Speech and Audio Coding Theory

Contents of lecture

- Pitch estimation
  - Time domain methods
  - Frequency domain methods
  - Time- and frequency-domain methods
  - Pre- and post-processing techniques
- Voiced-unvoiced classification
  - Hard-decision voicing
  - Soft-decision voicing

Introduction

- Three main speech features
  - Spectral envelope: from short-term correlation → LSFs
  - Pitch (period and gains): from long-term correlation
    - Especially for pitch period
      - Used in pitch predictor, to reduce the search space for LTP parameters (gains)
      - Used in the generation of excitation signal for a voiced region
  - Voiced-unvoiced (V-UV) classification
    - Voiced: high energy, periodicity
      - If incorrectly classified as unvoiced, the synthesized speech will sound rough and less intelligible.
    - Unvoiced: like random noise with no periodicity
      - If incorrectly classified as voiced, the synthesized speech will sound annoyingly metallic or robotic.
    - Transition region between voiced and unvoiced, or inherently mixed (i.e., /*d*/)
      - A soft decision voicing: frequency-band-dependent V-UV classification
      - The soft decision is usually carried out in the frequency domain.

Pitch estimation

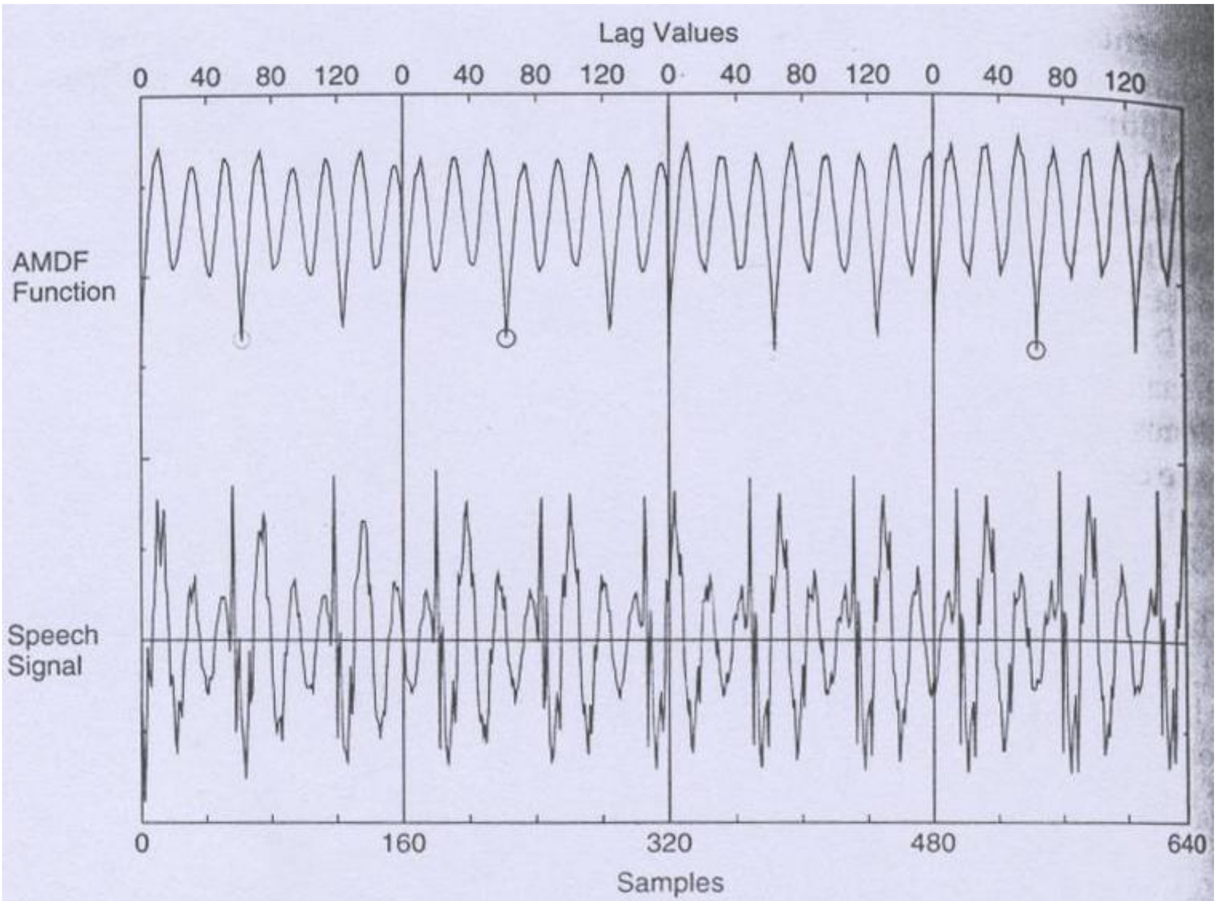
- Why accurate and reliable pitch period estimation is difficult?
  - No perfect train of periodic pulses, even in voiced regions
    - slowly evolves from cycle to cycle
  - Onset and offset regions of voiced speech are not stationary.
  - In some parts, the speech may contain a mixture of voiced and unvoiced signals.
  - Interaction with 1st formant as in the child or female speech
  - Background ambient noise
- Pitch determination algorithms (PDA) based on
  - Time domain properties
  - Frequency domain properties
  - Both the time and frequency domain properties

Time domain methods for PD

- Idea: using similarity of the waveform in time domain
- AMDF (Average Magnitude Difference Function) PDA
  - Definition: $$A(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} \left| s(n) - s(n+\tau) \right|$$
    - Anti-correlation measure (dissimilarity measure)
  - Merits
    - Simple computation
    - Not useful with DSPs optimized for multiplications and additions, but still useful with ASICs having no arithmetic component.
    - Smaller dynamic range due to no multiplications
      - Bounded to zero
    - Narrower valleys for stationary signals

Time domain methods for PD

- AMDF (Average Magnitude Difference Function) PDA



Time domain methods for PD

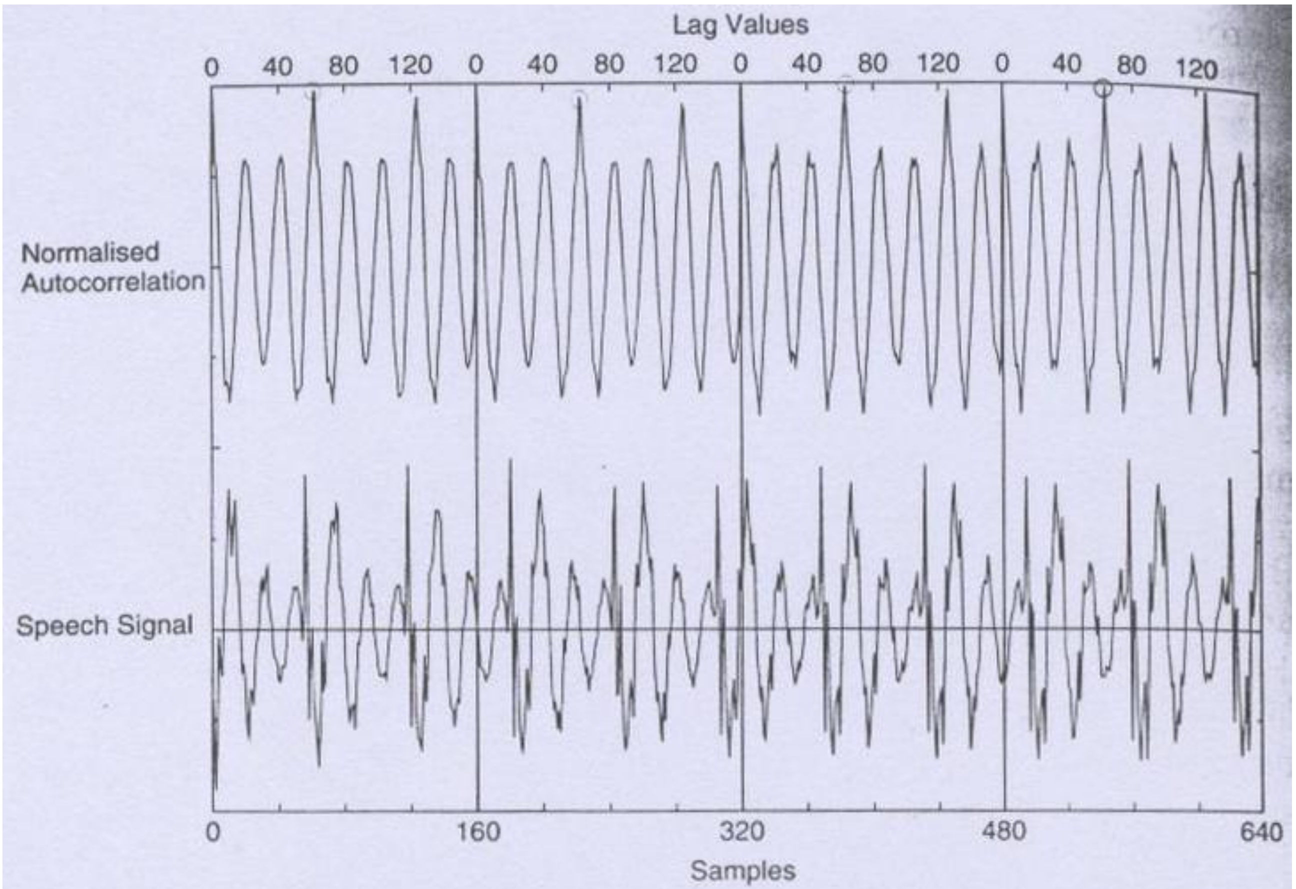
- Auto-correlation PDA
  - Definition
    - $E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n) - s(n+\tau)]^2$
  - Normalized criterion reflecting the non-stationary effect of pitch
    - $E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n) - \beta s(n+\tau)]^2$
    - β: scaling factor (pitch gain)

Time domain methods for PD

- Auto-correlation PDA
  - If the signal is stationary (that is, $\Sigma s^2(n) = \Sigma s^2(n+\tau)$), the similarity function becomes $E(\tau) = R(0) - R(\tau)$
    - Here, $R(\tau) = \sum_{n=0}^{N-1} s(n)s(n+\tau)$
  - Therefore, minimizing $E(\tau)$ corresponds approximately to maximizing $R(\tau)$.
  - Merits
    - Easy to implement in real-time with DSPs due to its regular form of multiplications
    - Phase insensitive

Time domain methods for PD

- Auto-correlation PDA (Normalized version)

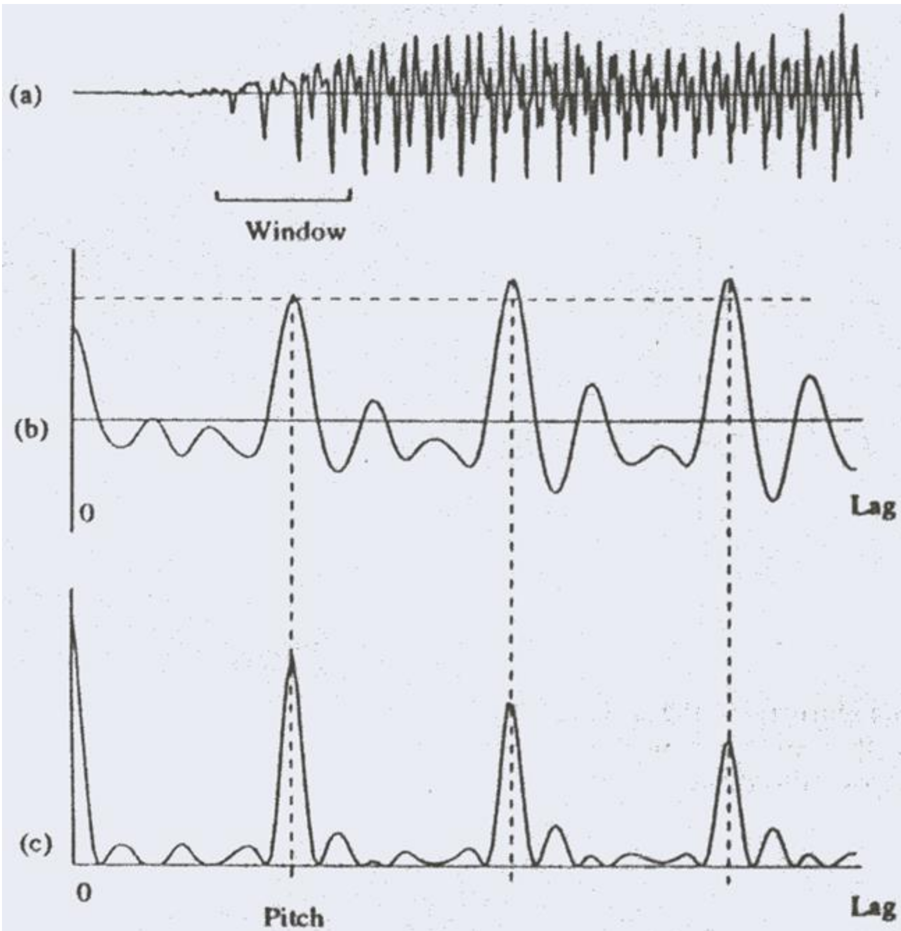


Time domain methods for PD

- Auto-correlation PDA
  - Generalized similarity measure
    - $E(\tau) = \frac{1}{N}\left\{\sum_{n=0}^{N-1}\left|s(n) - s(n+\tau)\right|^k\right\}^{\frac{1}{k}}$
    - From experiments, k=2 is best.
    - Since it corresponds to the auto-correlation method, it means that auto-correlation method is superior to AMDF method.

Time domain methods for PD

- Drawback of the direct auto-correlation method



(a) Original speech
- Window taken on an onset region

(b) Direct autocorrelation function
- Difficult to set an appropriate TH

(c) Normalized autocorrelation function
- Always a consistent pattern
  - Decreasing peaks
  - Bounded to zero
- Relatively easy to set the TH

Time domain methods for PD

- Normalized auto-correlation method
  - From $E(\tau)=\frac{1}{N}\sum_{n=0}^{N-1}[s(n)-\beta s(n+\tau)]^2$ and $\partial E(\tau)/\partial\beta = 0$,

    we get $$\beta=\frac{\sum_{n=0}^{N-1}s(n)s(n+\tau)}{\sum_{n=0}^{N-1}s^2(n+\tau)}$$
  - Substituting this to the normalized criterion, we obtain

    $$E(\tau)=\sum_{n=0}^{N-1}s^2(n)-\frac{\left[\sum_{n=0}^{N-1}s(n)s(n+\tau)\right]^2}{\sum_{n=0}^{N-1}s^2(n+\tau)}$$
  - Removing the negative correlation effects, the criterion becomes to maximize

    $$R(\tau)=\frac{\sum_{n=0}^{N-1}s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}s^2(n+\tau)}}$$

Frequency domain methods for PD

- Basic idea
  - Using the harmonic structure in frequency domain
  - Main drawback: high computational complexity

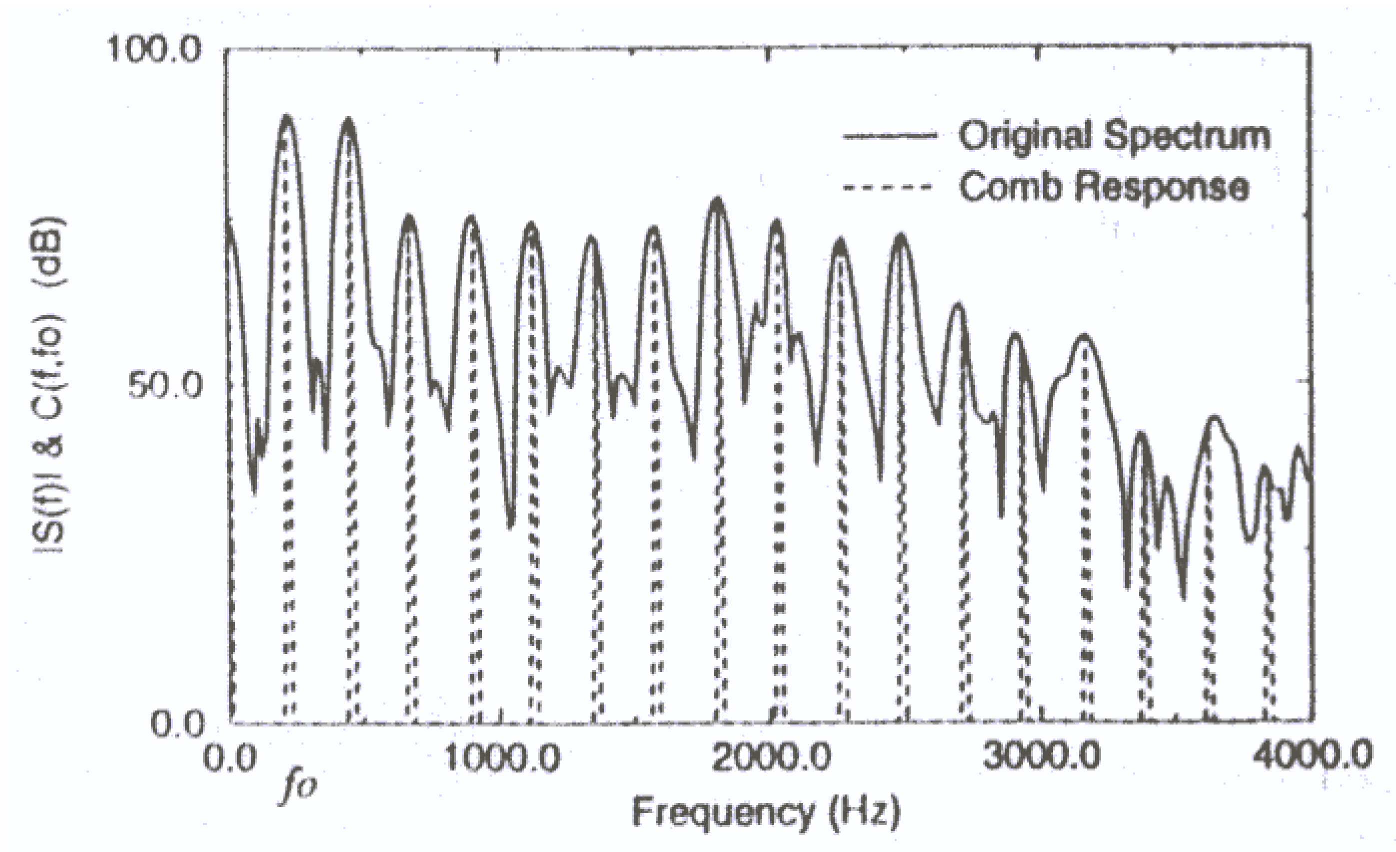
- Harmonic peak detection method
  - Using comb filter in the frequency domain as in the following.
    - That is, to maximize the following autocorrelation output.

$$A_c(\omega_0) = \frac{\omega_0}{\Omega_m} \sum_{k=1}^{\Omega_m/\omega_0} S(k\omega_0) W(k\omega_0) \qquad \frac{2\pi}{\tau_{\max}} \le \omega_0 \le \frac{2\pi}{\tau_{\min}}$$

    - Here, ω_0: fundamental freq., Ω_m: $(2\pi f_s)/2$, $W(k\omega_0)$: comb peaks
  - Actually, the first harmonic component is likely to disappear due to front-end filtering, therefore it is desirable to determine the period by utilizing the entire harmonics.

Frequency domain methods for PD

- Harmonic peak detection method using comb filter



Frequency domain methods for PD

- Spectrum similarity method
  - Comparing the reconstructed spectrum with the original speech spectrum
    - That is, to minimize the following error function.

$$E(\omega_0) = \sum_{m=0}^{M-1} \left(S(m) - \hat{S}(m,\omega_0) \right)^2$$

$$A_l(\omega_0) = \frac{\sum_{m=a_l}^{b_l} S(m) W\left(\frac{2\pi}{M} m - l\omega_0\right)}{\sum_{m=a_l}^{b_l} \left| W\left(\frac{2\pi}{M} m - l\omega_0\right) \right|^2}$$

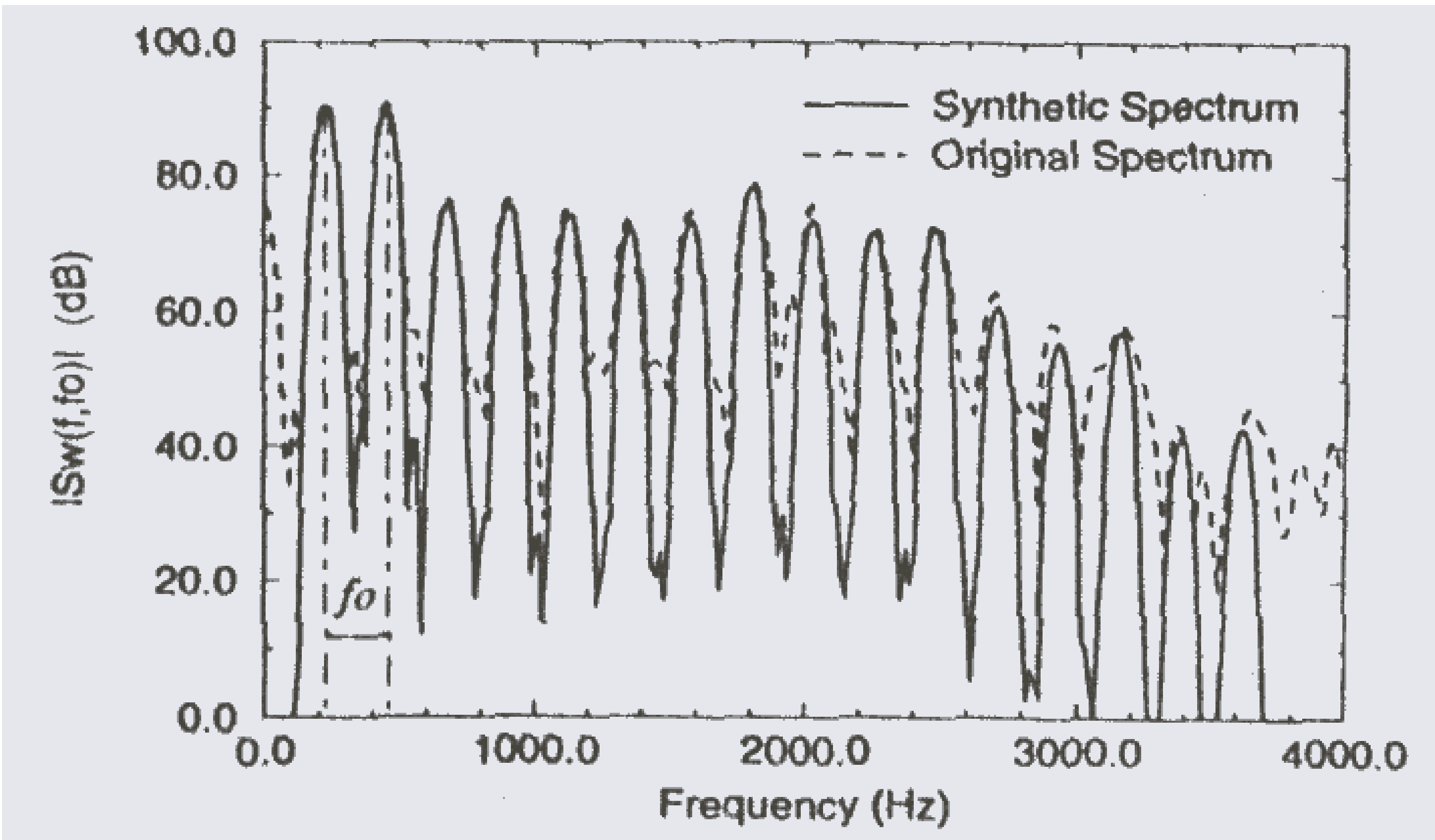
$$\hat{S}(m,\omega_0) = \begin{cases} A_0(\omega_0) W\left(\frac{2\pi}{M} m\right) \\ A_1(\omega_0) W\left(\frac{2\pi}{M} m - \omega_0\right) \\ \vdots \\ A_l(\omega_0) W\left(\frac{2\pi}{M} m - l\omega_0\right) \\ \vdots \end{cases}$$

$$a_l = \left\lceil \frac{M}{2\pi} \left(l - \frac{1}{2} \right) \omega_0 \right\rceil$$

$$b_l = \left\lfloor \frac{M}{2\pi} \left(l + \frac{1}{2} \right) \omega_0 \right\rfloor = a_{l+1} - 1$$

Frequency domain methods for PD

- Spectrum similarity method



Time- and frequency-domain methods for PD

- Pitch estimation using spectral autocorrelation
  - Redefine the normalized autocorrelation function as follows.

$$R_T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{N-\tau-1} s^2(n) \sum_{n=0}^{N-\tau-1} s^2(n+\tau)}}$$

Called normalized temporal autocorrelation (TA)

  - Similarly, define the normalized spectral autocorrelation (SA) function in the frequency domain.

$$S(m) = A(m)e^{j\theta(m)} \quad \text{for } 0 \le m \le M-1$$

$$A_z(m): \text{zero-crossing spectrum}$$

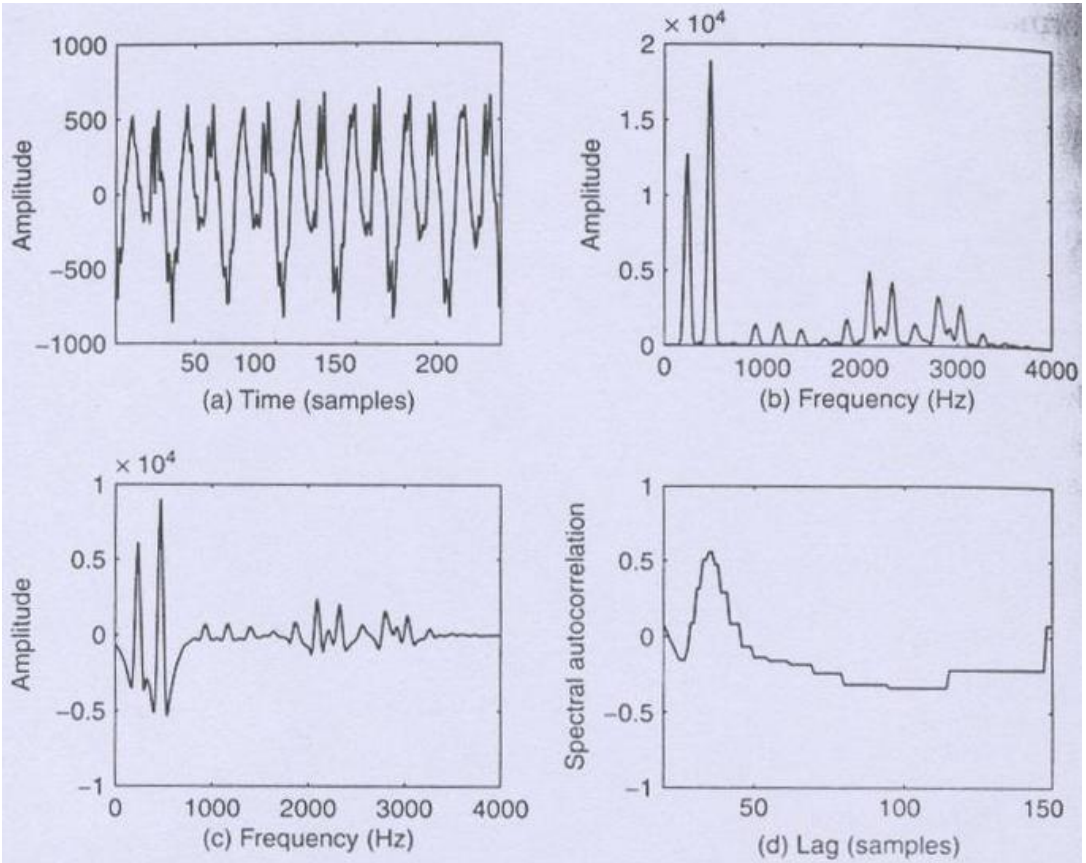
$$\omega_\tau = \lfloor M/\tau + 0.5 \rfloor$$

$$T_0^{(l)} \text{ and } T_0^{(u)}: \text{lower and upper limits}$$

$$R_S(\tau) = \frac{\sum_{m=0}^{\lfloor M/2 \rfloor - \omega_\tau} A_z(m) A_z(m+\omega_\tau)}{\sqrt{\sum_{m=0}^{\lfloor M/2 \rfloor - \omega_\tau} A_z^2(m) \sum_{m=0}^{\lfloor M/2 \rfloor - \omega_\tau} A_z^2(m+\omega_\tau)}} \quad \text{for } T_0^{(l)} \le \tau \le T_0^{(u)}$$

Time- and frequency-domain methods for PD

- Example of pitch estimation using spectral autocorrelation
  - T_0 = 34-sample (as in female speech)



(a) Speech signal (8 kHz)

(b) Magnitude spectrum

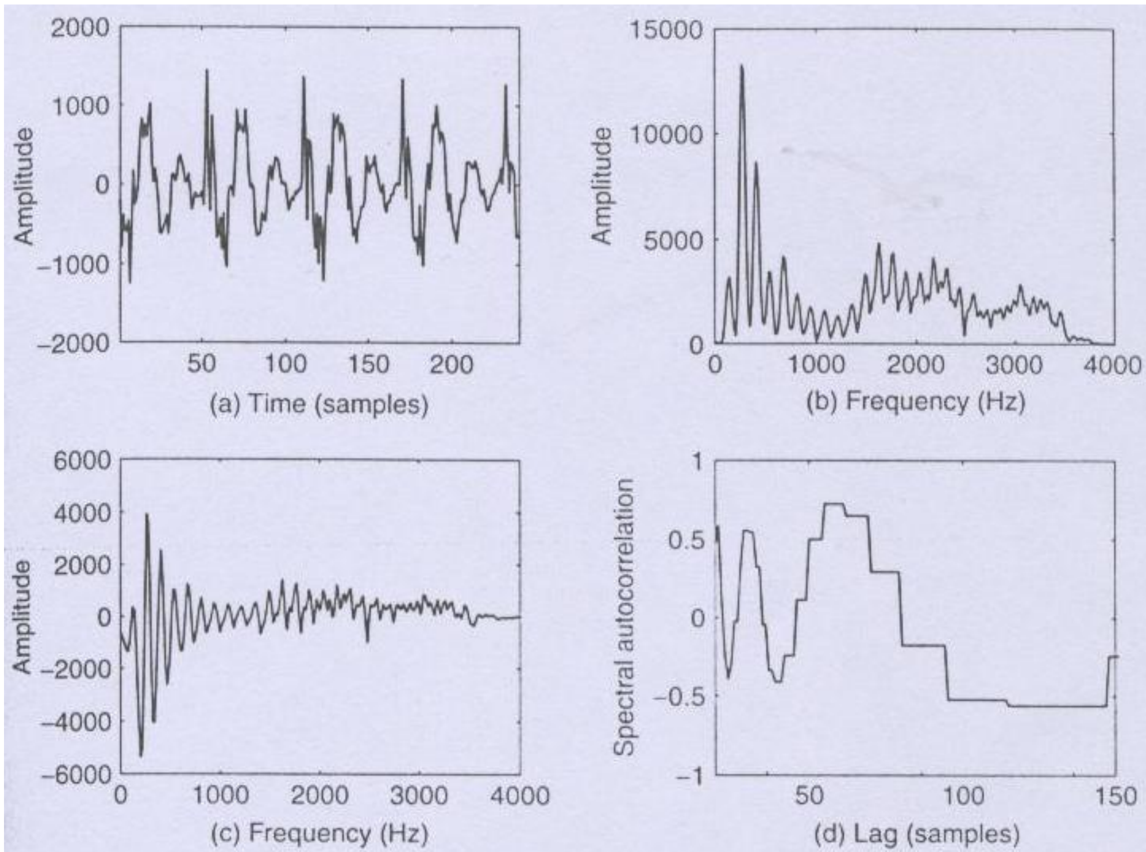
(c) Zero-crossing spectrum

(d) SA function

→ Good result

Time- and frequency-domain methods for PD

- Example of pitch estimation using spectral autocorrelation
  - T_0 = 59-sample (as in male speech)



(a) Speech signal (8 kHz)

(b) Magnitude spectrum

(c) Zero-crossing spectrum

(d) SA function

→ Not good result

Time- and frequency-domain methods for PD

- Analyzing the characteristics of the TA-based and SA-based PDAs, respectively,
  - TA-based PDA: likely to detect an unwanted pitch period multiple
  - SA-based PDA: likely to be pitch-halving
- Compensating for the problems by combining two methods,

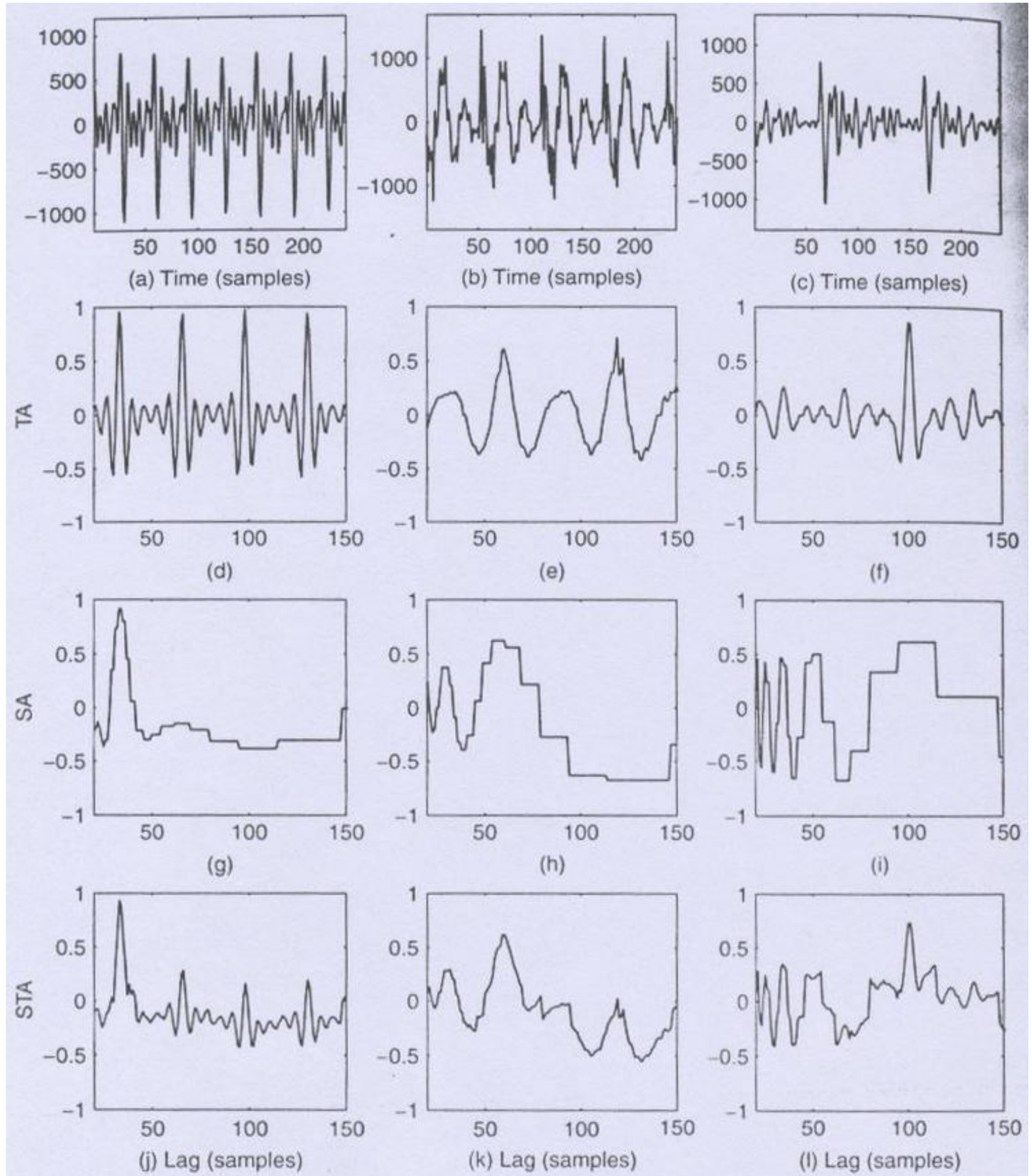
$$R_{ST}(\tau) = \alpha R_T(\tau) + (1-\alpha) R_S(\tau) \quad 0 \le \alpha \le 1$$

$$\hat{T}_0 = \arg\max_{\tau} \{R_{ST}(\tau)\}$$

  - Called the spectro-temporal autocorrelation (STA) PDA

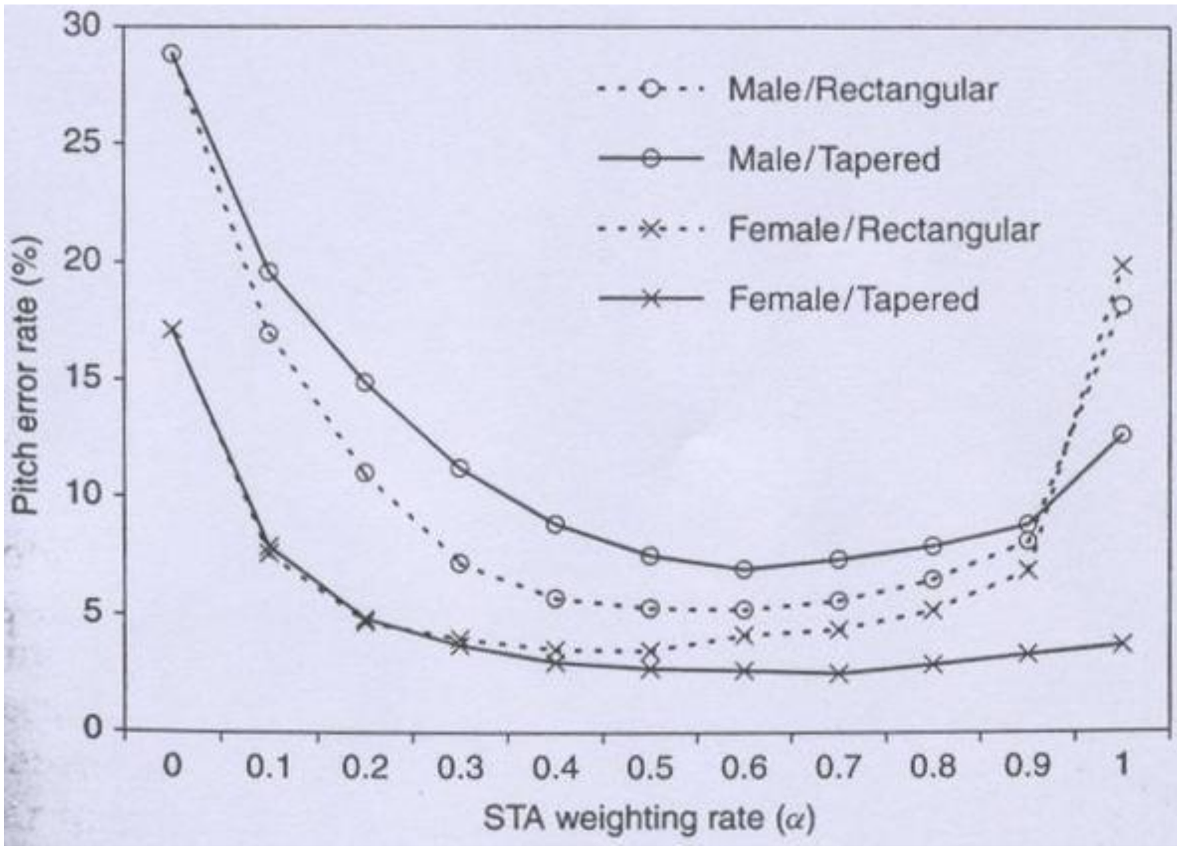
Time- and frequency-domain methods for PD

- Comparison of TA, SA, and STA (α=0.5)
  - Left: 32-sample T_0
  - Middle: 59-sample T_0
  - Right: 100-sample T_0



Time- and frequency-domain methods for PD

- Analysis of the effect of the STA weighting factor α in terms of the pitch error rate



Pre- and post-processing techniques

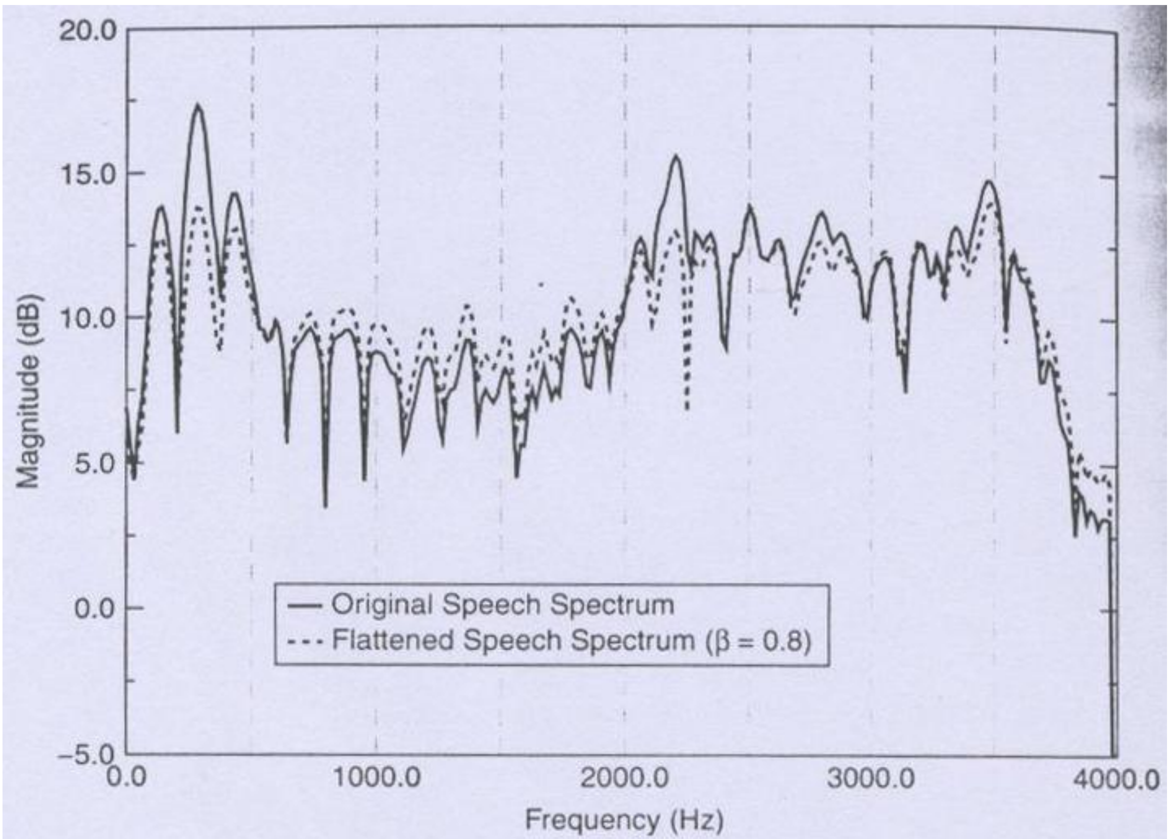
- Objectives
  - To improve the pitch period estimation performance
- Spectrum flattening
  - Removing the formants before pitch estimation process
  - Linear method: using LPC inverse filter
    - Drawback: The fundamental frequency and the first formant of high-pitch speech (like children or female) may be overlapped. → This may destroy the entire periodicity information in the residual signals.
    - Solution: obtaining the intermediate signal between the original and the LPC residual (even though high computations)

$$S_f(z) = \frac{A(z)}{A(z/\gamma)} S(z) \quad \text{for } 0 \le \gamma \le 1$$

      - $S_f(z)$: formant-suppressed signal, $A(z)$: inverse filter, γ : formant weighting factor

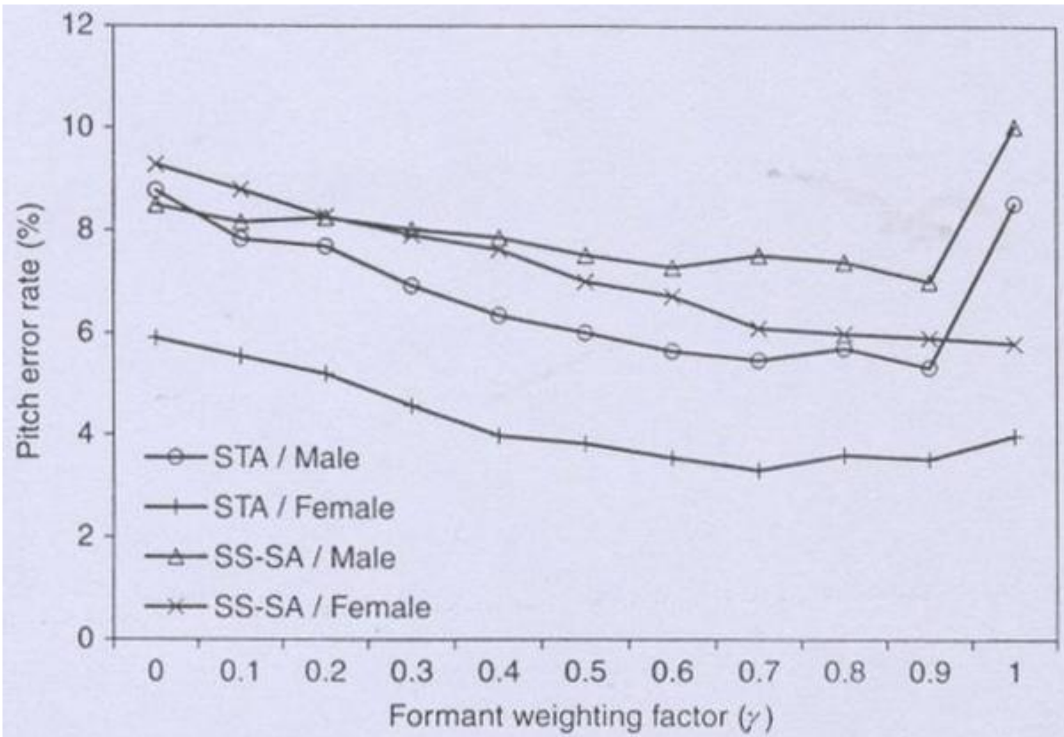
Pre- and post-processing techniques

- Spectrum flattening
  - Influence of the spectrum-flattening filter



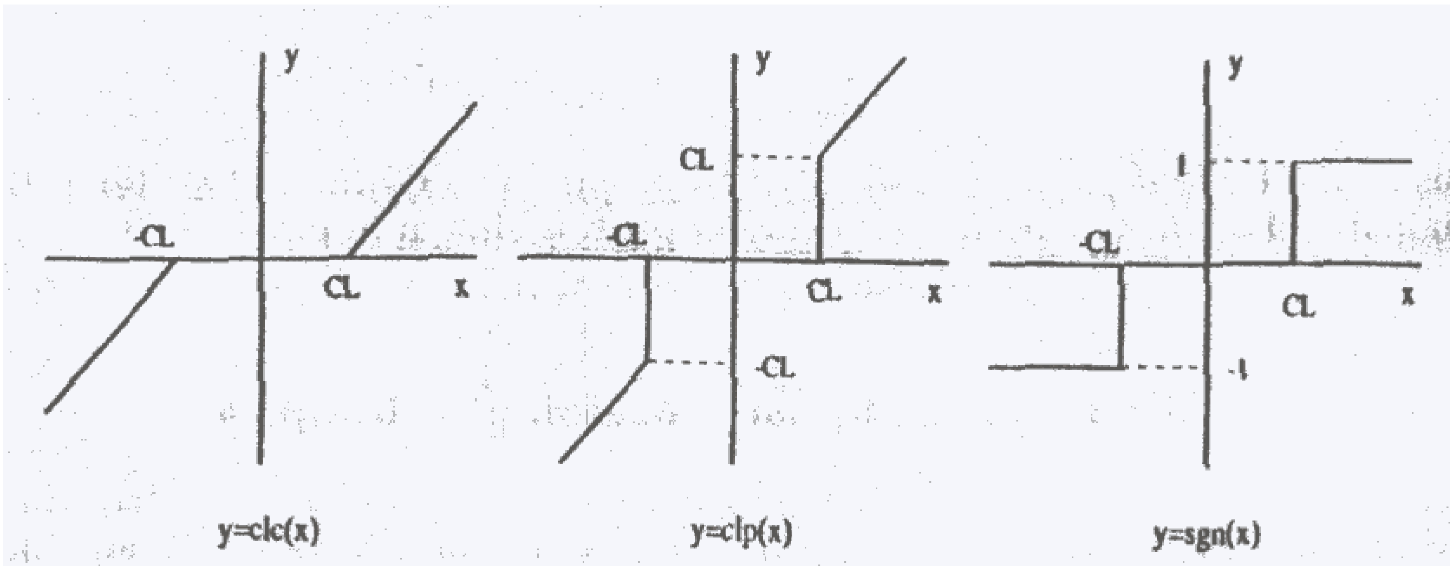
Pre- and post-processing techniques

- Spectrum flattening
  - Analysis of the effect of γ in terms of the pitch error rate
    - Here, SS-SA is a PDA using spectral synthesis – spectral autocorrelation method.



Pre- and post-processing techniques

- Spectrum flattening
  - Non-linear method: using center clipping functions
    - Several clipper functions for spectrum flattening



    - Key problem: How to choose optimum clipping threshold (CL)

Pre- and post-processing techniques

- Pitch tracking
  - Principle: To utilize continuity characteristics of pitch in restricting the search space for pitch detection
    - For voiced speech, the variation of pitch period is small.
  - Passive way: Smoothing the pitch periods after main determination
    - Drawback: Smoothing out an original abrupt change
  - Active way: Applying a path penalty to main pitch determination process
    - Forward tracking & backward tracking
    - For example, once a pitch period of the current frame was estimated, the search for the pitch period of the next frame may be restricted to a range of a constant weighting of the current period.

Pre- and post-processing techniques

- Correction of multiple-pitch errors
  - Pitch determination process in time-domain PDA (e.g. auto-correlation method) probably results in those errors.
    - First, a maximum peak is picked.
    - Then, sub-multiple positions are checked by examining whether the ratio

$$\frac{R(\tau_0 / i)}{R(\tau_0)} > TH$$

    - That is, if any, select a minimum integer i (≥ 2) satisfying the above condition, and then determine τ_0 / i as the final pitch period.
    - There is no optimum solution. → The threshold is determined by tuning.

Pre- and post-processing techniques

- Correction of half-pitch errors
  - Pitch determination process in frequency-domain PDA (e.g. spectral auto-correlation method) probably results in those errors.
  - Even in the time-domain PDA, if the previous ratio test is passed wrongly, pitch halving will take place.
  - Therefore, for the vocoder sensitive to pitch period, another solution not using pitch detector is required.

Voiced-unvoiced classification

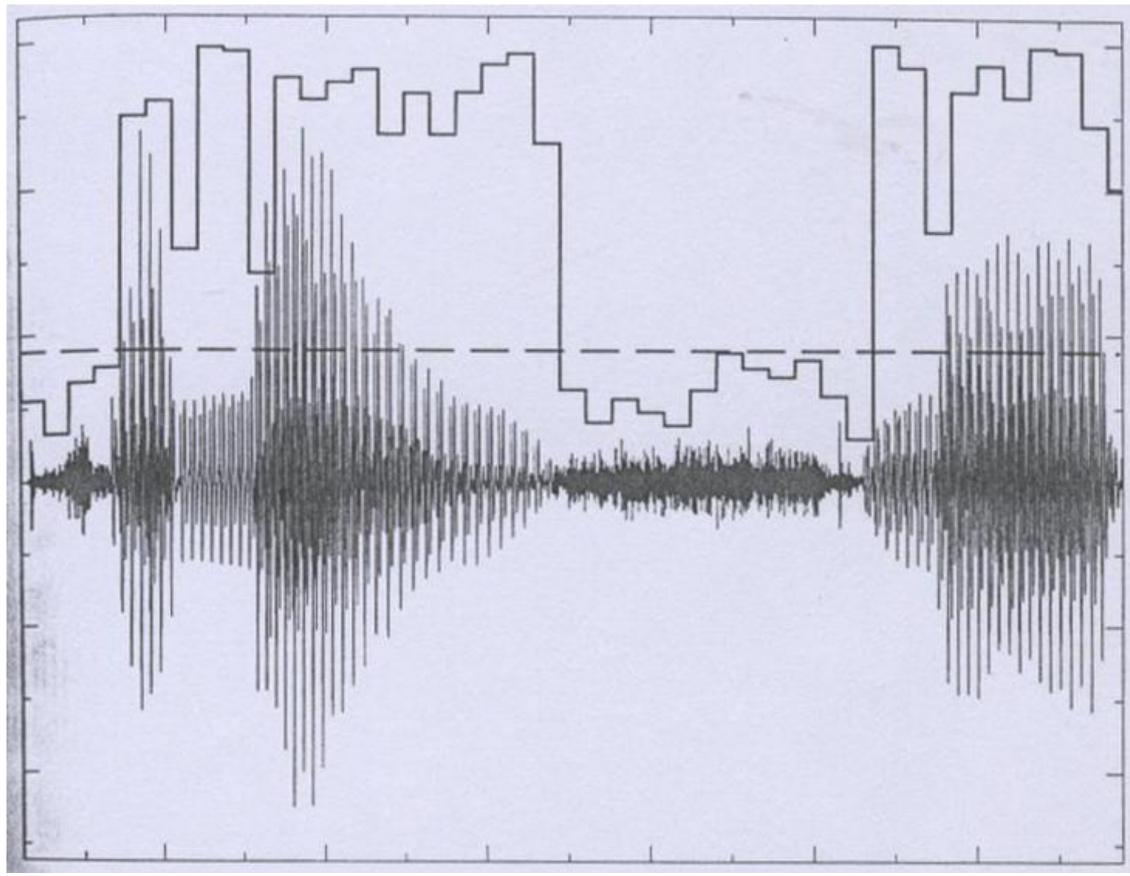
- Classifying the frame as either voiced or unvoiced
- Hard-decision voicing (binary voicing decision)
  - Periodic similarity (high for voiced)
  - Peakiness of speech (high)
  - Zero crossing rate (low)
  - Spectrum tilt (high)
  - Pre-emphasized energy ratio (low)
  - Low-band to full-band energy ratio (high)
  - Frame energy (high)
- Soft-decision voicing (mixed decision of voicing)
  - MBE mixed voicing
  - Split-band mixed voicing

Hard-decision voicing

- Periodic similarity
  - Measuring the regularity of waveform in terms of pitch period

$$Ps = \frac{\left[\sum_{i=1}^{N} s(i)s(i-T)\right]^2}{\sum_{i=1}^{N} s^2(i)\sum_{i=1}^{N} s^2(i-T)}$$

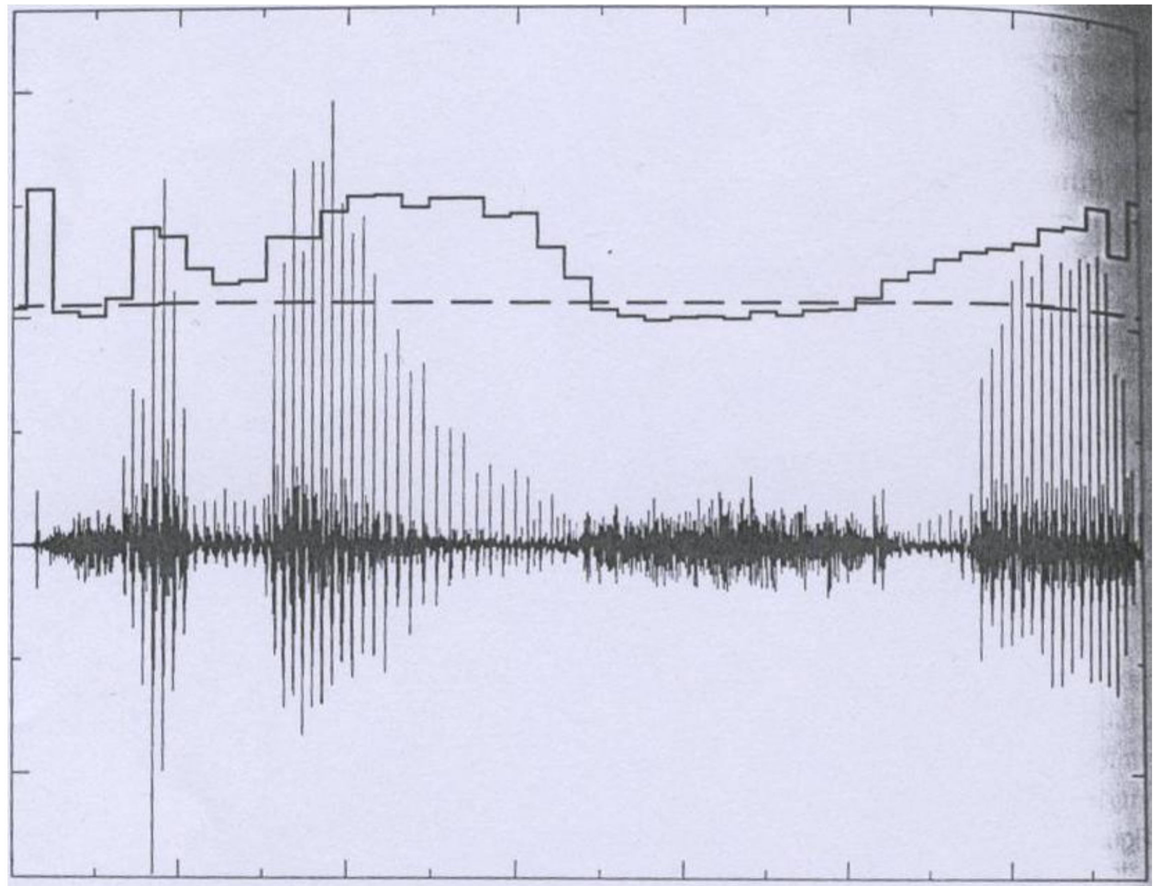
  - T: pitch period
  - A possible TH: 0.5



Hard-decision voicing

- Peakiness of speech
  - Measuring the peakiness of the LPC residual

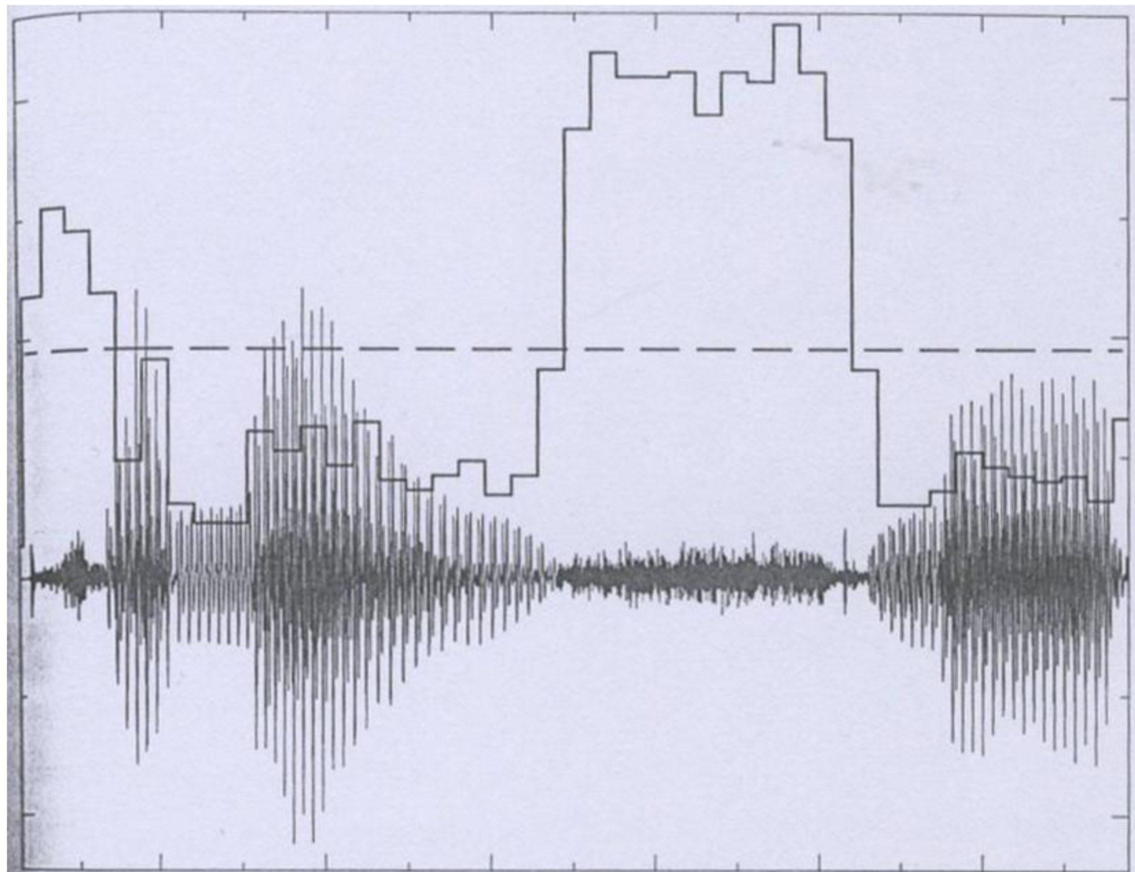
$$Pk = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} r^2(i)}}{\frac{1}{N}\sum_{i=1}^{N} |r(i)|}$$



- A possible TH: 1.4

Hard-decision voicing

- Zero crossing rate
  - Measuring the number of times the signal crosses the zero line
    - A possible TH: 60

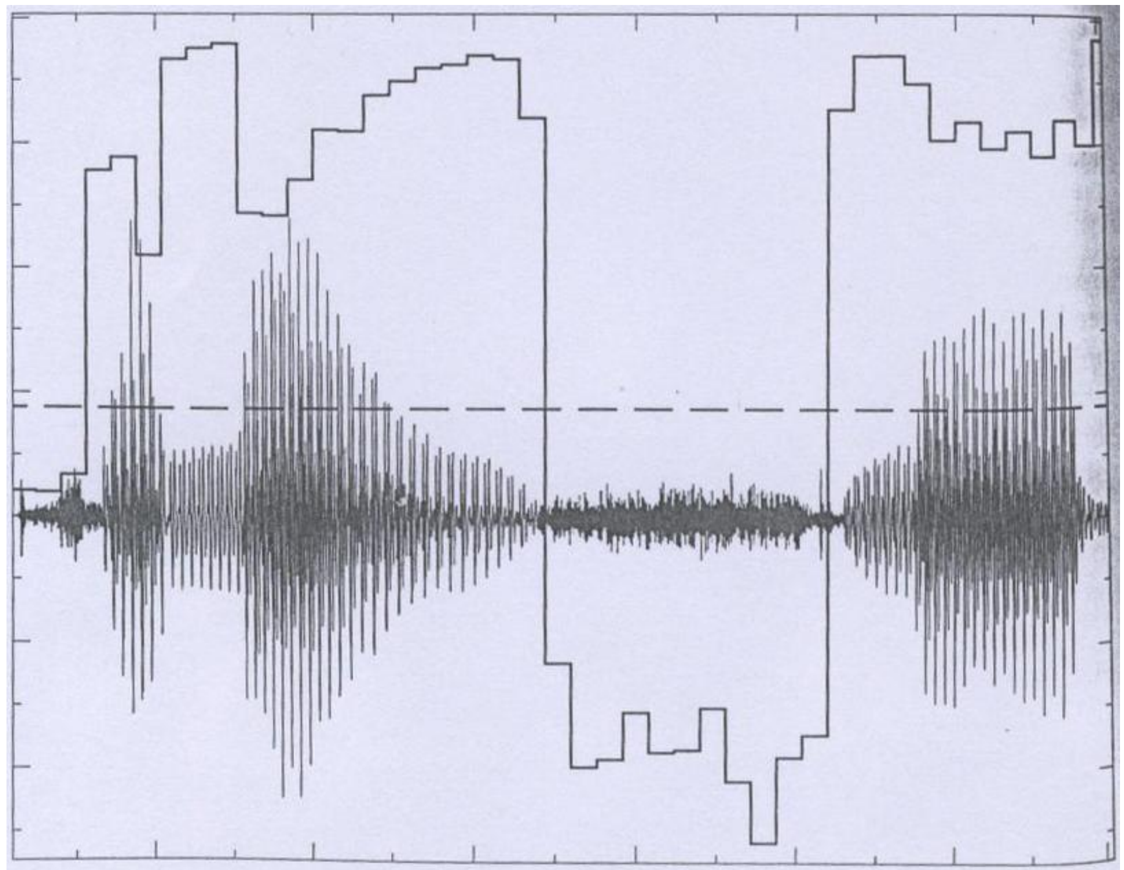


Hard-decision voicing

- Spectrum tilt
  - Voiced speech has higher energy in low frequencies.
  - Measuring the first-order normalized autocorrelation

$$St = \frac{\sum_{i=1}^{N} s(i)s(i-1)}{\sum_{i=1}^{N} s^2(i)}$$

    - A possible TH: 0.25

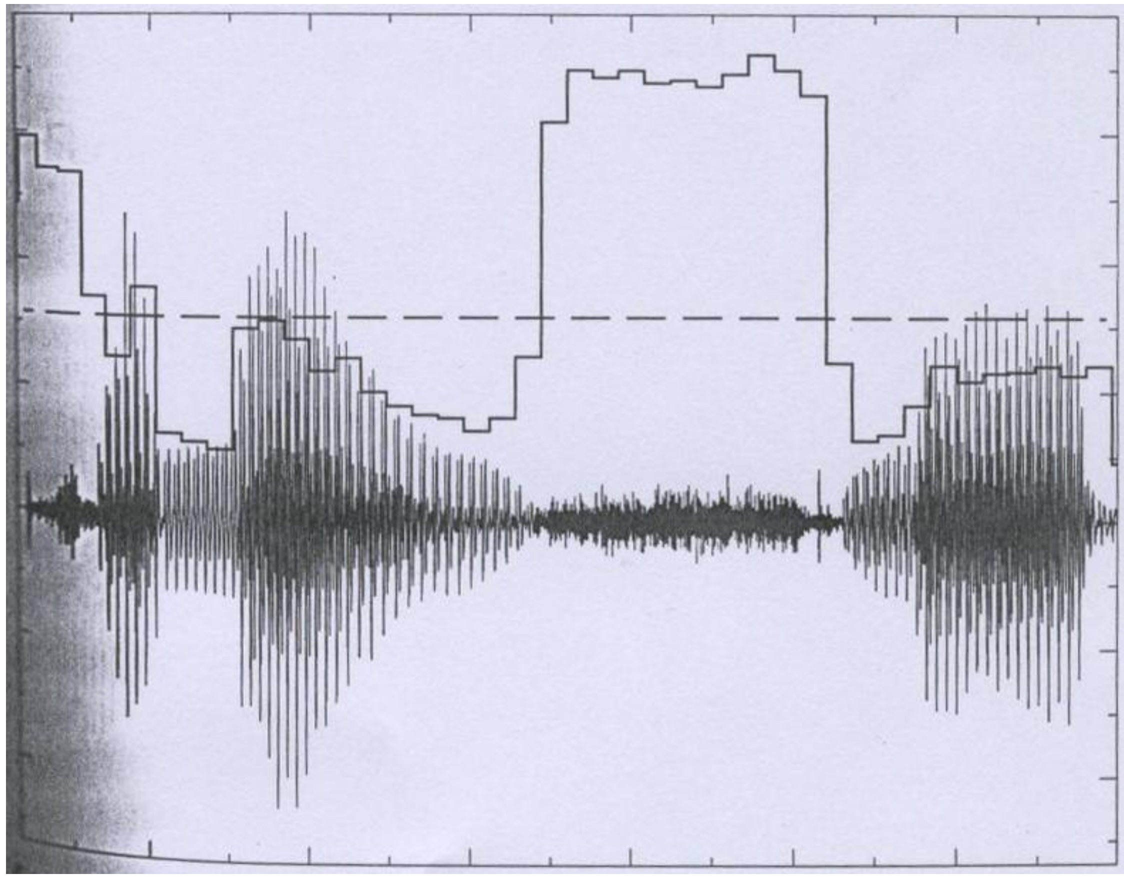


Hard-decision voicing

- Pre-emphasized energy ratio
  - The first-order correlation of voiced samples is much higher than that of unvoiced.
  - Measuring the ratio of the pre-emphasized energy to the original

$$Pr = \frac{\sum_{i=1}^{N} \left| s(i) - s(i-1) \right|}{\sum_{i=1}^{N} \left| s(i) \right|}$$

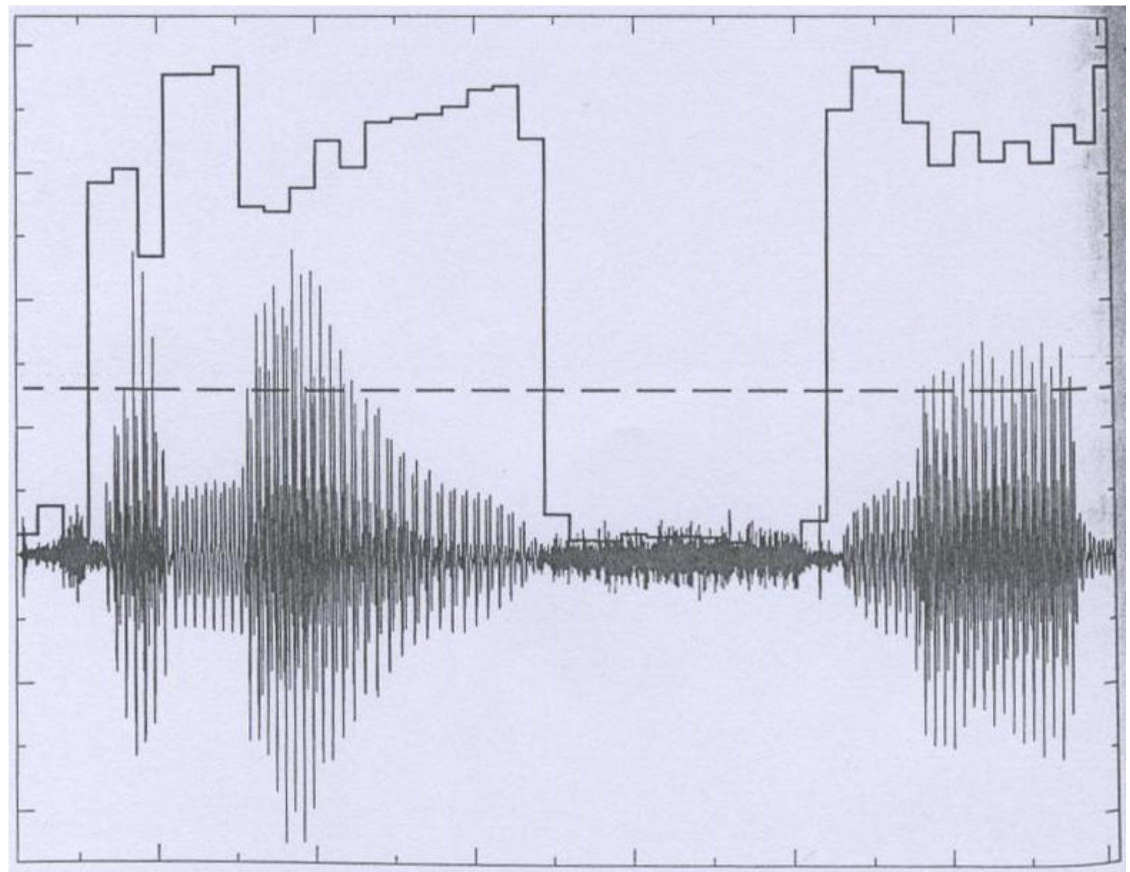
- A possible TH: 0.9



Hard-decision voicing

- Low-band to full-band energy ratio
  - Measuring the energy ratio of the first 1 kHz to the full-band energy

$$LF = \frac{\sum_{i=1}^{N} s_{lpf}^2(i)}{\sum_{i=1}^{N} s^2(i)}$$



- A possible TH: 0.4

Hard-decision voicing

- Frame energy
  - Voiced speech usually has a higher energy not in the absolute value but in a relative amount.
    - That is, a comparison of current frame energy with the tracked maximum and minimum energies, given as follow, would useful.
      - $E_{max}(n)$ can go up fast and come down slowly.

$$E_{\max}(n)=\begin{cases}\alpha E_{\max}(n-1)+(1-\alpha)E_0 & \text{if } E_0>E_{\max}(n-1)\\ \gamma E_{\max}(n-1)+(1-\gamma)E_0 & \text{otherwise}\end{cases}$$

$$\text{where } E_0 : \text{current frame energy, and typically } \alpha=0.5,\ \gamma=0.98$$

      - $E_{min}(n)$ can come down fast and go up slowly.

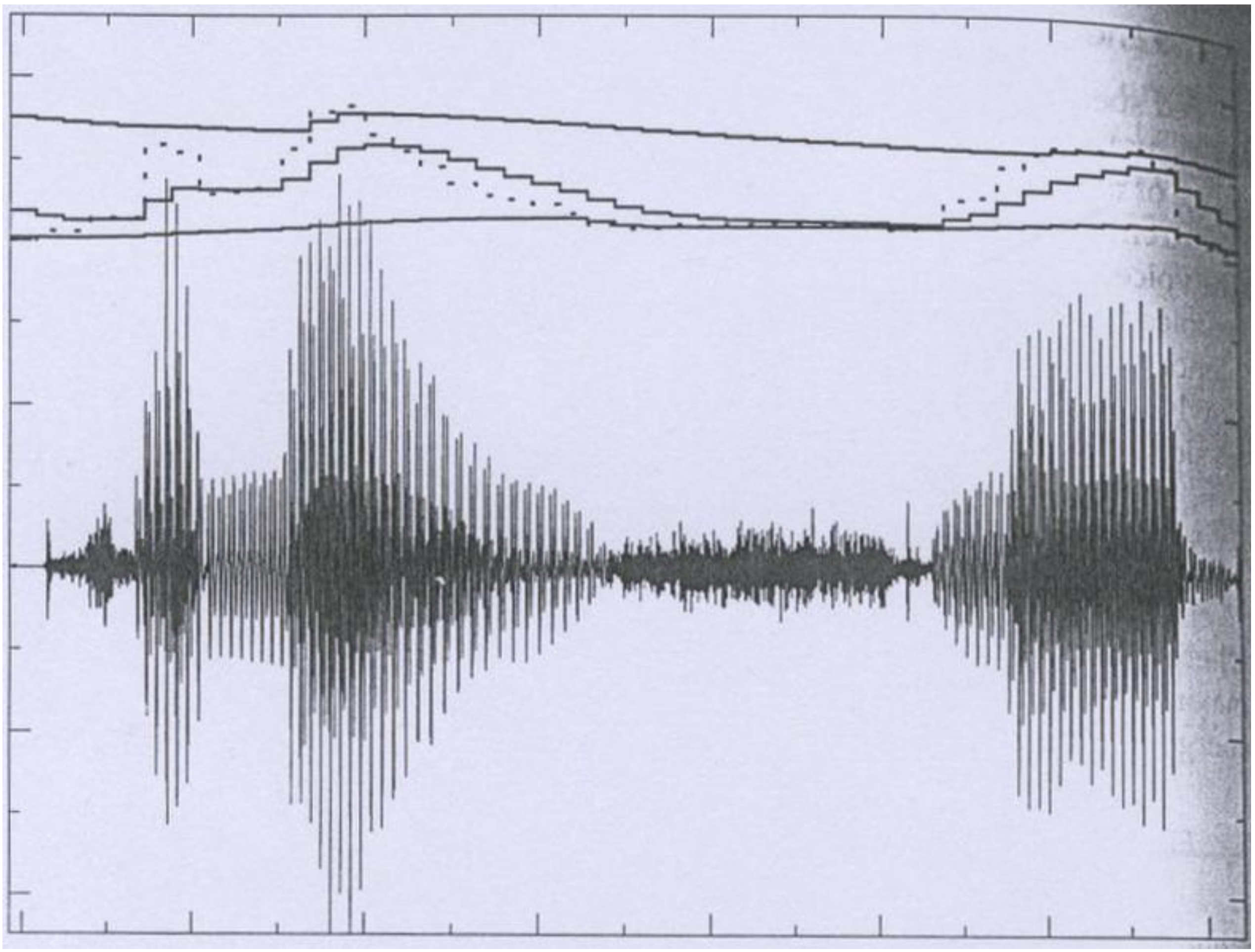
$$E_{\min}(n)=\begin{cases}\zeta E_{\min}(n-1)+(1-\zeta)E_0 & \text{if } E_0<E_{\min}(n-1)\\ \beta E_{\min}(n-1)+(1-\beta)E_0 & \text{otherwise}\end{cases}$$

$$\text{where typically } \zeta=0.55,\ \beta=0.98$$

      - Tracked average energy: $E_{av}(n)=0.75E_{av}(n-1)+0.25E_0$

Hard-decision voicing

- Frame energy



$E_{max}(n)$ track

$E_{av}(n)$ track

$E_{min}(n)$ track

Dotted: frame energy

Speech waveform

Decision logic:

If $\{(E_0 > E_{max} - TH1)$ or $(E_0 > E_{av})\}$ Voiced,

Else if $(E_0 < E_{min} + TH2)$ Unvoiced,

Else Not-sure.

Hard-decision voicing

- Decision-making
  - Combined decision using the voicing indicators
    - The simplest way: majority vote
    - Better rule: weighted combination
  - Two-step normalization
    - To compensate for differences of each parameter from the optimum decision threshold

$$Ps' = \begin{cases} (Ps - Th_{ps}) / (Ps_{\max} - Th_{ps}) & \text{if } Ps > Th_{ps} \\ (Ps - Th_{ps}) / (Th_{ps} - Ps_{\min}) & \text{if } Ps < Th_{ps} \end{cases}$$

$$Zc' = \begin{cases} (Th_{zc} - Zc) / (Th_{zc} - Zc_{\min}) & \text{if } Zc < Th_{zc} \\ (Th_{zc} - Zc) / (Zc_{\max} - Th_{zc}) & \text{if } Zc > Th_{zc} \end{cases}$$

$$\vdots$$

where $Th_{ps}, Th_{pk}, Th_{zc}, \cdots$ are fixed voicing thresholds

Hard-decision voicing

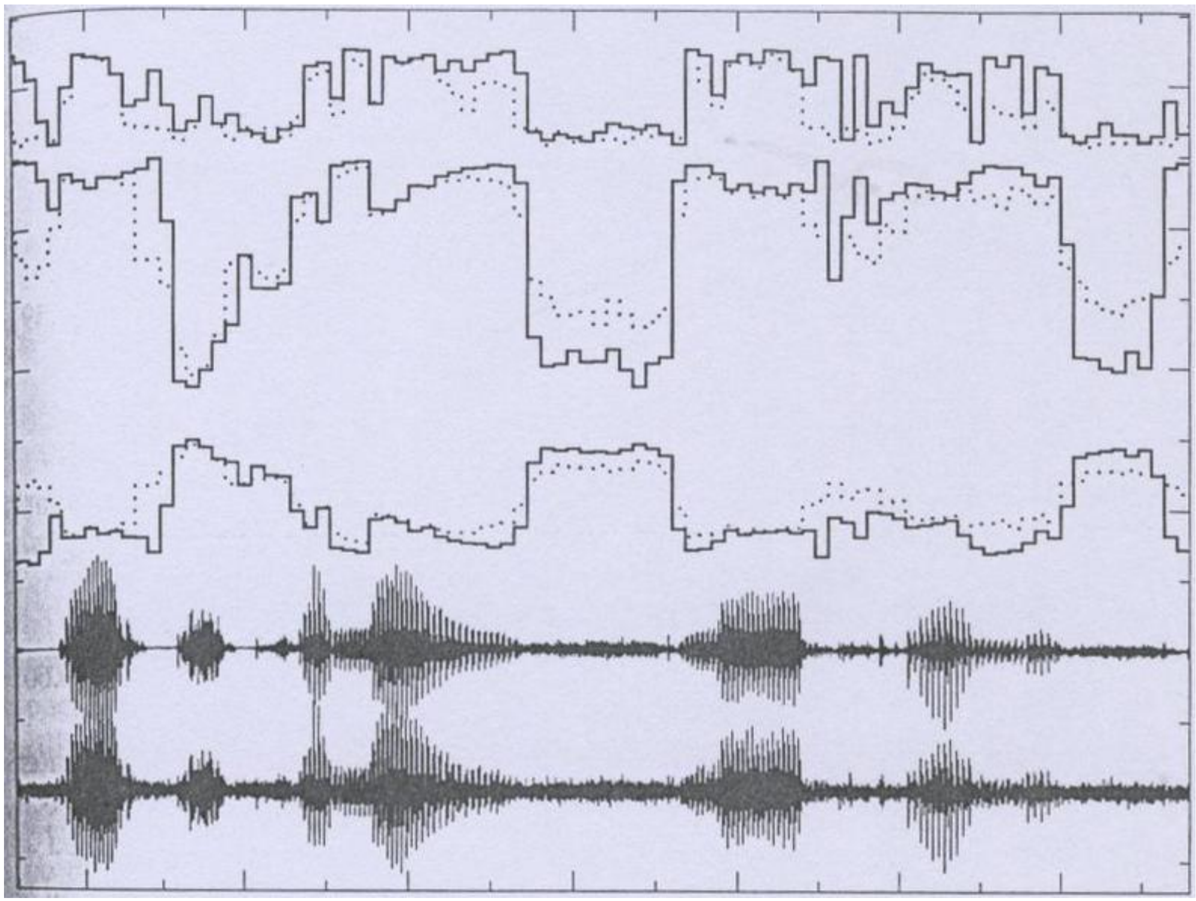
- Decision-making
  - Two-step normalization (cont.)
    - To compensate for different degrees of reliability, the overall voicing indicator V is

      $$V = w_1 Ps' + w_2 Pk' + w_3 Zc' + w_4 St' + w_5 LF' + w_6 Pr' + w_7 Fe'$$

      - The weights are chosen according to the reliability of each indicator.
  - Decision
    - When distinctively positive → voiced
    - When distinctively negative → unvoiced
    - If close to zero → unsure case → further checking
  - Works very well with clean speech without background noises

Hard-decision voicing

- Problems
  - When speech is mixed with background noise, the thresholds may not be valid anymore.
  - When there is a transition from V to UV or vice versa even in clean speech,



Ps plot

St plot

Pr plot

Clean speech waveform

Noisy speech with 10 dB SNR vehicle noise

Dotted: the corresponding plots for noisy speech

Soft-decision voicing

- Alternative approach is to use a soft-decision voicing.
  - A frequency-domain voicing-decision process using the harmonic and random structures of voiced and unvoiced sounds, respectively
- Two methods
  - Multi-band excitation (MBE) mixed voicing
  - Split-band mixed voicing

MBE mixed voicing

- Voicing decision
  - Define the normalized distance D_k between the original and the estimated speech spectra in each frequency band k.

$$D_k = \frac{\sum_{m=a_k}^{b_k} |S(m) - \hat{S}(m,\omega_0)|^2}{\sum_{m=a_k}^{b_k} |S(m)|^2}$$

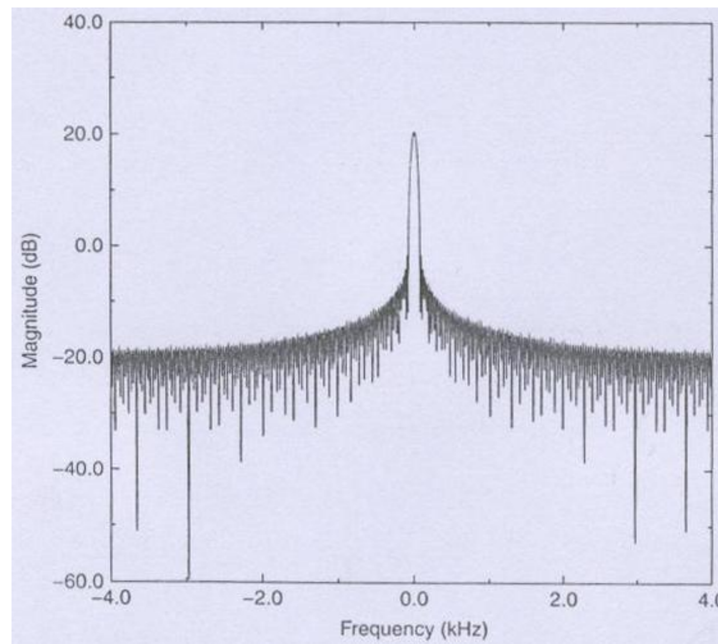
    - ω_0: the refined fundamental frequency after a post-processing
    - a_k, b_k: the first and last harmonic freq. bin indices in the k^{th} band
    - $S(m)$: the original speech spectrum
    - $\hat{S}(m,\omega_0)$: the reconstructed speech spectrum
    - Bandwidth of each band: a multiple (e.g. 3) of ω_0
      - Thus, number of bands is dependent on the pitch period of the frame.

MBE mixed voicing

- Voicing decision
  - The reconstructed speech spectrum is given by

$$\hat{S}(m,\omega_0) = \sum_{l=1}^{L} A_l(\omega_0) W_{l\omega_0}(m), \quad \lceil a_l \rceil \le m < \lceil b_l \rceil$$

    - $a_l = (l - 0.5)\omega_0$, $b_l = (l + 0.5)\omega_0$
    - L : the number of harmonics within the 4 kHz bandwidth
    - $W(m)$: the frequency response of a suitable window that will be centered at the l^{th} harmonic of ω_0
    - $A_l(\omega_0)$: the l^{th} harmonic amplitude



$$A_l(\omega_0) = \frac{\sum_{m=\lceil a_l \rceil}^{\lceil b_l \rceil} S(m) W_{l\omega_0}(m)}{\sum_{m=\lceil a_l \rceil}^{\lceil b_l \rceil} |W_{l\omega_0}(m)|^2}$$

MBE mixed voicing

- Voicing decision
  - Compare with the adaptive threshold from listening tests

$$\Delta_k(\omega_0) = (\alpha + \beta\omega_0)[1.0 - \varepsilon(k-1)\omega_0]M(E_0, E_{av}, E_{min}, E_{max})$$

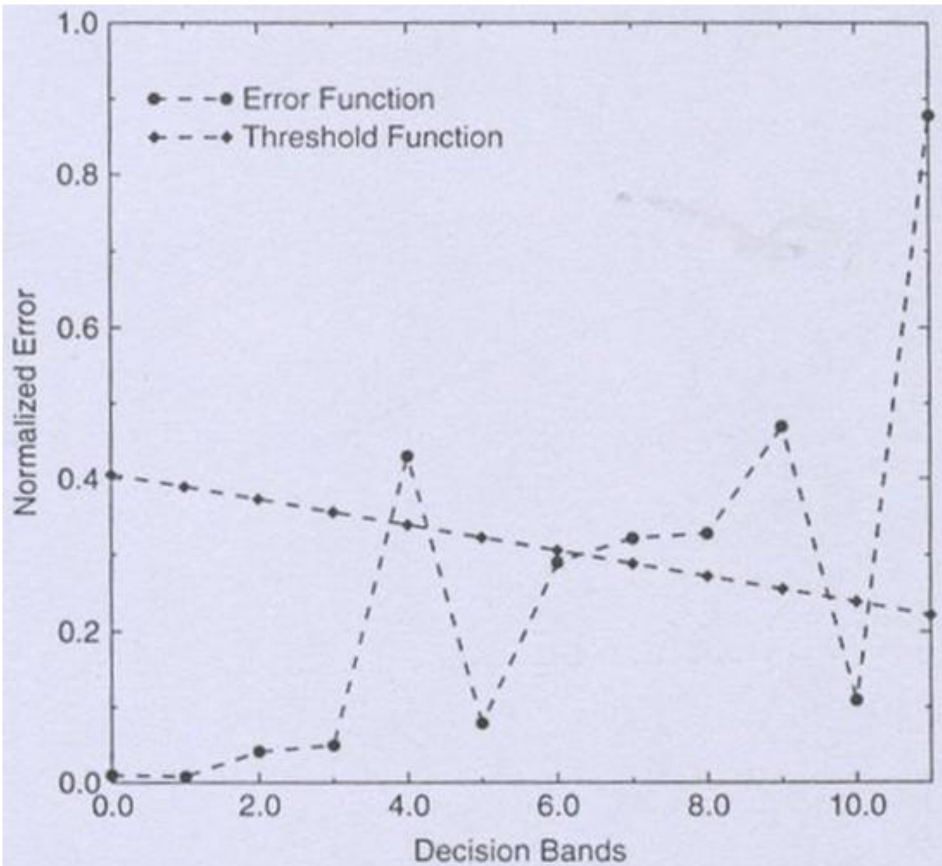
- α = 0.35, β = 0.557, ε = 0.4775 are the factors that give good subjective quality.
- $M()$ is the adaptation factor that controls the decision threshold for V/UV decisions with μ = 0.0075,

$$M(E_0, E_{av}, E_{min}, E_{max}) = \begin{cases} 0.5, & E_{av} < 200 \\ \dfrac{(E_0 + E_{min})(2E_0 + E_{max})}{(E_0 + \mu E_{max})(E_0 + E_{max})}, & E_{av} \geq 200 \text{ and } E_{min} < \mu E_{max} \\ 1.0, & \text{otherwise} \end{cases}$$

- So, if $D_k < \Delta_k(\omega_0)$, then the band is regarded as voiced, elsewhere as unvoiced.

MBE mixed voicing

- Voicing decision
  - Typical example of the error and threshold functions for one frame



From the threshold function, since ω_0 in male speech is relatively low, a lower band of male speech will be likely to be declared voiced, and a higher band of female speech will be likely to be declared unvoiced.

Split-band mixed voicing

- One drawback of MBE mixed voicing
  - More than one bit (12 bits in the previous) will be needed.
- Observation from experiments
  - If a spectrum contains an unvoiced band between two voiced bands, the unvoiced signal in the middle is usually small.
  - Thus if it is declared as voiced, subjectively it would not make much difference in speech quality.
- So, simply split the full band into low frequency band for voiced and high frequency band for unvoiced. → Split-band mixed voicing
  - Based on a more reliable measure such as voicing likelihood
  - Simply transmit the quantized voicing cut-off frequency.
    - Only 4 bits for the previous case

Summary of lecture

- Pitch estimation
  - Detection of pitch period
  - Time domain methods
    - AMDF, ACF, N-ACF
  - Frequency domain methods
    - Harmonic peak detection method, Spectrum similarity method
  - Time- and frequency-domain methods
    - Spectro-temporal autocorrelation (STA) PDA
  - Pre- and post-processing techniques
    - Spectrum flattening, Pitch tracking, Correction of multiple- or half-pitch errors
- Voiced-unvoiced classification
  - Hard-decision voicing
  - Soft-decision voicing