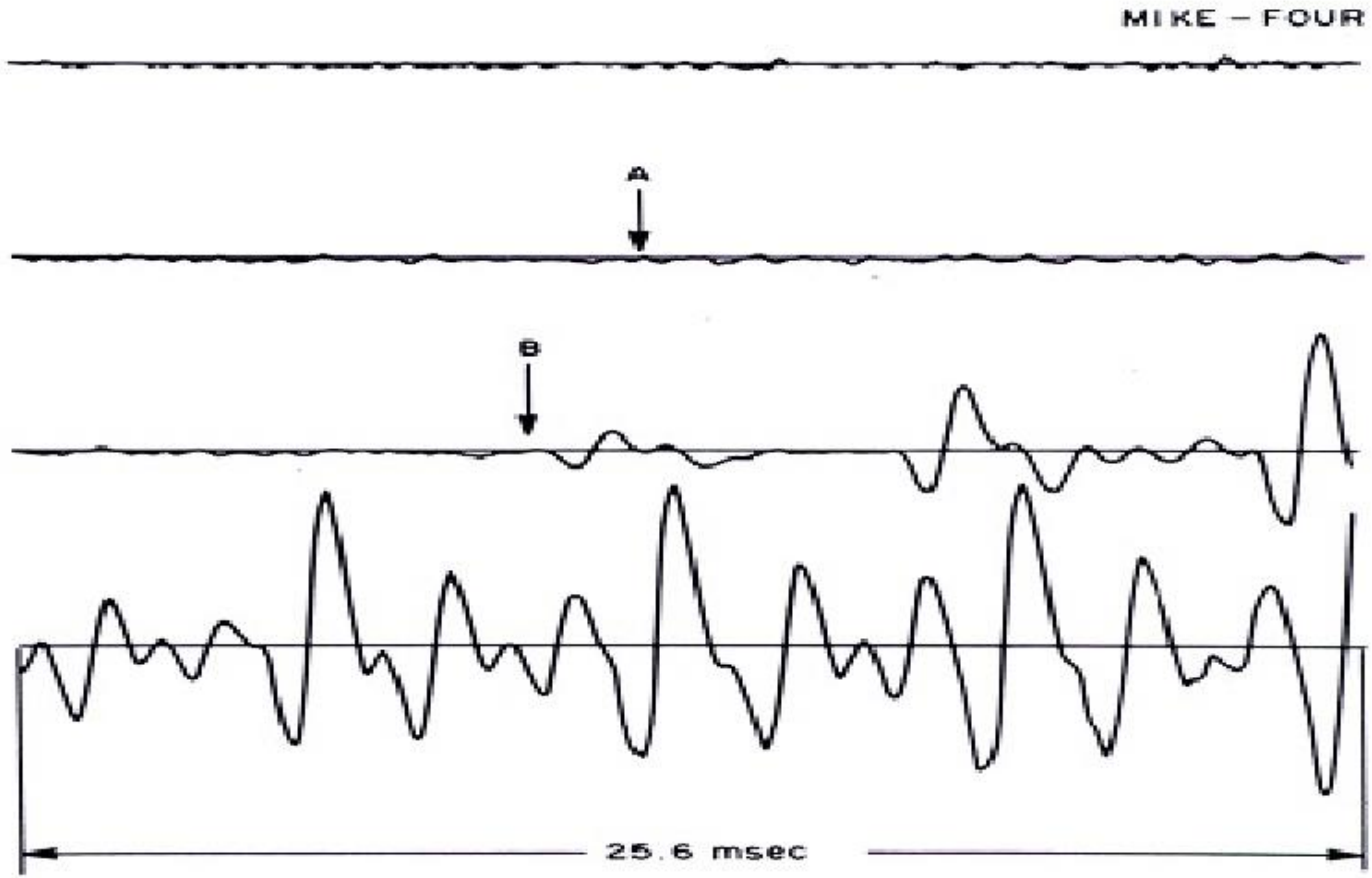




case of low background noise => simple case

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

Speech Detection Examples

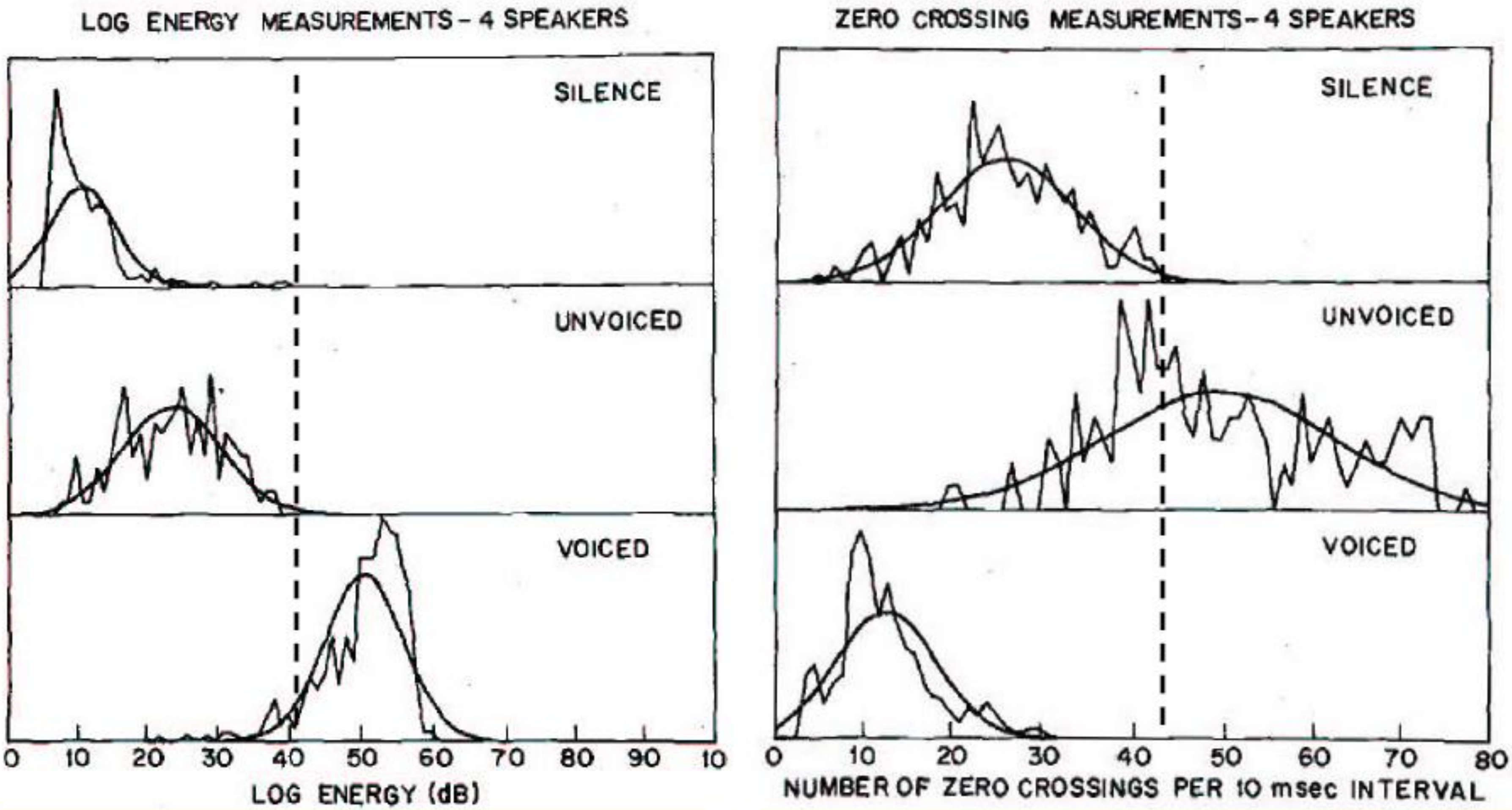


difficult case because of weak fricative sound, /f/, at beginning of speech

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 (逐渐减小) of vowel sounds at end of utterance

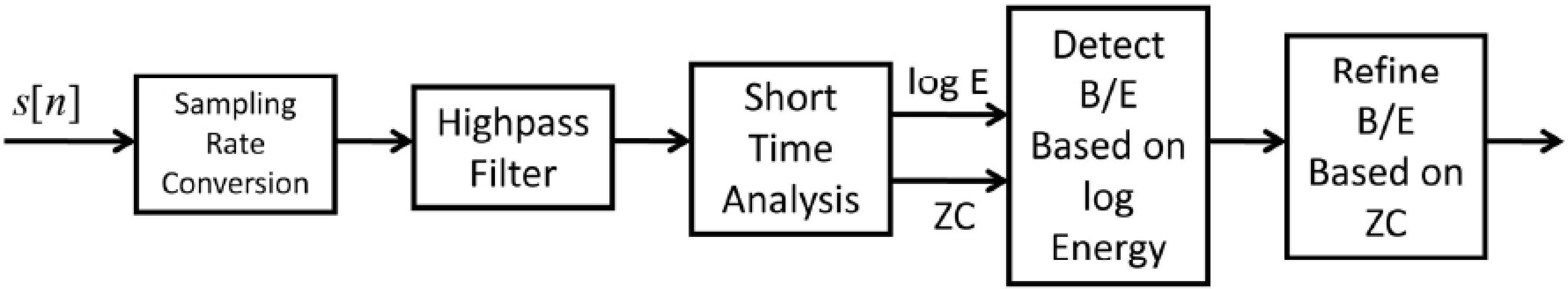
Speech/Non-Speech Detection



Log energy separates Voiced from Unvoiced and Silence

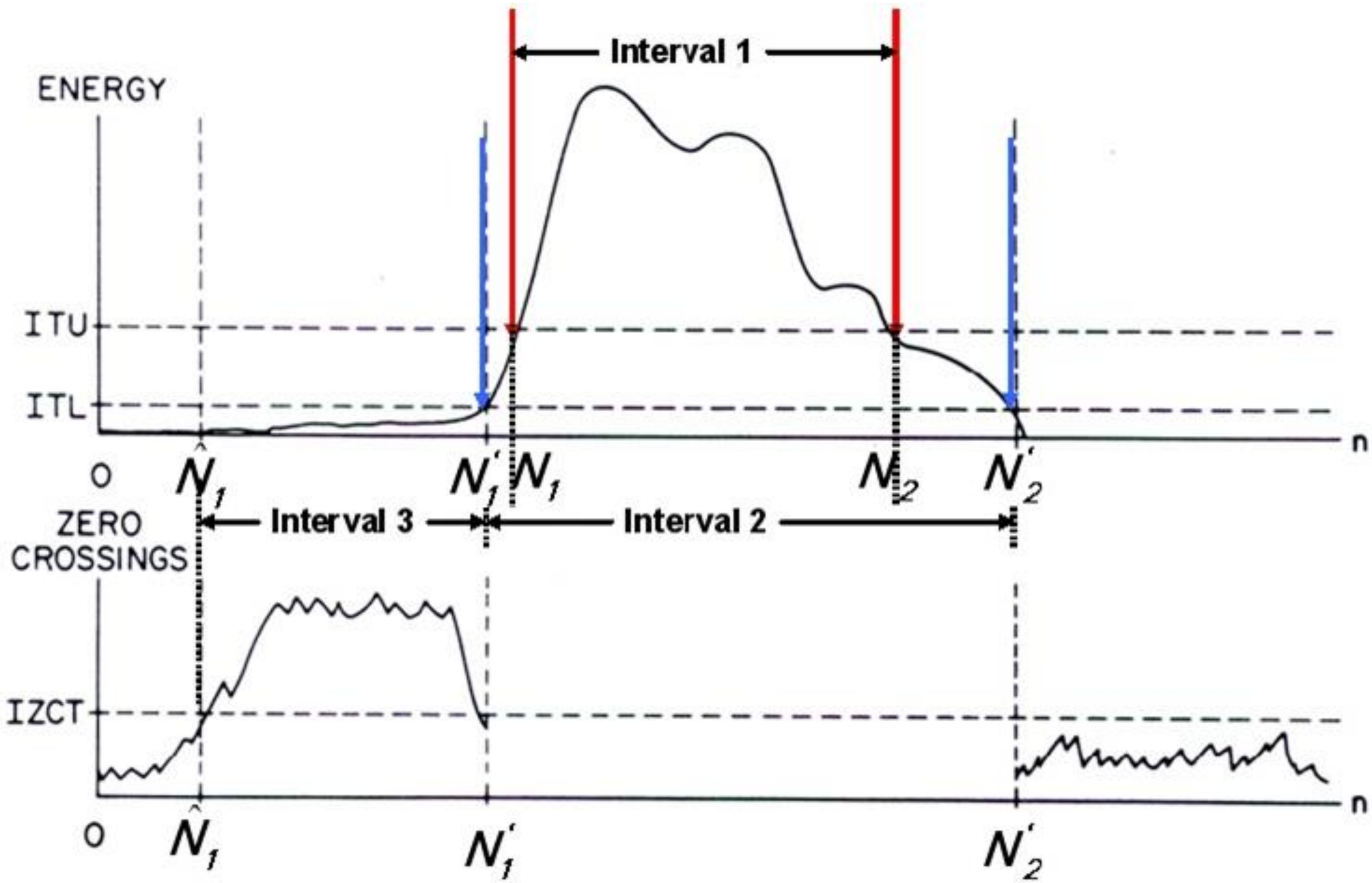
Zero crossings separate Unvoiced from Silence and Voiced

Speech/Non-Speech Detection



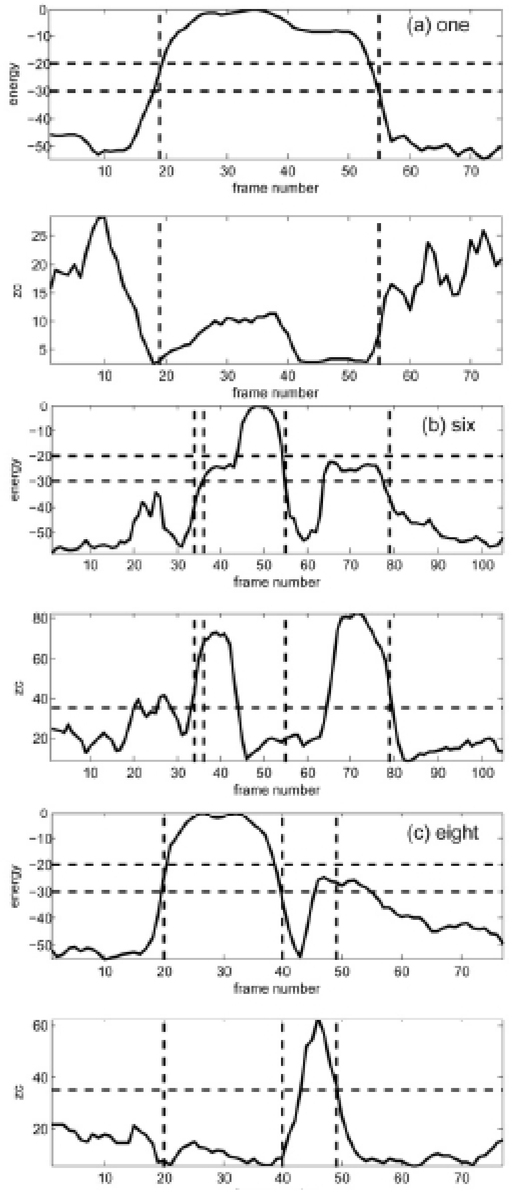
- sampling rate conversion to standard rate (10 kHz)
- highpass filtering to eliminate DC offset and hum
- short-time analysis using frame size of 40 msec, with a frame shift of 10 msec; compute short-time log energy and short-time zero crossing rate
- detect beginning and ending frames based entirely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

Endpoint Detection Algorithm



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

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

Bayesian Classifier

- Class definition

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

- Feature extraction: vector *x* for each frame
- Distribution estimation

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

$\mathbf{m}_i = E[x]$ 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

- Make decision by maximizing the probability

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

where

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

Feature Extraction

$X = [x_1, x_2, x_3, x_4, x_5]$ feature vector for each frame, including

$x_1 = \log E_S$ -- 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

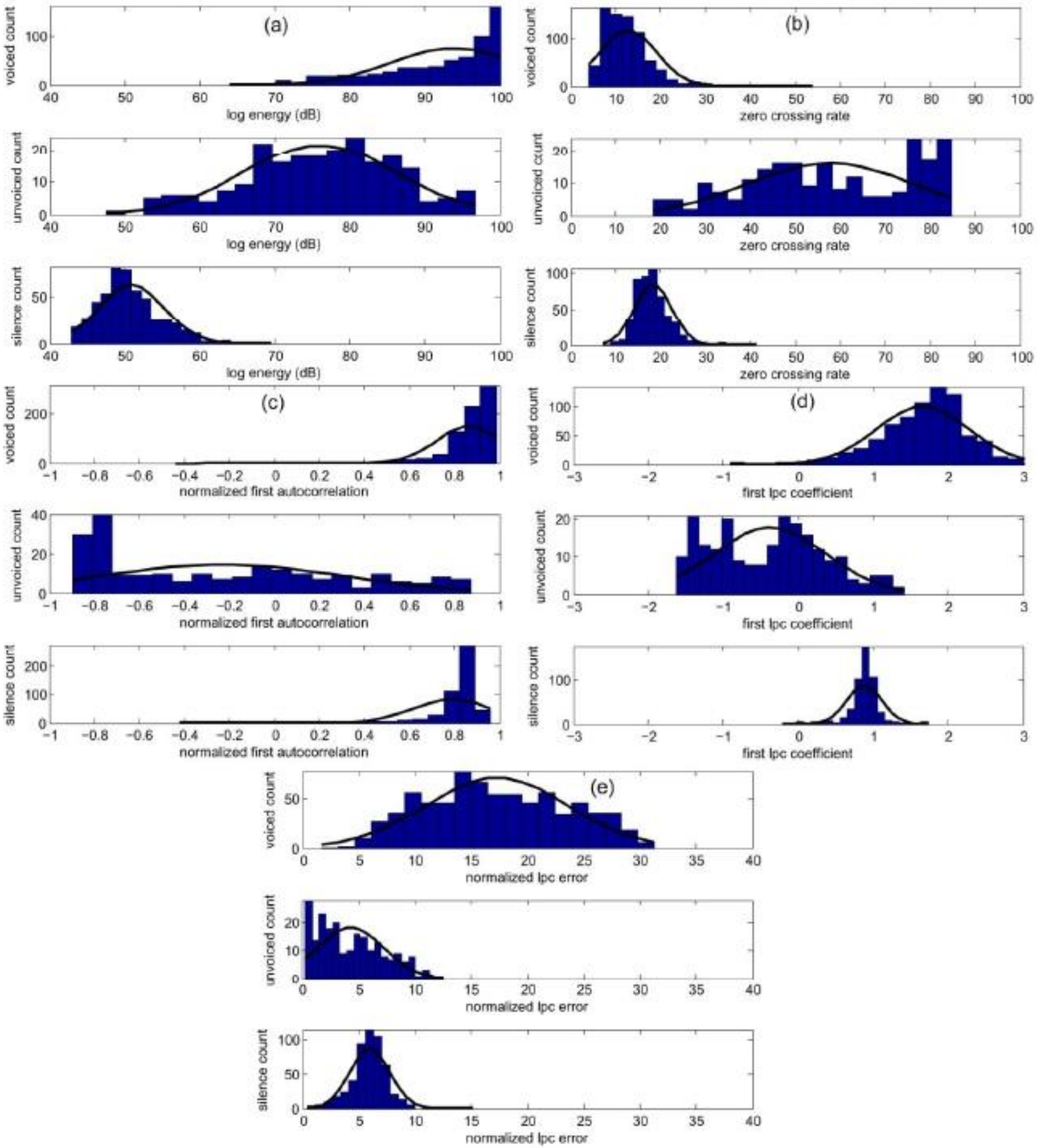
$x_5 = E_p$ -- normalized energy of the prediction error of a p^{th} order linear predictor

Feature Extraction

- Frame-based measurements
- Frame size of 40 msec (10kHz sampling rate)
- Frame shift of 10 msec
- 200 Hz highpass filter used to eliminate any residual low frequency hum or DC offset in signal

Distribution Estimation

- 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

Make Decision

- 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

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

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

Make Decision

- Ignore bias term, c_i , and a priori class probability, $P(\omega_i)$. Then we can convert maximization to a minimization by reversing the sign, giving the decision rule:

  Decide class ω_i if and only if

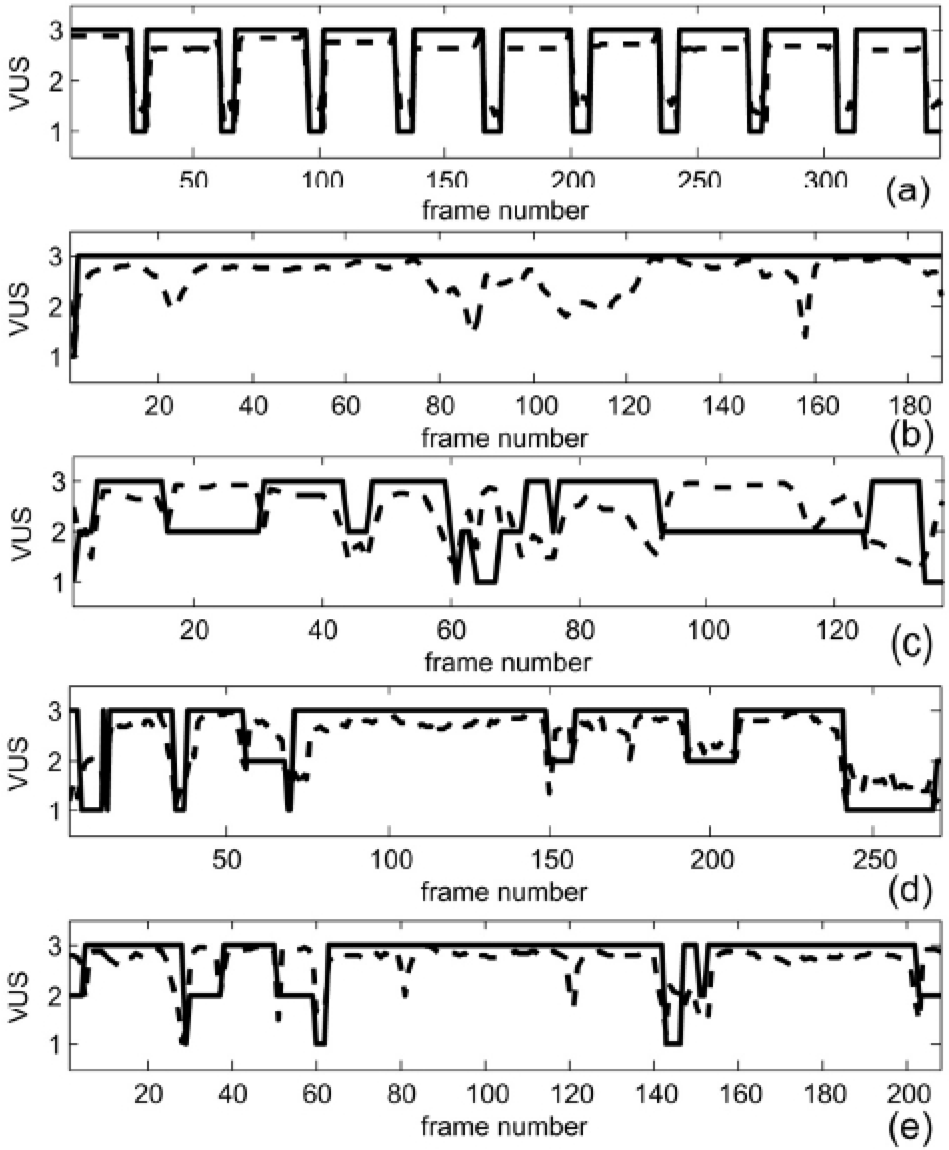
  $$d_i(x) = (x - m_i)^T W_i^{-1} (x - m_i) \leq d_j(x) \;\forall\; j \neq i$$

- Utilizing 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 “we were away a year ago”

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

The solid line indicates decision and the dashed line indicates the corresponding confidence measure (multiplied by 3 for plotting)

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

Algorithm #3

F0 Detection

(F0 Period Estimation Methods)

F0 Period Estimation

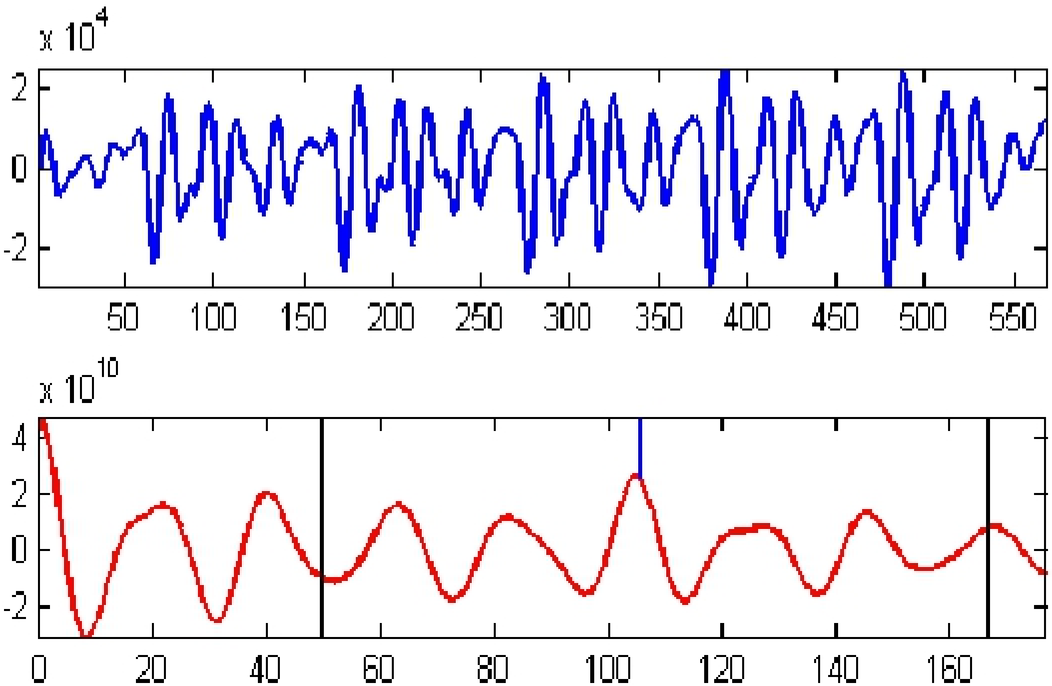
- Essential component of general synthesis model for speech production
- Major component of excitation source information (along with voiced-unvoiced decision, amplitude)
- F0 period estimation involves two problems, simultaneously; determination as to whether the speech is periodic, and, if so, the resulting F0 (period or frequency)
- A range of F0 detection methods have been proposed including several time domain/frequency domain/cepstral domain/LPC domain methods

Autocorrelation Method of F0 Detection

Autocorrelation F0 Detection

- basic principle – a periodic function has a periodic autocorrelation –just find the correct peak
- basic problem – the autocorrelation representation of speech is just too rich
  - it contains information that enables you to estimate the vocal tract transfer function (from the first 10 or so values)
  - many peaks in autocorrelation in addition to F0 periodicity peaks
  - some peaks due to rapidly changing formants
  - some peaks due to window size interactions with the speech signal
- need some type of spectrum flattening so that the speech signal more closely approximates a periodic impulse train => center clipping (中心削波) spectrum flattener

Autocorrelation of Voiced Speech Frame

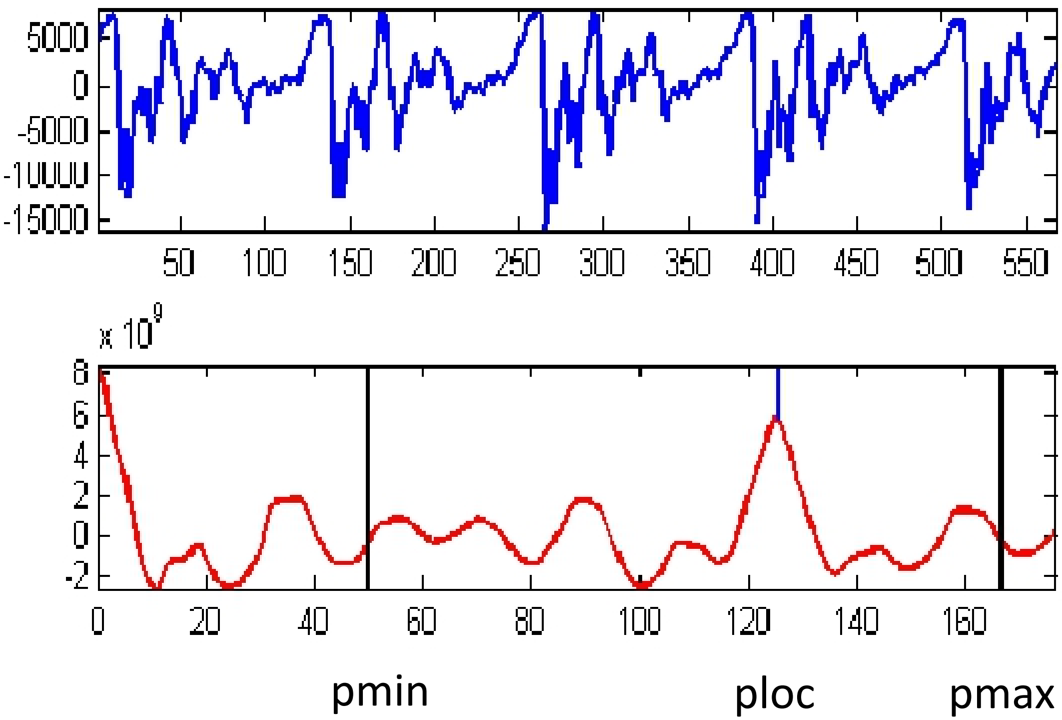


$\tilde{x}[n],\ n = 0, 1, \ldots, 559$

$\tilde{R}[k],\ k = 0, 1, \ldots, \text{pmax} + 10$

pmin ploc pmax

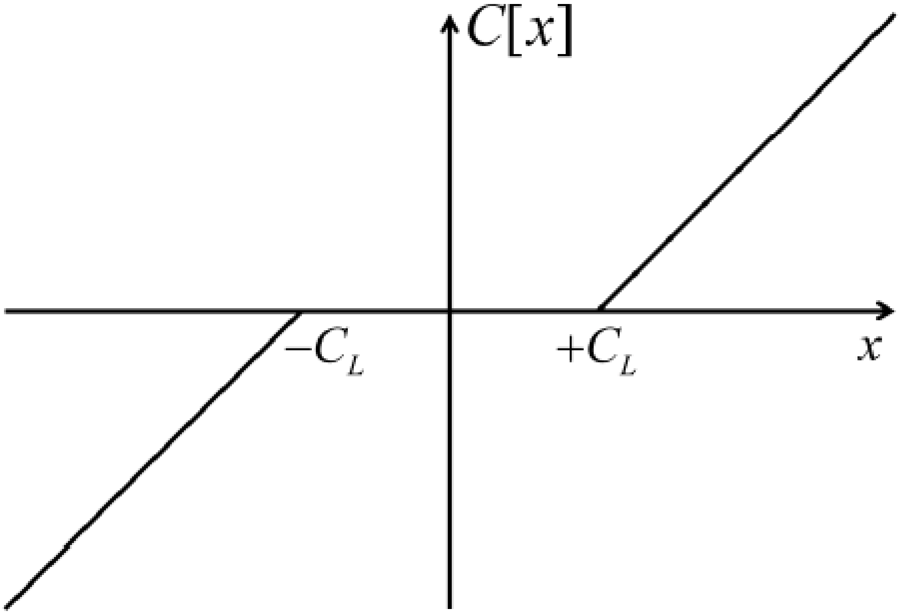
Autocorrelation of Voiced Speech Frame

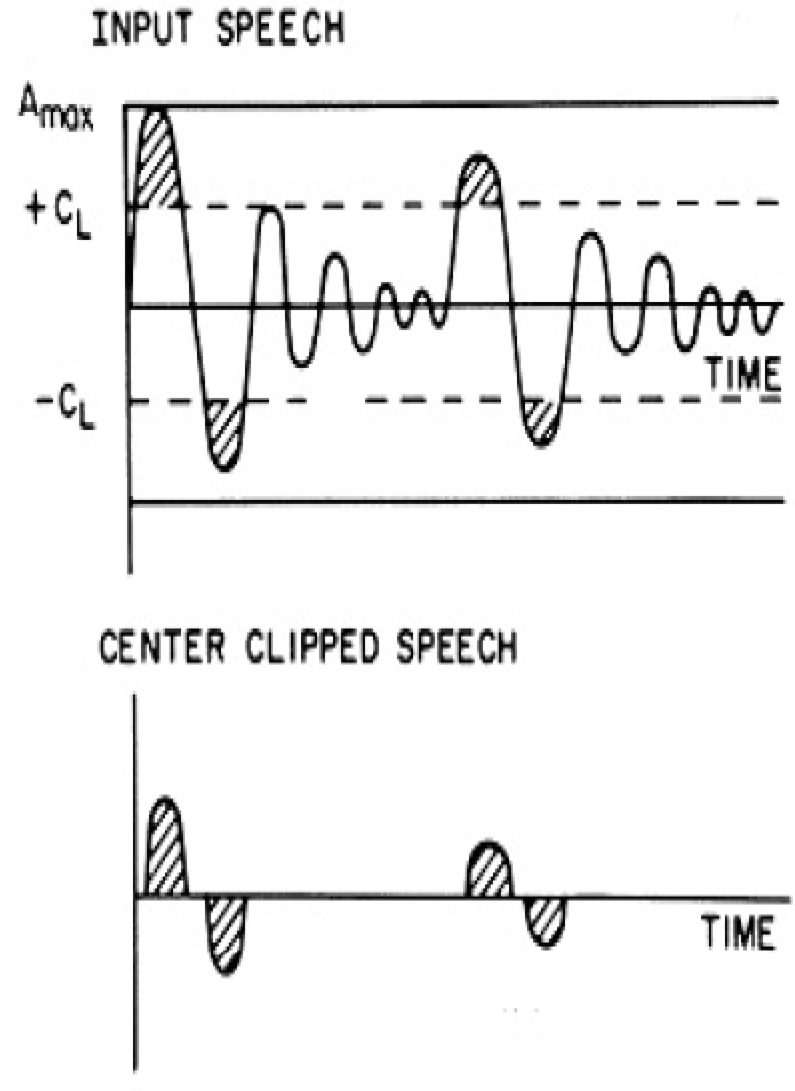


$$\tilde{x}[n],\ n = 0, 1, \ldots, 559$$

$$\tilde{R}[k],\ k = 0, 1, \ldots, \text{pmax} + 10$$

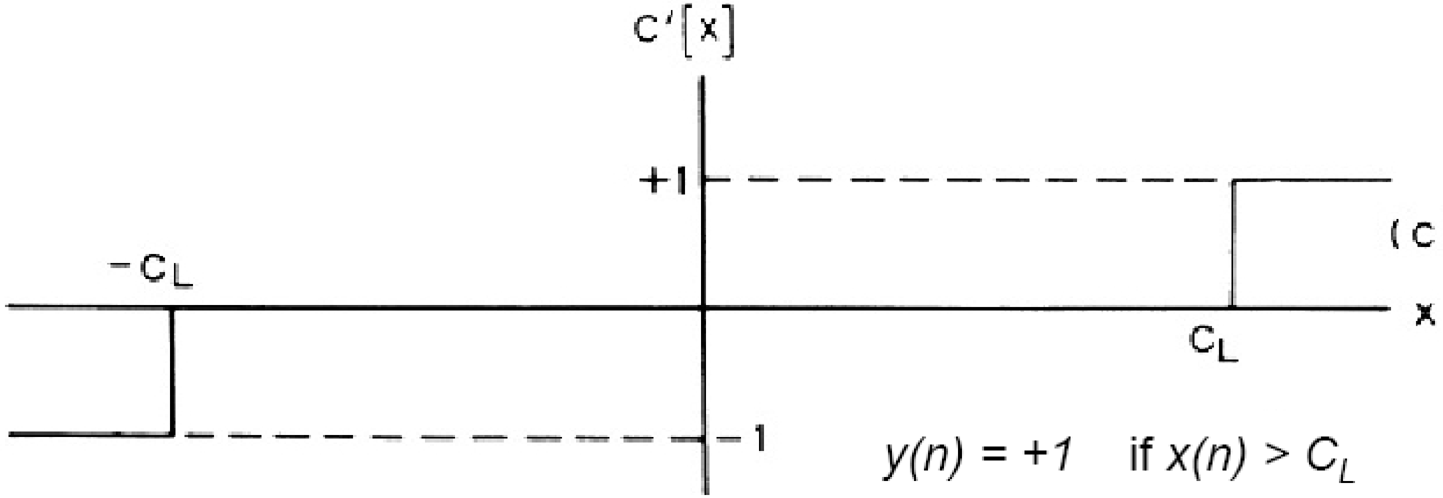
Center Clipping





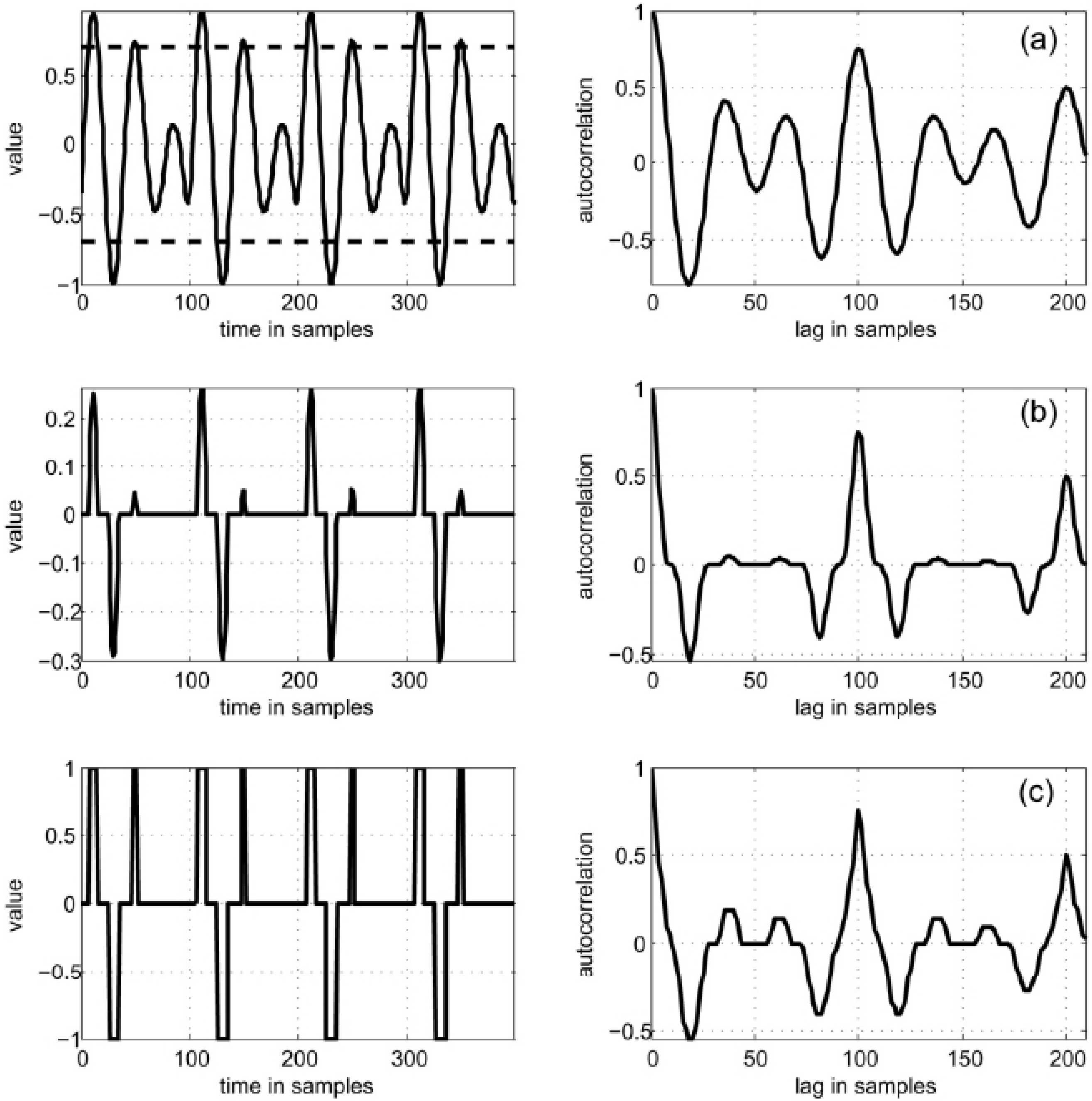
- C_L=% of A_{max} (e.g., 30%)
- Center Clipper definition:
  - if $x(n) > C_L$, $y(n)=x(n)-C_L$
  - if $x(n) \le C_L$, $y(n)=0$

3-Level Center Clipper



$$y(n) = \begin{cases} +1 & \text{if } x(n) > C_L \\ -1 & \text{if } x(n) < -C_L \\ 0 & \text{otherwise} \end{cases}$$

- significantly simplified computation (no multiplications)
- autocorrelation function is very similar to that from a conventional center clipper => most of the extraneous peaks are eliminated and a clear indication of periodicity is retained



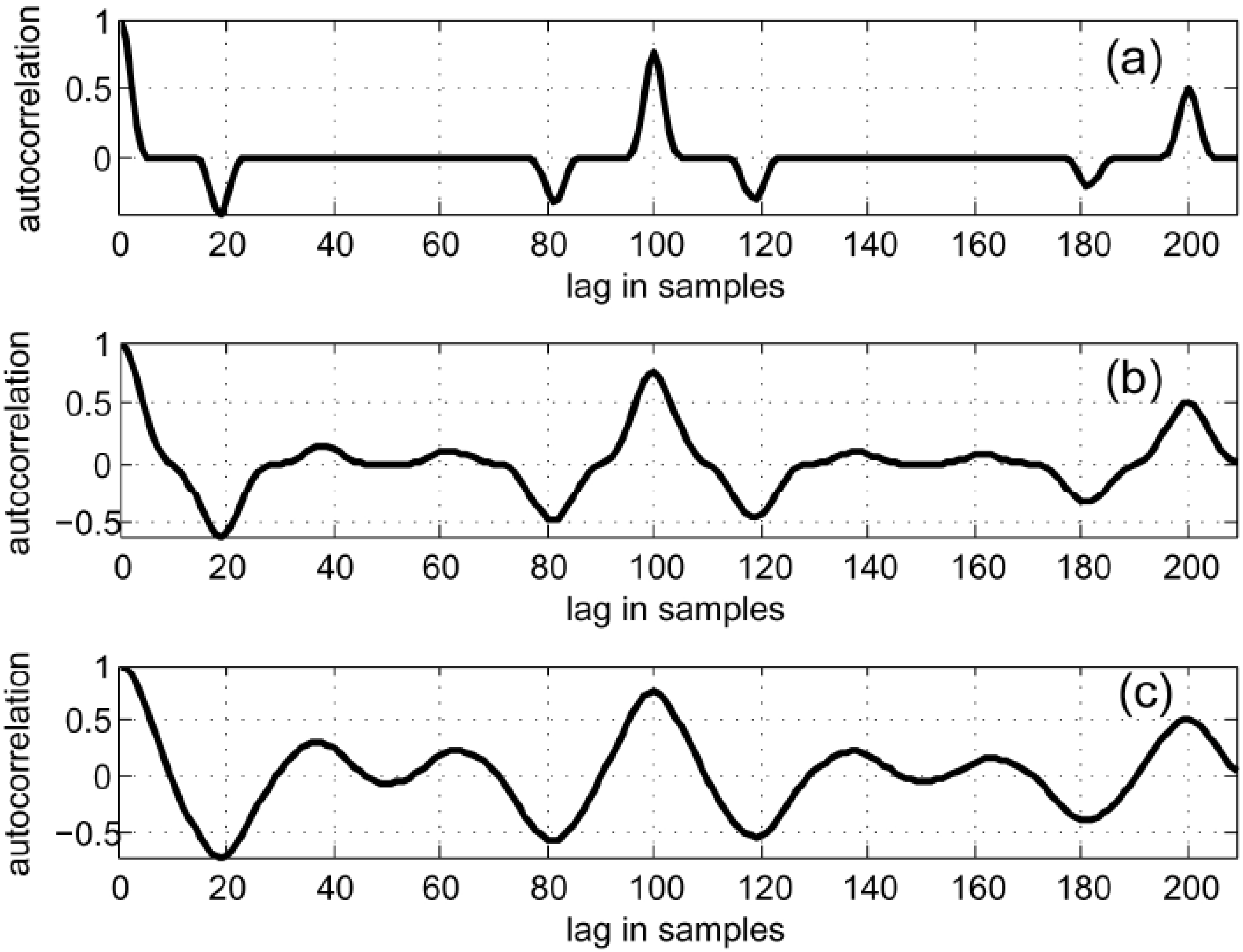
Waveforms and Autocorrelations

First row: no clipping (dashed lines show 70% clipping level)

Second row: center clipped at 70% threshold

Third row: 3-level center clipped

Autocorrelations of Center-Clipped Speech



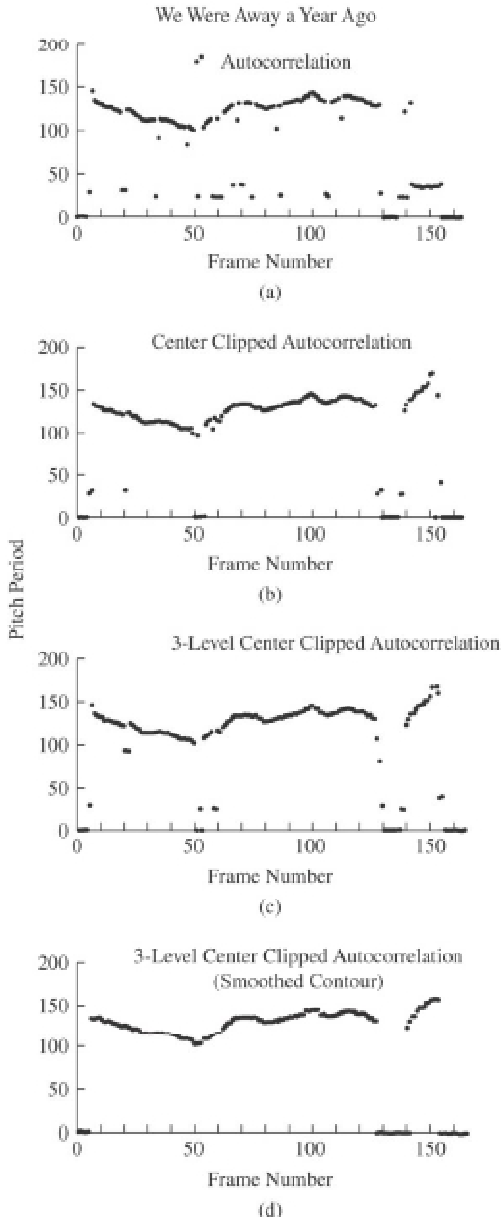
Clipping Level:

(a) 90%

(b) 60%

(c) 30%

Autocorrelation Pitch Detector



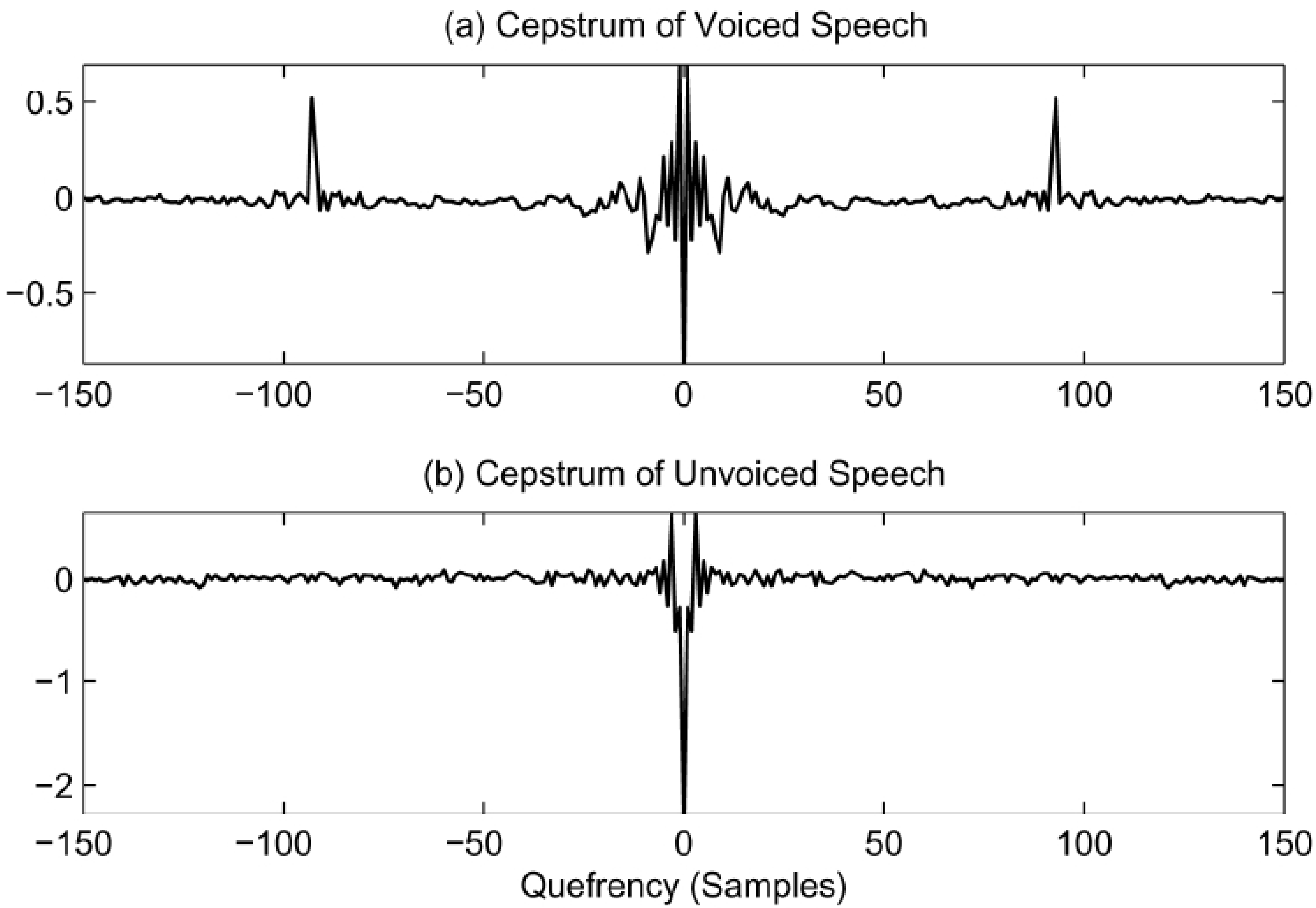
- lots of errors with conventional autocorrelation—especially short lag estimates of pitch period
- center clipping eliminates most of the gross errors
- nonlinear smoothing fixes the remaining errors

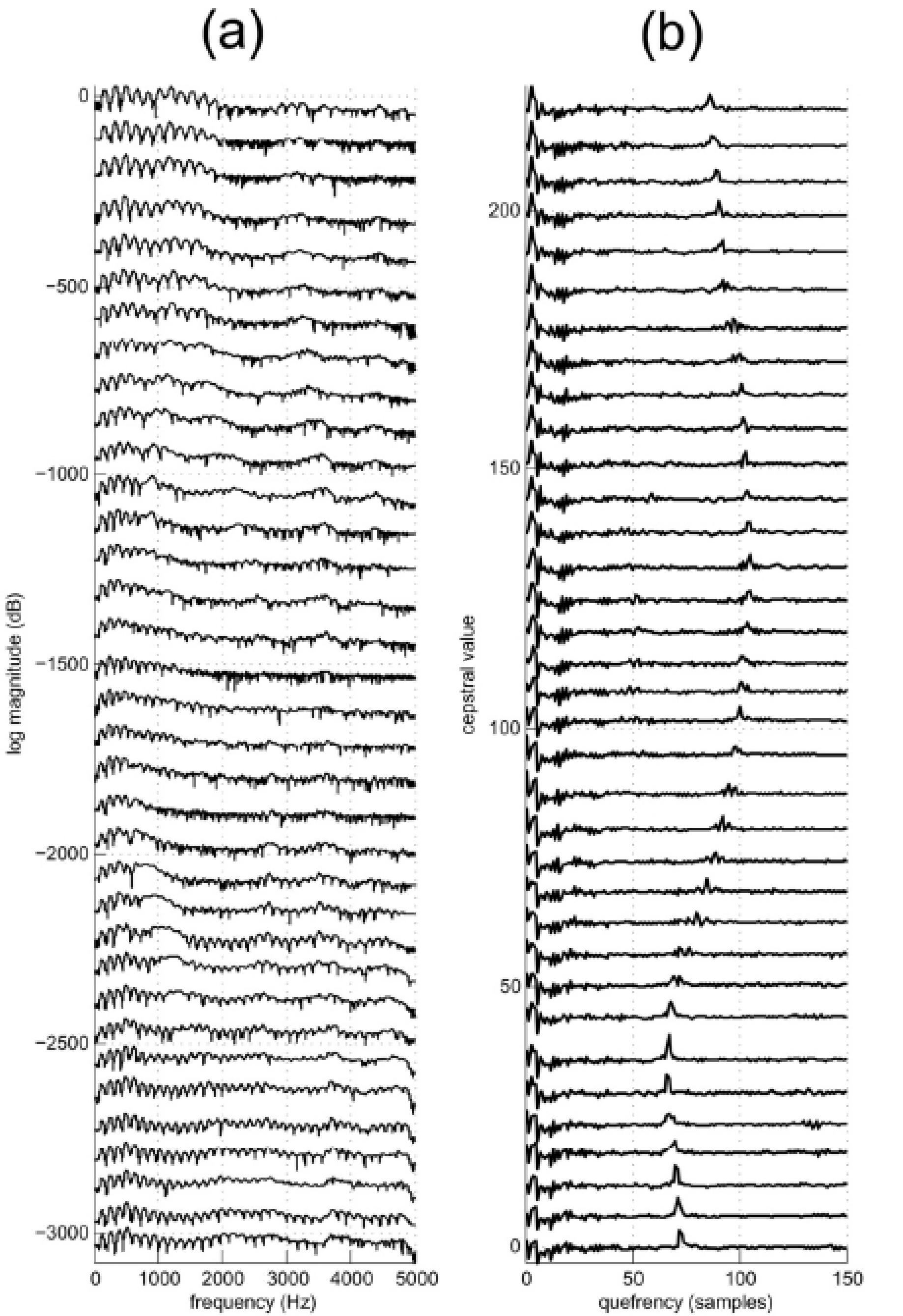
Cepstral F0 Detector

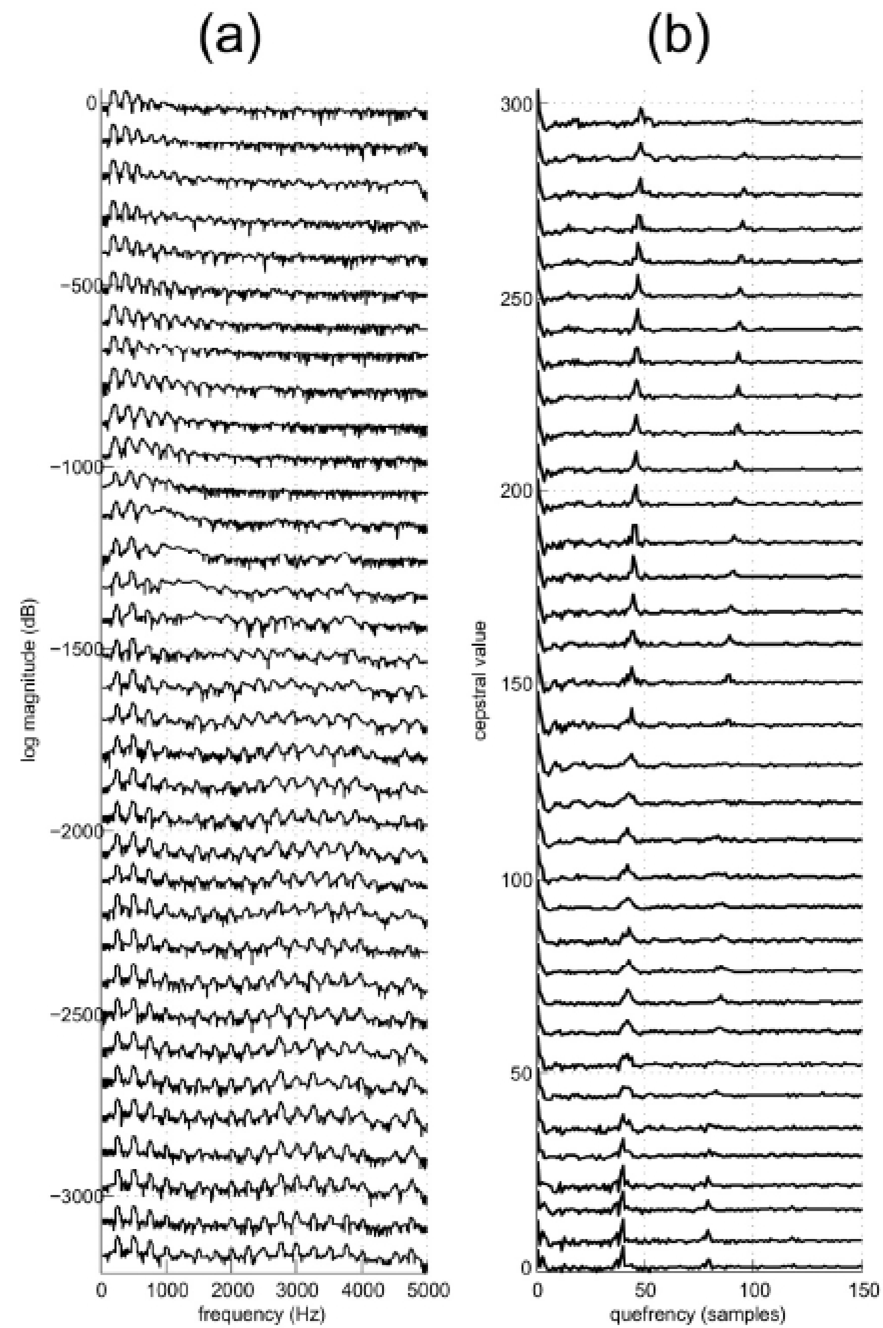
Cepstral F0 Detection

- simple procedure for cepstral F0 detection
  1. compute cepstrum every 10-20 msec
  2. search for periodicity peak in expected range of *n*
  3. if found and above threshold => voice, F0 period =location of cepstral peak
  4. if not found => unvoiced

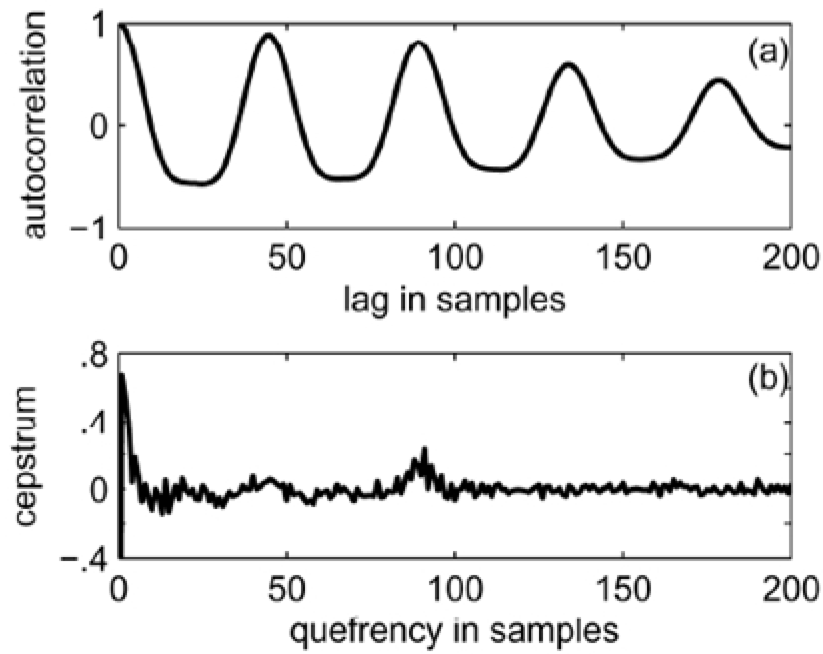
Cepstral Sequences for Voiced and Unvoiced Speech




(a)
(b)
log magnitude (dB)
0
-500
-1000
-1500
-2000
-2500
-3000
0 1000 2000 3000 4000 5000
frequency (Hz)
cepstral value
0
50
100
150
200
0 50 100 150
quefrency (samples)


(a)
(b)
log magnitude (dB)
0
-500
-1000
-1500
-2000
-2500
-3000
0 1000 2000 3000 4000 5000
frequency (Hz)
cepstral value
0
50
100
150
200
250
300
0 50 100 150
quefrency (samples)

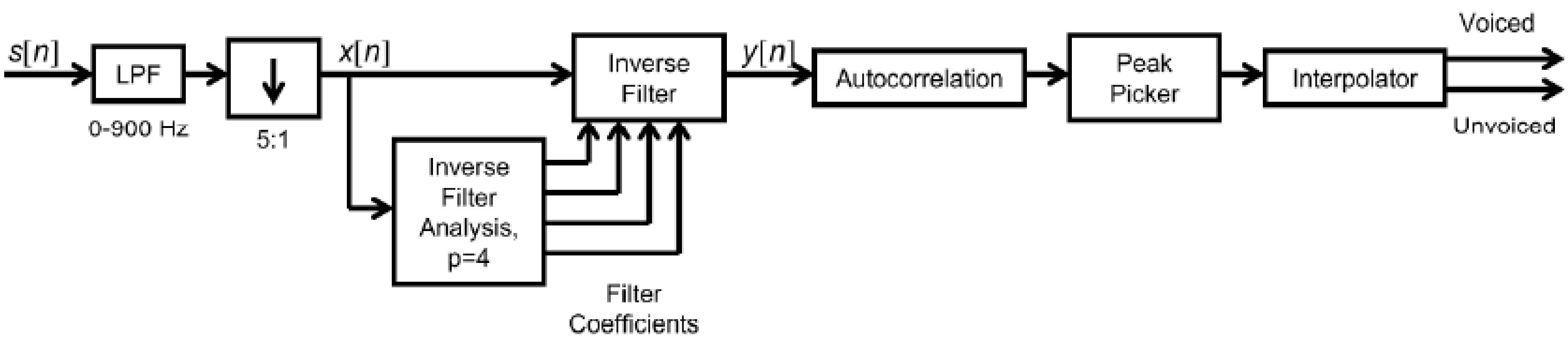
Comparison of Cepstrum and ACF



Pitch doubling errors eliminated in cepstral display, but not in autocorrelation display. Weak cepstral peaks still stand out in cepstral display.

LPC-Based F0 Detector

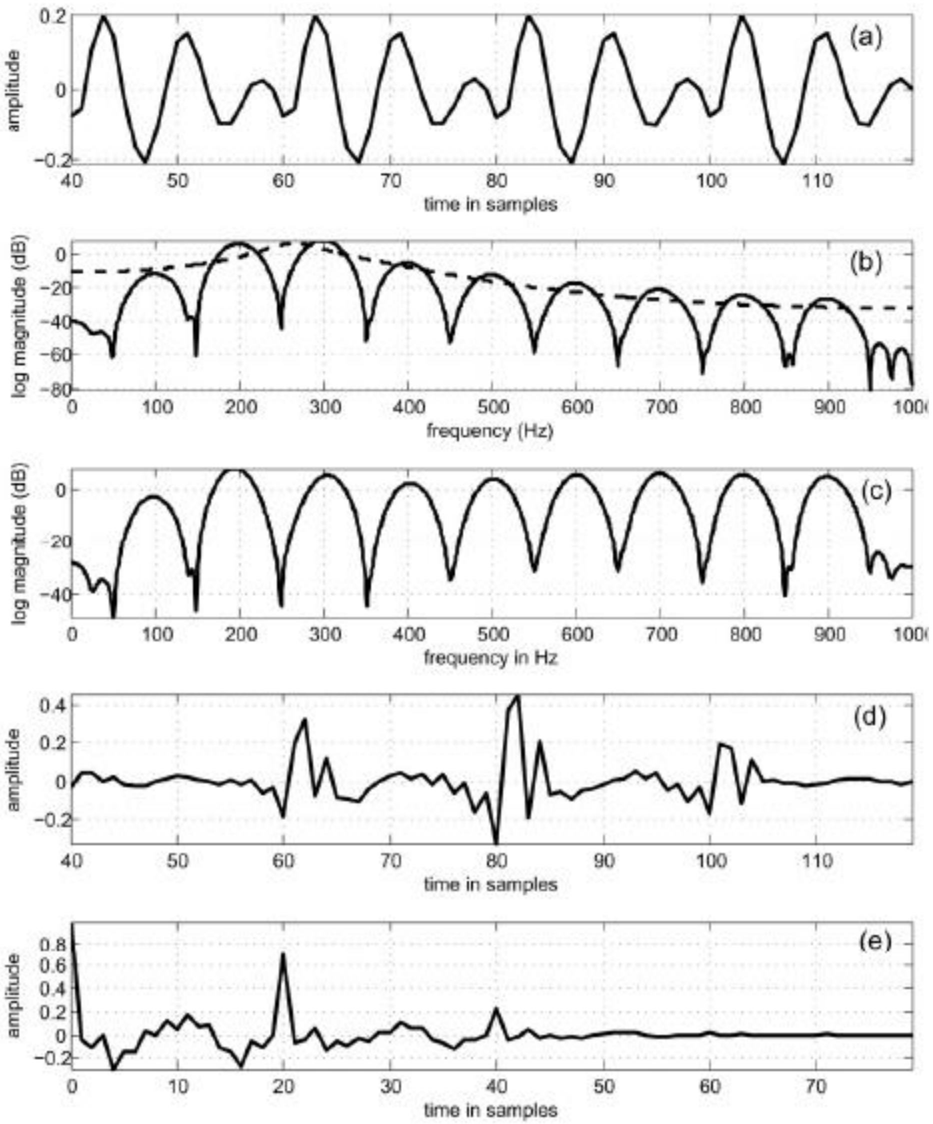
LPC F0 Detection



Simple Inverse Filtering Track

- sampling rate reduced from 10 kHz to 2 kHz
- *p=4* analysis
- inverse filter signal to give spectrally flat result
- compute short time autocorrelation and find strongest peak in estimated pitch region

LPC F0 Detection



- part a: section of input waveform being analyzed
- part b: input spectrum and reciprocal of the inverse filter
- part c: spectrum of signal at output of the inverse filter
- part d: time waveform at output of the inverse filter
- part e: normalized autocorrelation of the signal at the output of the inverse filter

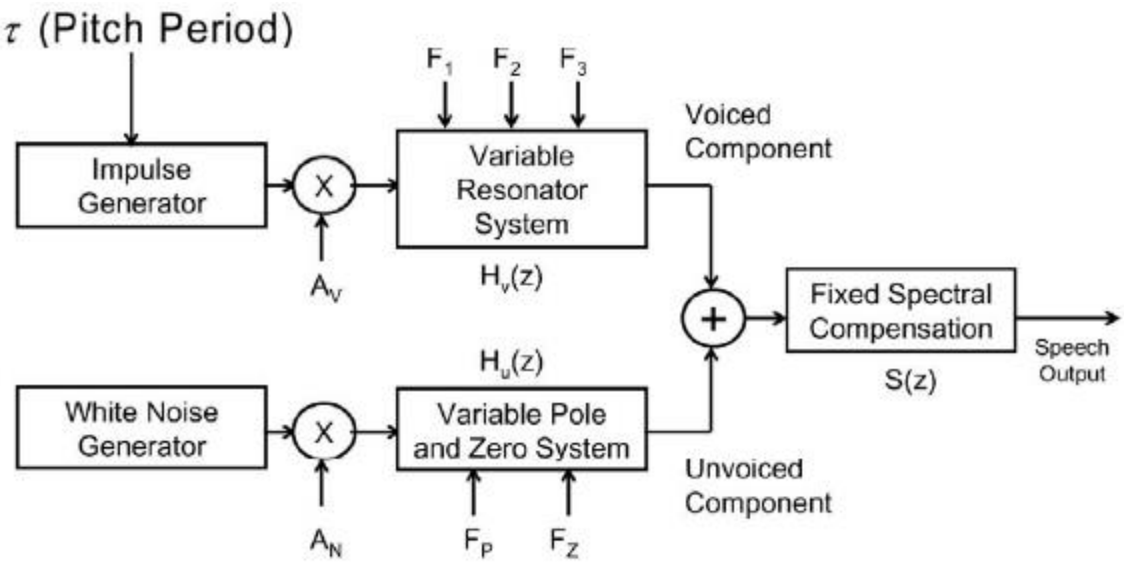
Algorithm #4 – Formant Estimation

Cepstral-Based Formant Estimation

Cepstral Formant Estimation

- the low-time cepstrum corresponds primarily to the combination of vocal tract, glottal pulse, and radiation, while the high time part corresponds primarily to excitation

  => use lowpass liftered cepstrum to give smoothed log spectra to estimate formants

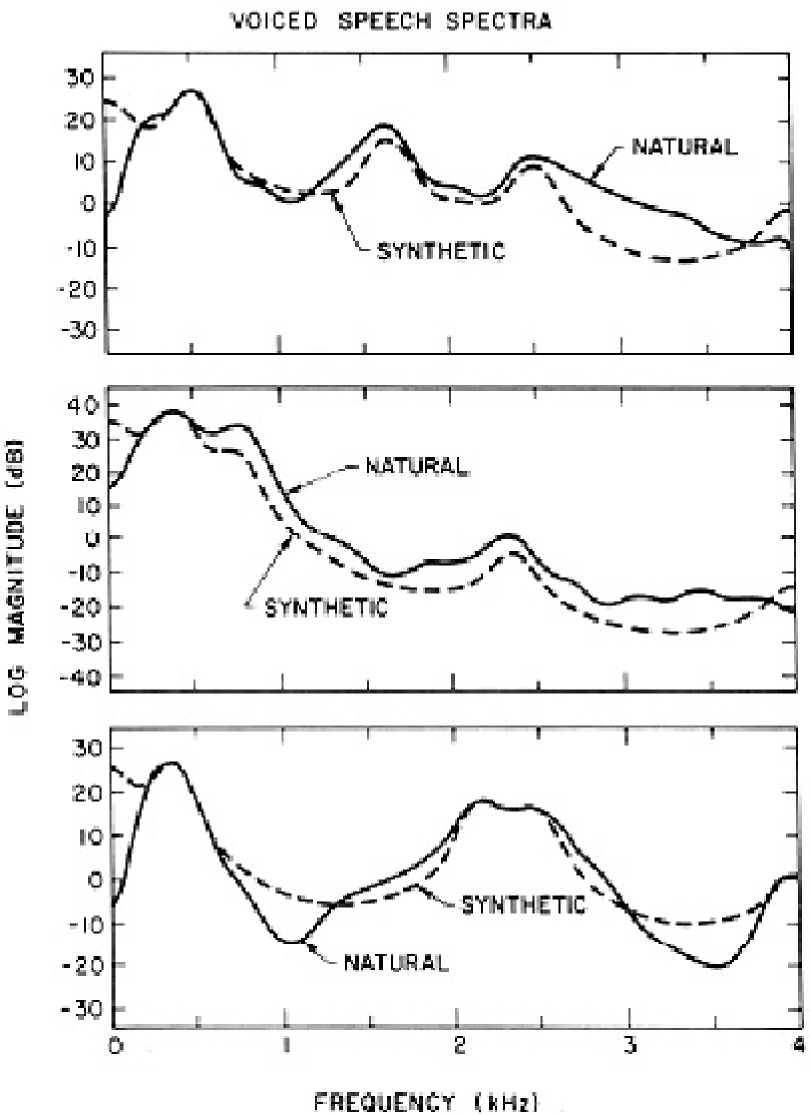


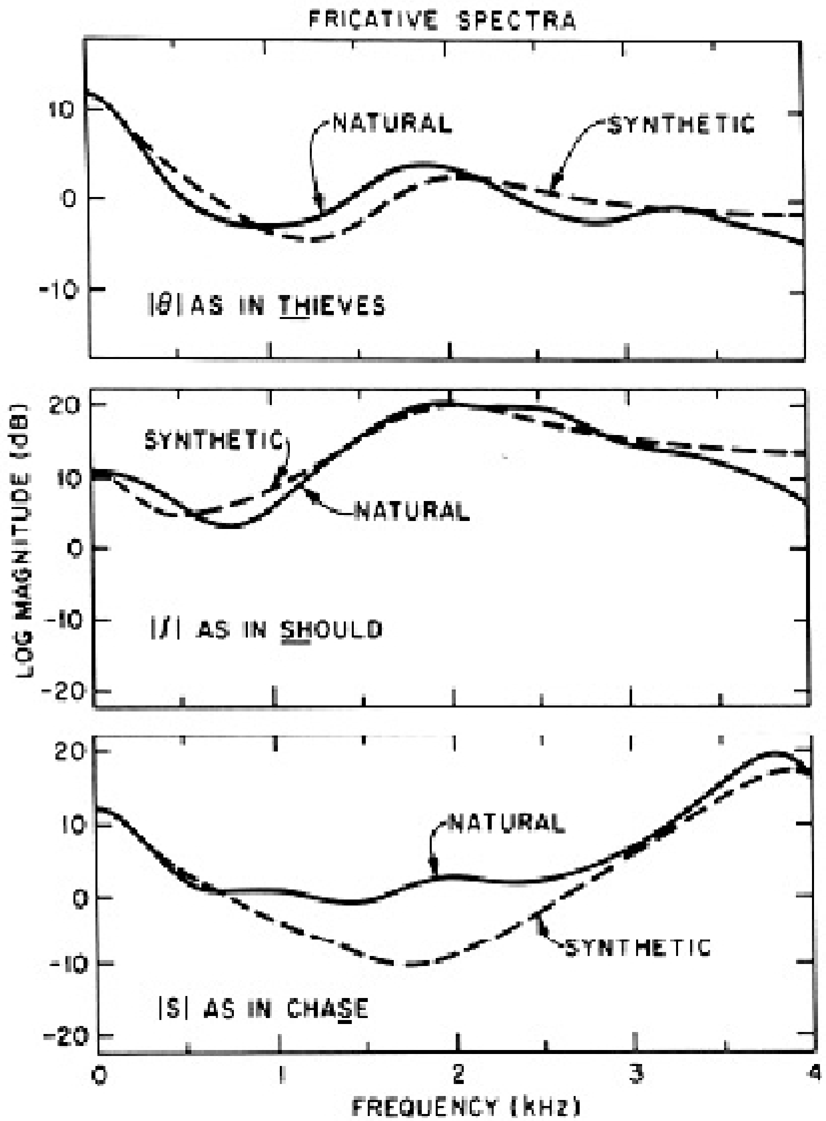
want to estimate time-varying model parameters every 10-20 msec

Cepstral Formant Estimation

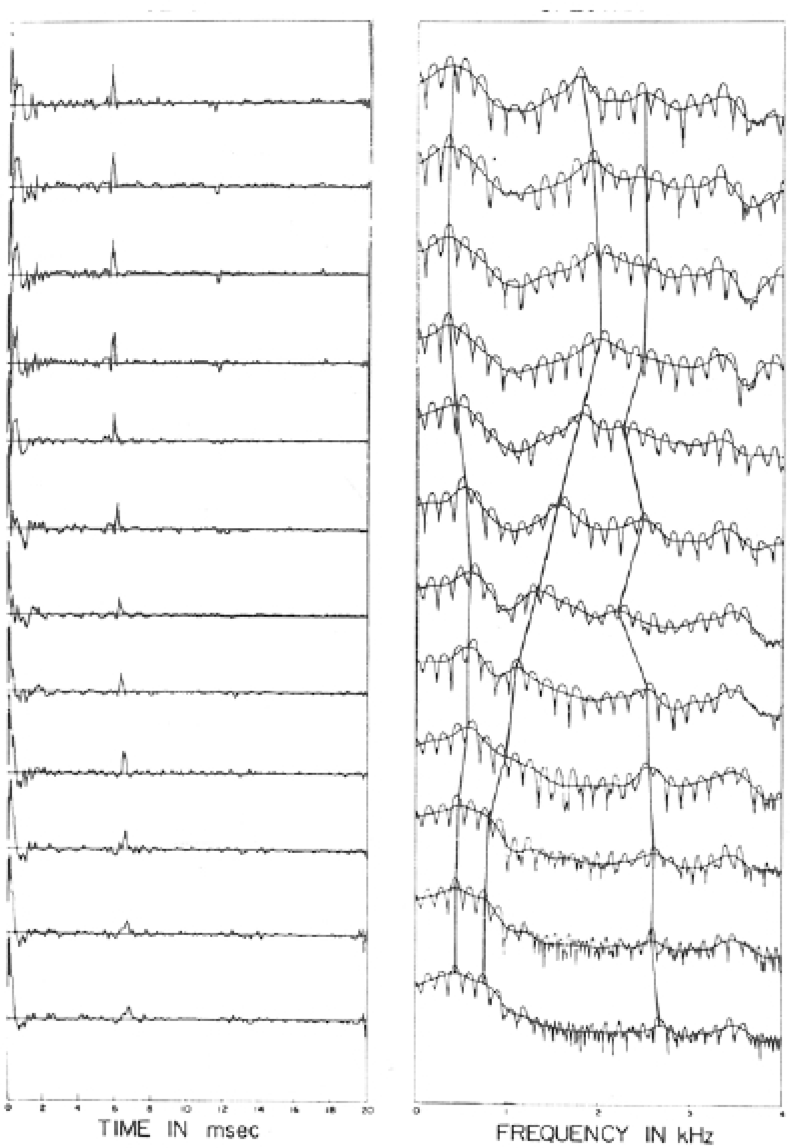
1. fit peaks in cepstrum—decide if section of speech voiced or unvoiced
2. if voiced-estimate pitch period, lowpass lifter cepstrum, match first 3 formant frequencies to smooth log magnitude spectrum
3. if unvoiced, set pole frequency Fp to highest peak in smoothed log spectrum; choose zero Fz to maximize fit to smoothed log spectrum

Cepstral Formant Estimation



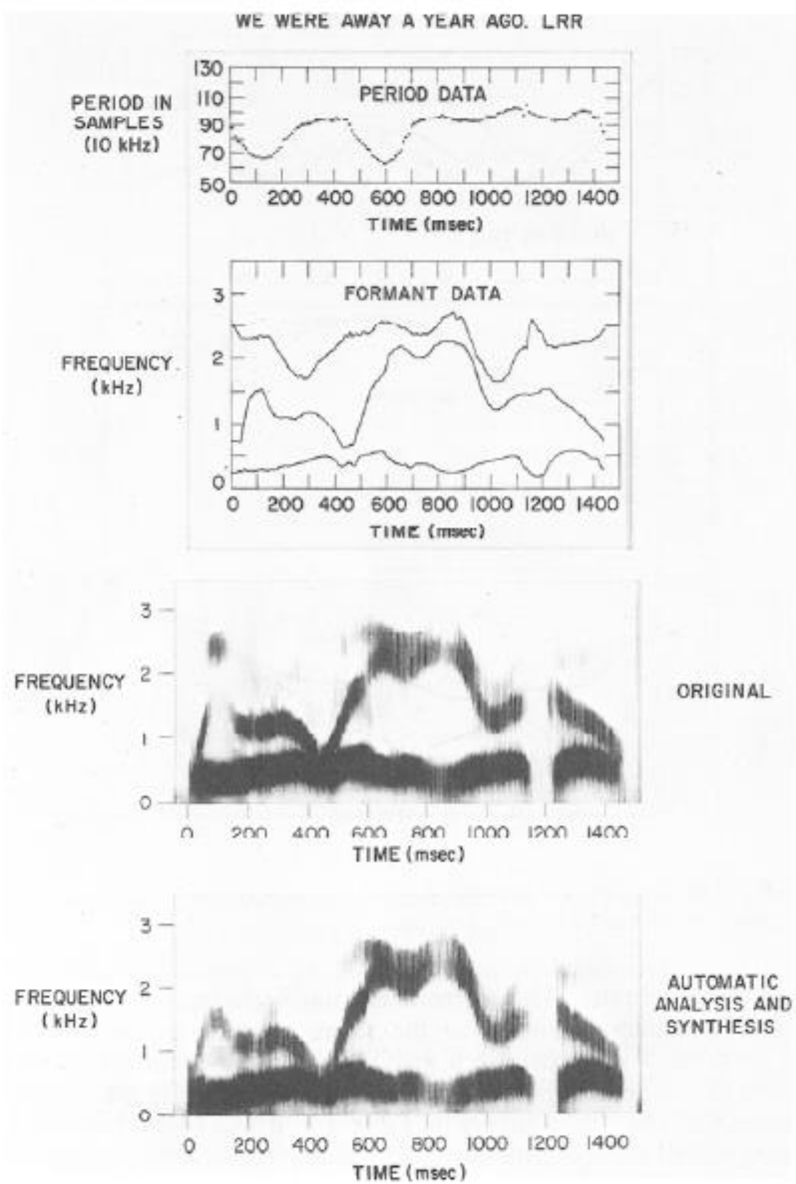


Cepstral Formant Estimation



- sometimes 2 formants get so close that they merge and there are not 2 distinct peaks in the log magnitude spectrum

Cepstral Speech Processing



Cepstral pitch detector – median smoothed

Cepstral formant estimation

Formant synthesizer – 3 estimated formants for voiced speech; estimated formant and zero for unvoiced speech

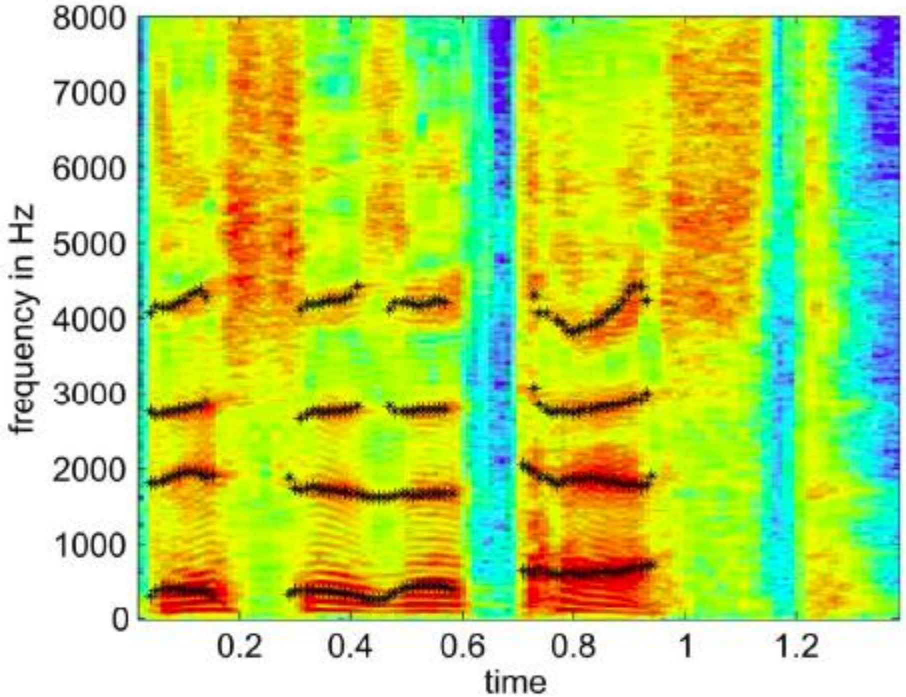
All parameters quantized to appropriate number of levels

LPC-Based Formant Estimation

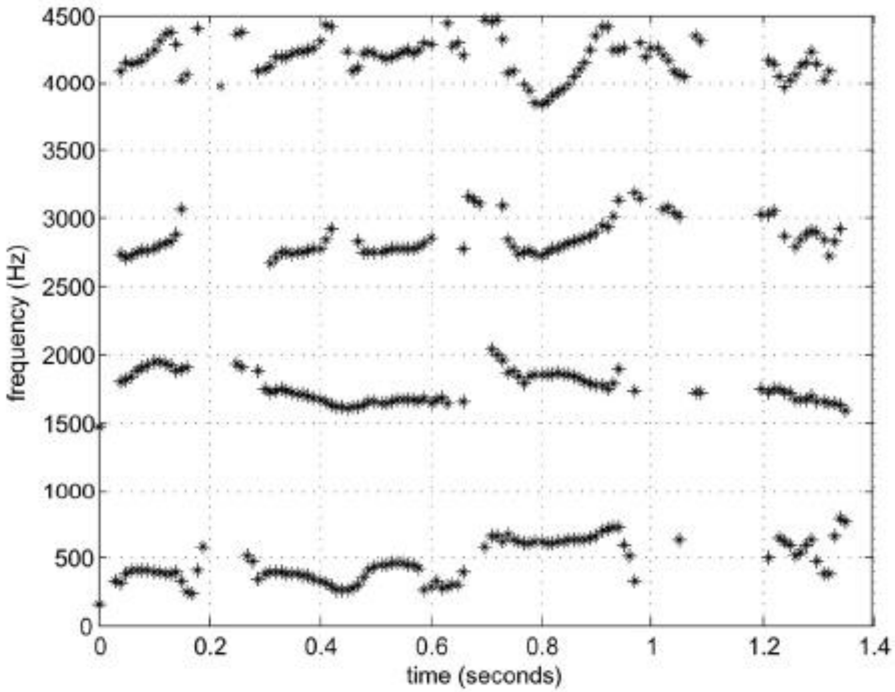
Formant Analysis Using LPC

- factor predictor polynomial—assign roots to formants
- pick prominent peaks in LPC spectrum
- problems on nasals which should be described by poles and zeros

" This is a test. " Formants estimated from p = 16



Polynomial Roots



Algorithms for Speech Processing

- Based on the various representations of speech we can create algorithms for measuring features that characterize speech and estimating properties of the speech signal, e.g.,
  - presence or absence of speech (Speech/Non-Speech Discrimination)
  - classification of signal frame as Voiced/Unvoiced/ Background signal
  - estimation of F0 for a voiced speech frame
  - estimation of the formant frequencies (resonances and anti-resonances of the vocal tract) for both voiced and unvoiced speech frames