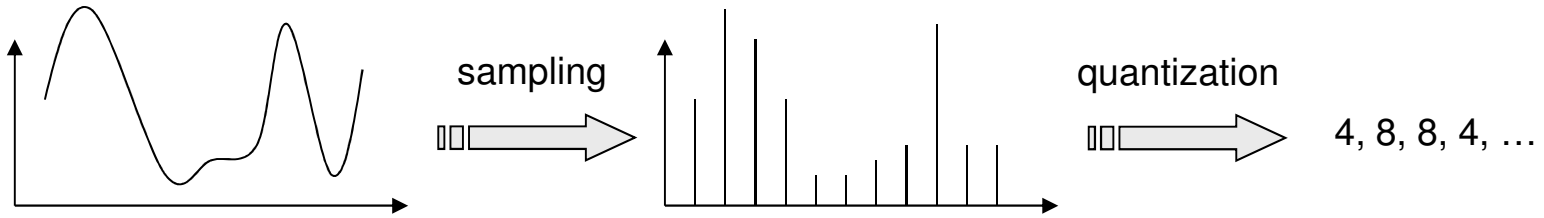
Speech Codecs

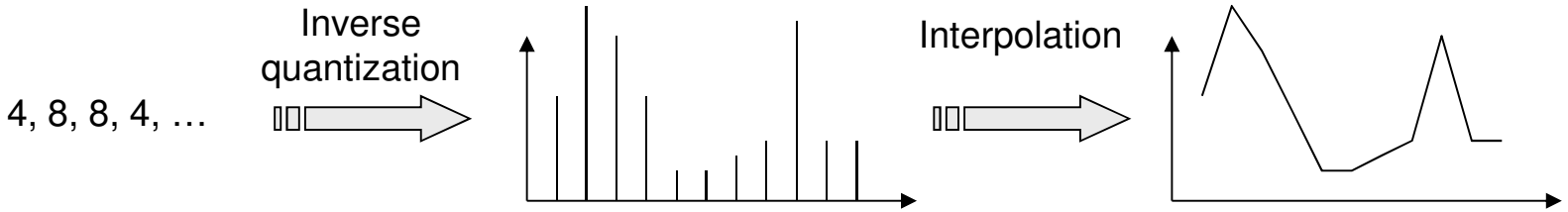
National Chiao Tung University
Chun-Jen Tsai
12/15/2014

Digital Representation of Sound

- Pulse Code Modulation (PCM) is the simplest representation of digital sound:
  - Encode:

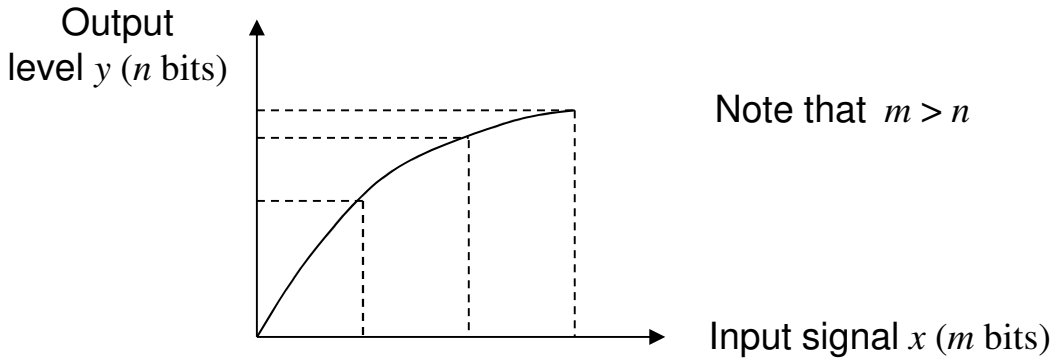


  - Decode:



Audio Quantization Law

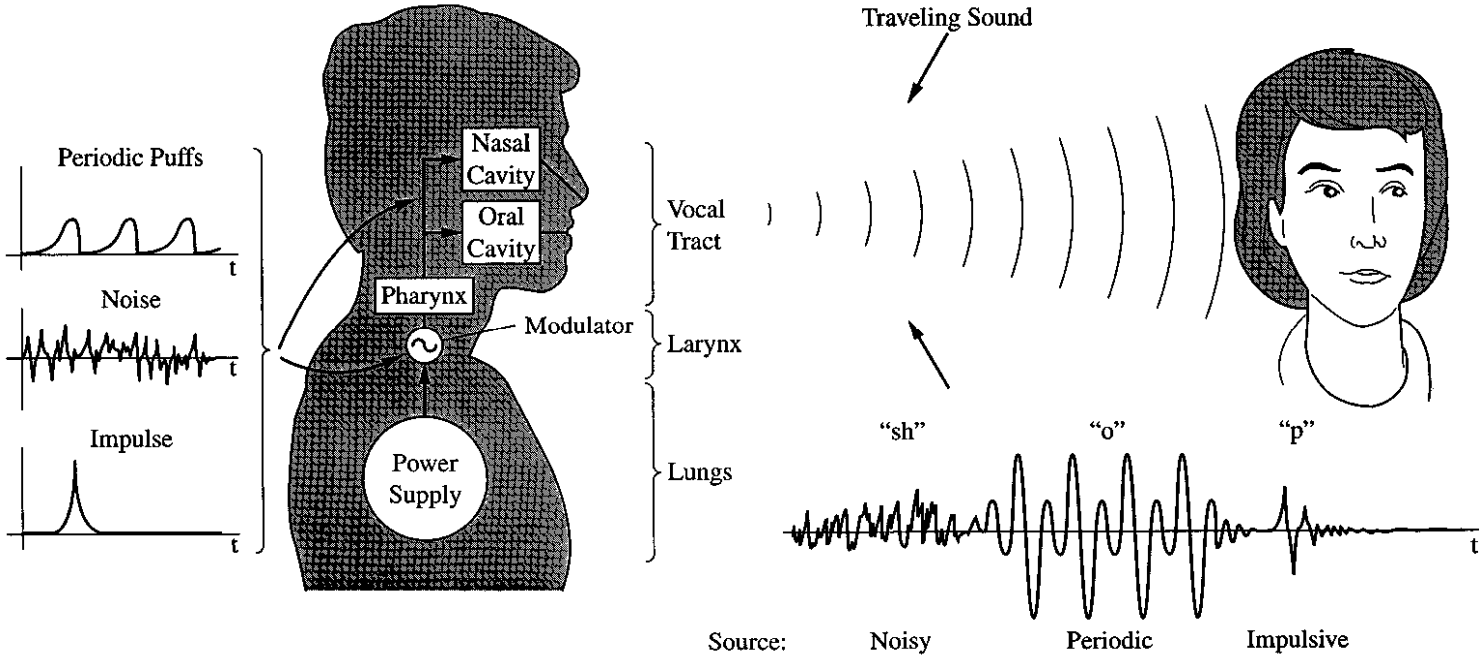
- Non-uniform quantizer are often used for audio:



  - Common non-linear quantization functions:
    - power-law ($y = |x|^p$, $p = 0.75$)
    - logarithmic law (A-Law, μ-law)

Speech Model

- Speech signal can be modeled as the output of a signal sequence passing though a liner filter†



† T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002

Speech Coding Principle

- A vocal tract model, driven by coded bits (excitations) produces synthesized speech:

- Types of speech
  - Voiced:
    - From vibration of vocal cords, e.g., “A”, “E”, “I”, “O”, “U”
    - Characterized by “pitch frequency” associated with the tone
  - Unvoiced Sound:
    - Breathing sound, e.g., “shhhh,” “thhhh,” …, etc.

Wide Band v.s. Narrow Band Speech

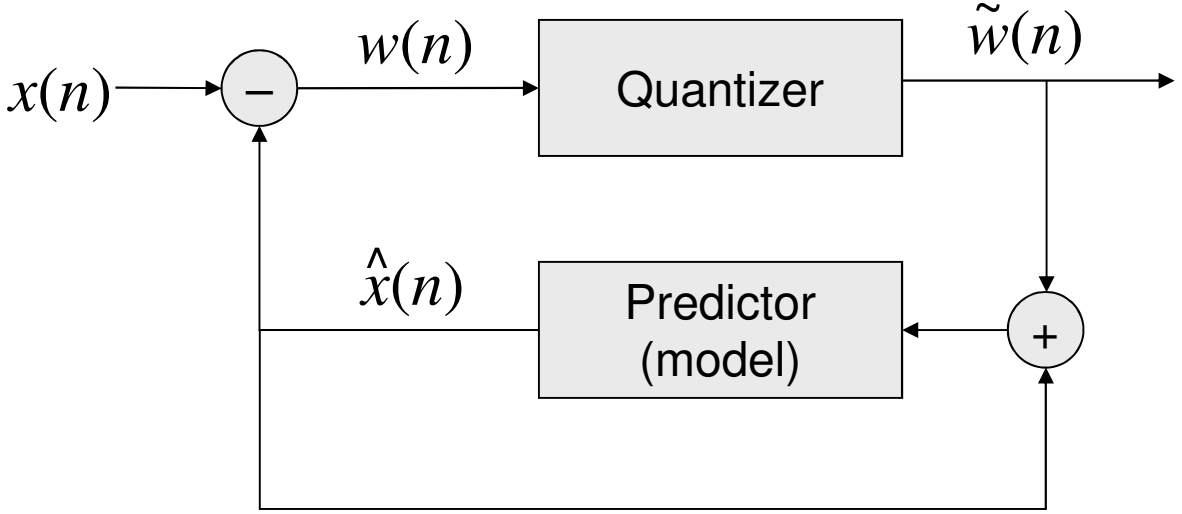
- Narrow band (phone quality) speech signal
  - Signal bandwidth 0.2~3.4 kHz
  - Sampling rate 8 kHz
  - Coded bitrate 3~13 kbps
- Wide band (ISDN quality) speech signal
  - Signal Bandwidth 0.05~7 kHz
  - Sampling rate 16 kHz
  - Coded bitrate 16~64 kbps

Classification of Speech Coding

- Vocoders
  - Subband coders (frequency domain coders)
  - Linear predictive coders (LPC)
- Waveform coders
  - Delta modulation (DM)
  - Differential pulse code modulation (DPCM)
- Analysis-by-Synthesis
  - Code-excitation Linear Predictive Coders (CELP)

Predictive Coding Revisited

- Predictive coding is used in video as well as in speech coding:



Predictive Coding Model (1/2)

- A signal is "predicted" by some linear time-invariant (LTI) model, such as the auto-regressive (AR) model:

  Signal $s(n)$ is a zero-mean, Gauss-Markov sequence

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + u(n)$$
$$= \hat{s}(n) + u(n) = \text{prediction} + \text{innovation},$$

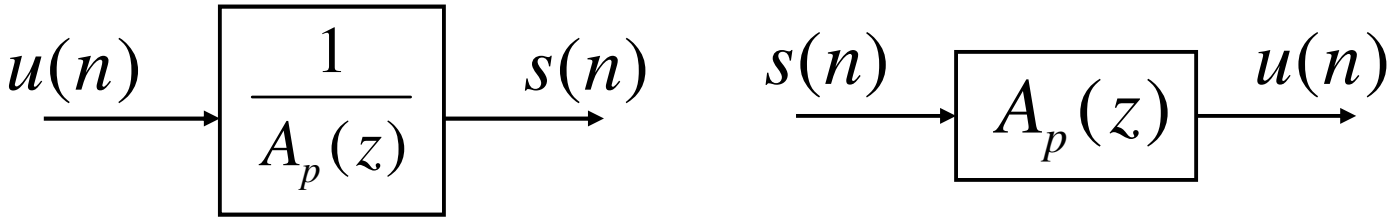
  where $u(n)$ is white (i.i.d.) Gaussian with

$$E[u(n)] = 0, \quad E[u^2(n)] = \text{constant}.$$

Predictive Coding Model (2/2)

- In z-domain, the model is described as follows:

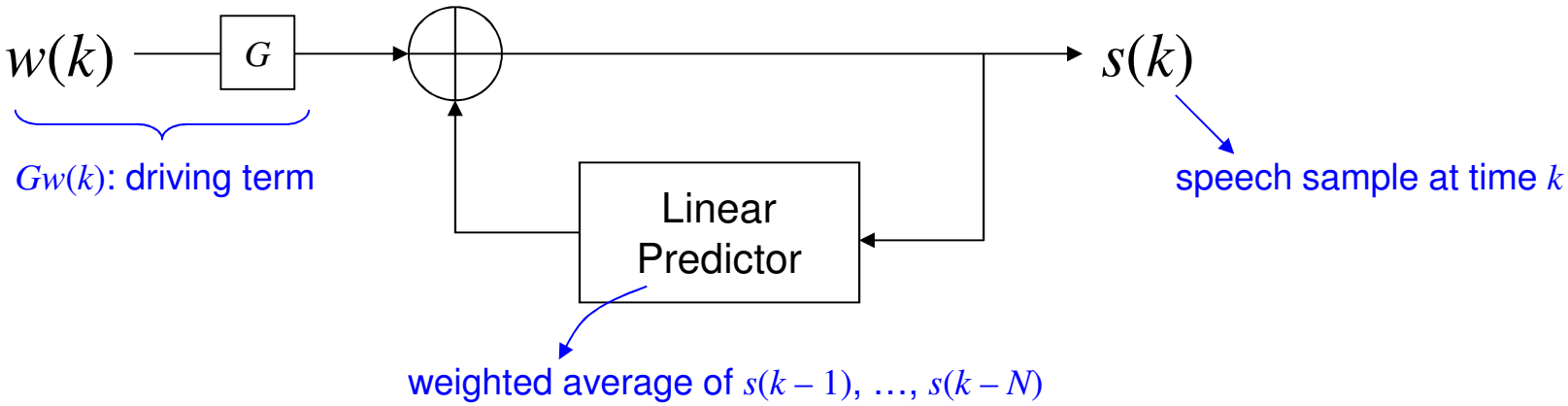
$$A_p(z) \equiv 1 - \sum_{k=1}^{p} a(k) z^{-k}$$



- Question: how to find $A_p(z)$ for an input signal $s(n)$?

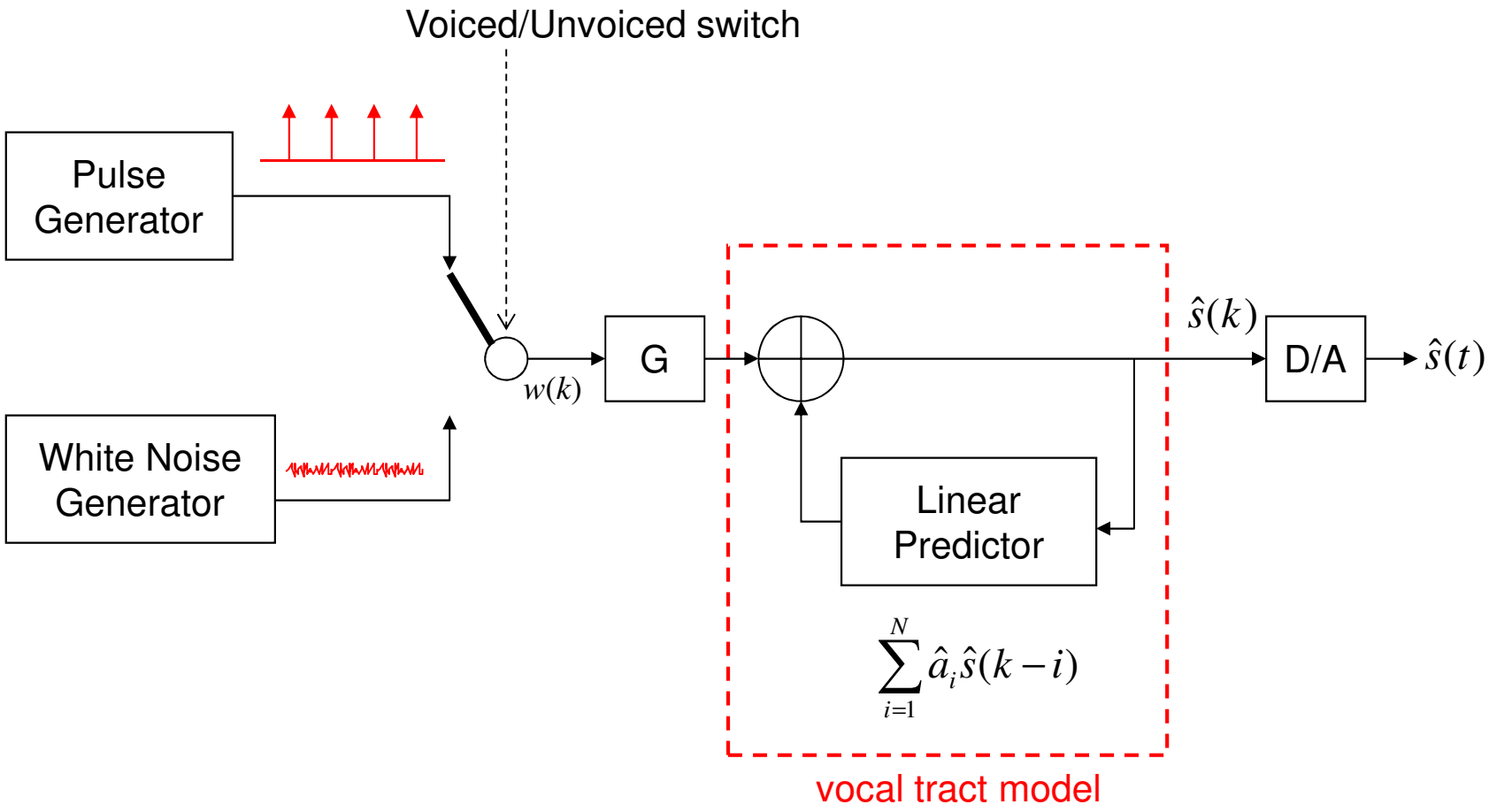
Linear Predictive Coders

- Linear Prediction Model



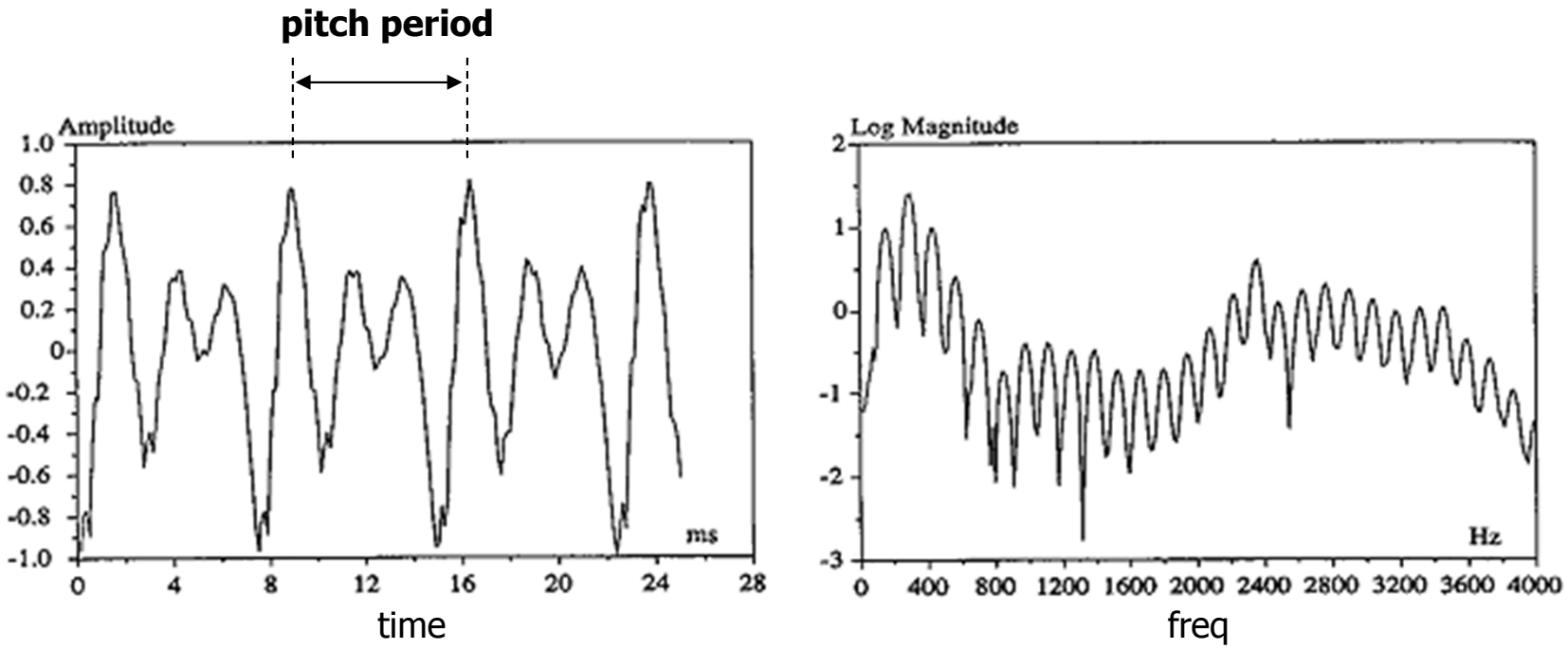
$$\Rightarrow \quad s(k) = \sum_{i=1}^{N} a_i s(k-i) + Gw(k)$$

LPC Receiver (Decoder)



Spectrum of Voiced Speech

- The frequency spectrum of a voiced signal has distinctive structure†:



† T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002

Parameters to Be Transmitted

- Coefficients of the model $\{ a_i : i=1,2,\ldots,N \}$
- Voiced/unvoiced (1-bit flag)
- Gain factor G
- Pitch period for voiced frame

$\rightarrow$ Estimation of a_i's is called AR model identification

LPC Principle

- LPC is based on least-squares estimation, that is, minimize the mean square error:

$$\varepsilon = \frac{1}{M}\sum_{k=1}^{M}\left[s(k)-\hat{s}(k)\right]^2$$

$$= \frac{1}{M}\sum_{k=1}^{M}\left[\, s(k)-\sum_{i=1}^{N} a_i s(k-i)\right]^2$$

where

$$\hat{s}(k) = \sum_{i=1}^{N} a_i s(k-i)$$

M is the block of samples used in coding analysis, usually called a frame (typically 10-25 msec)

Solution I: Covariance Method

- Compute $\partial\varepsilon/\partial a_j=0$:

$$\sum_{i=1}^{N} a_i \left\{ \sum_{k=1}^{M} s(k-i)s(k-j) \right\} = \sum_{k=1}^{M} s(k)s(k-j)$$

$$\Longrightarrow \quad \sum_{i=1}^{N} a_i \phi_{ij} = \phi_{j0}, \quad j = 1,2,\ldots,N$$

$$\Longrightarrow \quad \mathbf{\Phi a} = \boldsymbol{\psi}$$

predictor coefficients

covariance matrix

Solution II: Autocorrelation Method

- Compute $\partial\varepsilon/\partial a_j=0$, but assume samples outside current frame are zeros:

$$\sum_{i=1}^{N} a_i \sum_{k=-\infty}^{\infty} s(k-i)s(k-j) = \sum_{k=-\infty}^{\infty} s(k)s(k-j)$$

Let $m = k - j$, for $j = 1, 2, \ldots, N$.

$$\rightarrow \sum_{i=1}^{N} a_i \sum_{m=-\infty}^{\infty} s(m)s(m+j-i) = \sum_{m=-\infty}^{\infty} s(m)s(m+j)$$

Define $$R(j) = \sum_{m=-\infty}^{\infty} s(m)s(m+|j|)$$

$$\rightarrow \sum_{i=1}^{N} a_i R(|j-i|) = R(j), \quad j = 1,2,\ldots,N \qquad \text{Eq (1)}$$

Durbin Solution† of Eq (1)

1. Initial Error: $E^{(0)} = R(0)$
2. Compute
$$k_i = \left\{ R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right\} \Big/ E^{(i-1)}, \quad 1 \le i \le N$$
3. Then
$$a_i^{(i)} = k_i$$
$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \le j \le i-1$$
4. Update MSE: $E^{(i)} = (1 - k_i^2) E^{(i-1)}$
5. Repeat Step 2-4 for $i = 1, 2, \ldots, N$

$\rightarrow k_i$'s are called partial correlation (PARCOR) coefficients

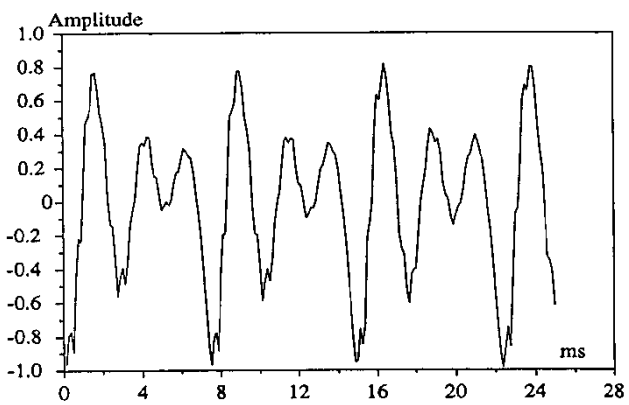
† J. Durbin, "The Fitting of Time Series Models," *Rev. Inst. Inter. Statist.* 28:233-243, 1960

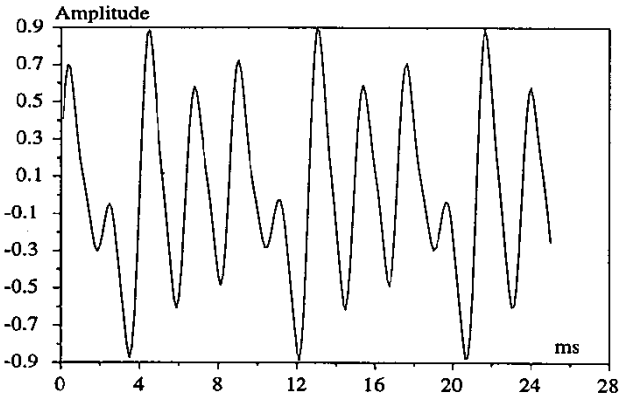
Voiced/Unvoiced Decision

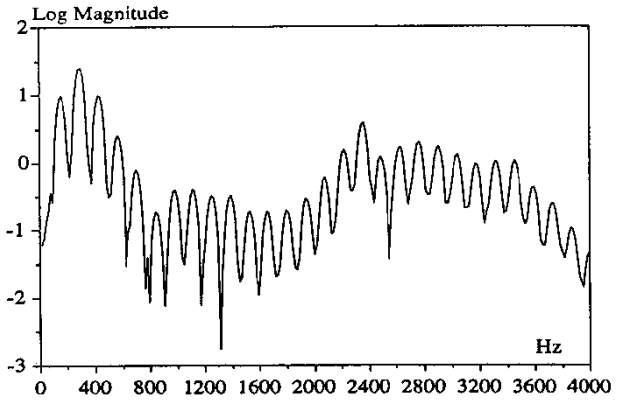
- Voiced/unvoiced decision controls the type of excitation function at the receiver
- “Unvoiced too much” $\rightarrow$ “ breathy”
- “Voiced too much”$\rightarrow$ “ buzzy”
- V/UV decision is the most problematic part of an LPC system
- Pitch period estimation for voiced excitation is also very difficult

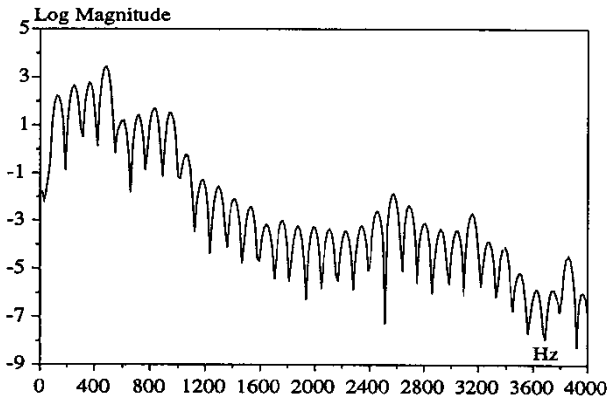
Voiced Sounds

- **Left**: /I/ in “bit”; **Right**: /U/ in “foot”









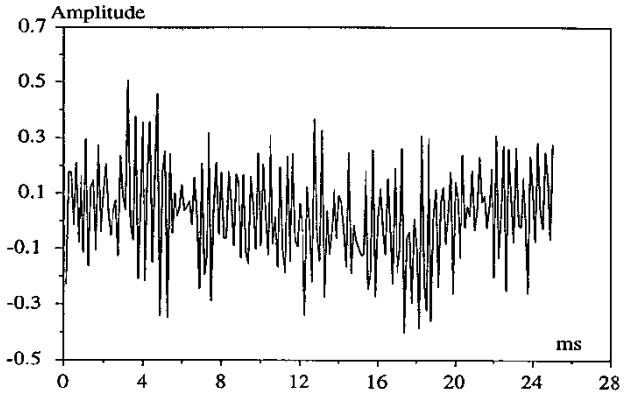
waveforms

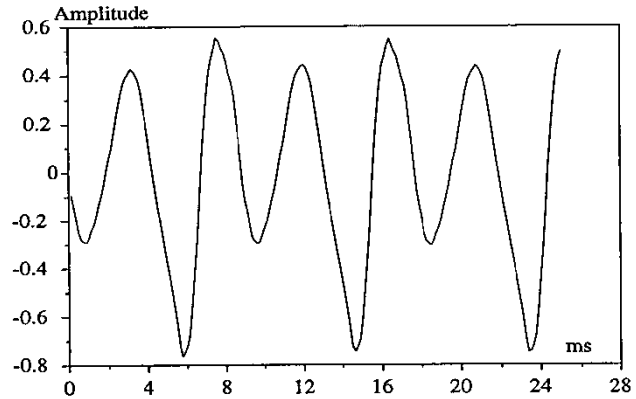
spectra

† T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002

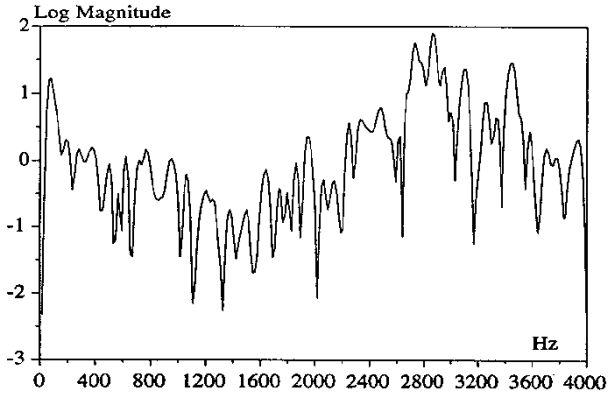
Unvoiced Sounds

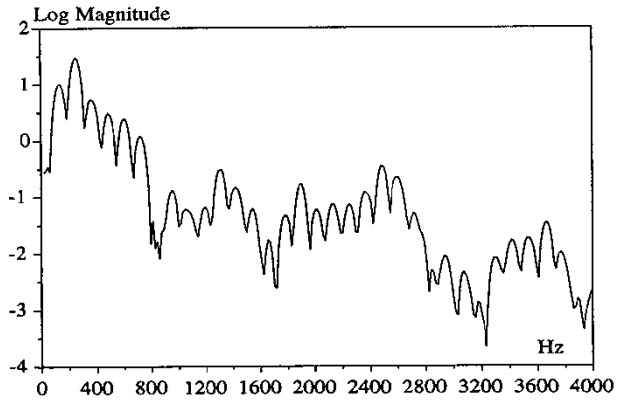
- **Left**: /sh/ in “shop”; **Right**: /m/ in “map”





waveforms





spectra

† T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002

Excitation Gain

- During "carry-over phase," G is calculated from one pitch period to the next to ensure input energy equals output energy
- During "quenching phase," G is calculated based on:

$$\begin{cases} G = \sqrt{P} \cdot \left[R(0) - \sum_{i=1}^{N} a_i R(i) \right]^{\frac{1}{2}}, & \text{if voiced} \\ G = \left[R(0) - \sum_{i=1}^{N} a_i R(i) \right]^{\frac{1}{2}}, & \text{if unvoiced} \end{cases}$$

† B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, 50(2):637-655, 1971

Differential Pulse Code Modulation

- Predict the current original (source) sample, s, using the previously coded samples, $\{\tilde{s}\}$.

  Prediction:

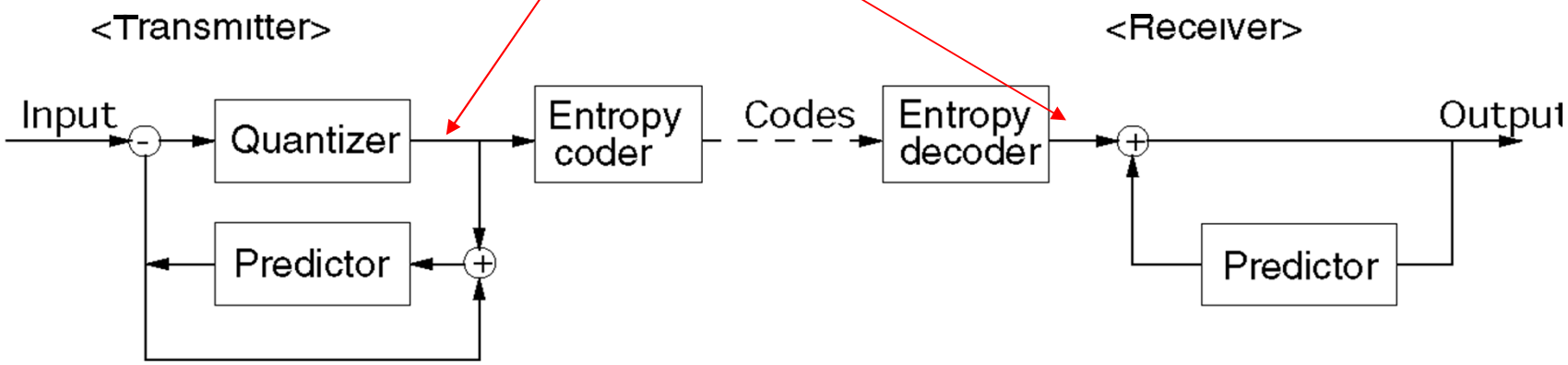
$$\hat{s}(n) = \sum_{k=1}^{N} h(k)\tilde{s}(n-k)$$

  In the above equation, $h(k)\tilde{s}(n-k)$ is *not* equal to $a(k)s(n-k)$ in the source model. Hence, in general, the optimum $h(k) \neq a(k)$.

DPCM Mechanism

- Prediction error: $e(k) = s(k) - \hat{s}(k)$
- Quantization:

$$e_q(k) = e(k) // \Delta_k$$

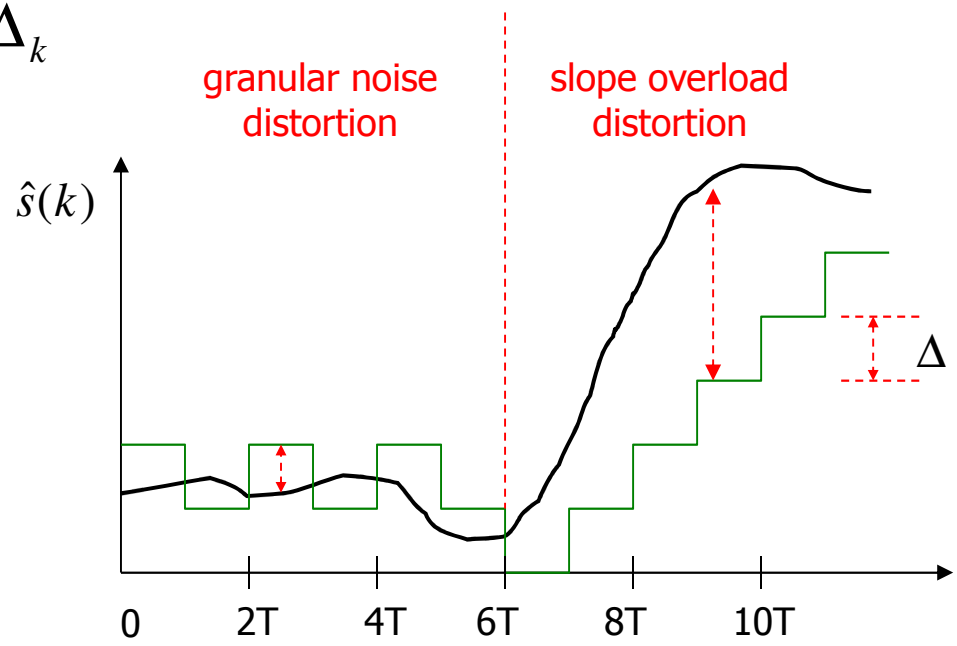


Delta Modulation

- Delta modulation is a DPCM with a first order predictor $\hat{s}(k) = a\hat{s}(k-1),\ \therefore e(k) = s(k) - a\hat{s}(k-1)$

  The quantizer has only two levels, $\pm\Delta$

$$e_q(k) = \mathrm{sgn}\{e(k)\}\Delta_k$$



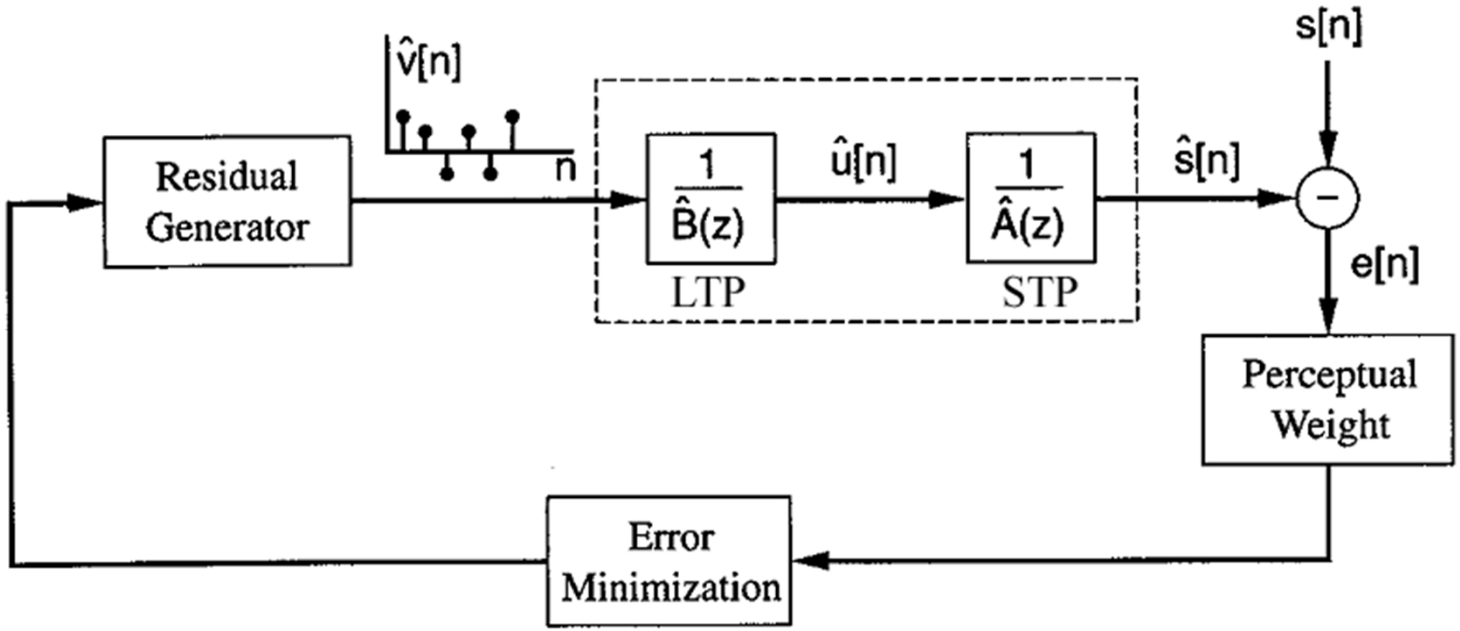
Multi-Pulse LPC

- Two main problems with LPC:
  - Voiced/Unvoiced decision
  - Pitch extraction
- Multi-Pulse LPC (MPLPC) use a more sophisticated excitation model:

  $\rightarrow$ allow several impulses to be used as the synthesis filter excitations for each speech frame

MPLPC Block Diagram

- MPLPC is an analysis-by-synthesis (AbS) coder



Features of AbS Coders

- Two types of linear predictors:
  - Short-term predictor (STP)

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + u(n)$$

  - Long-term predictor (LTP), for pitch prediction

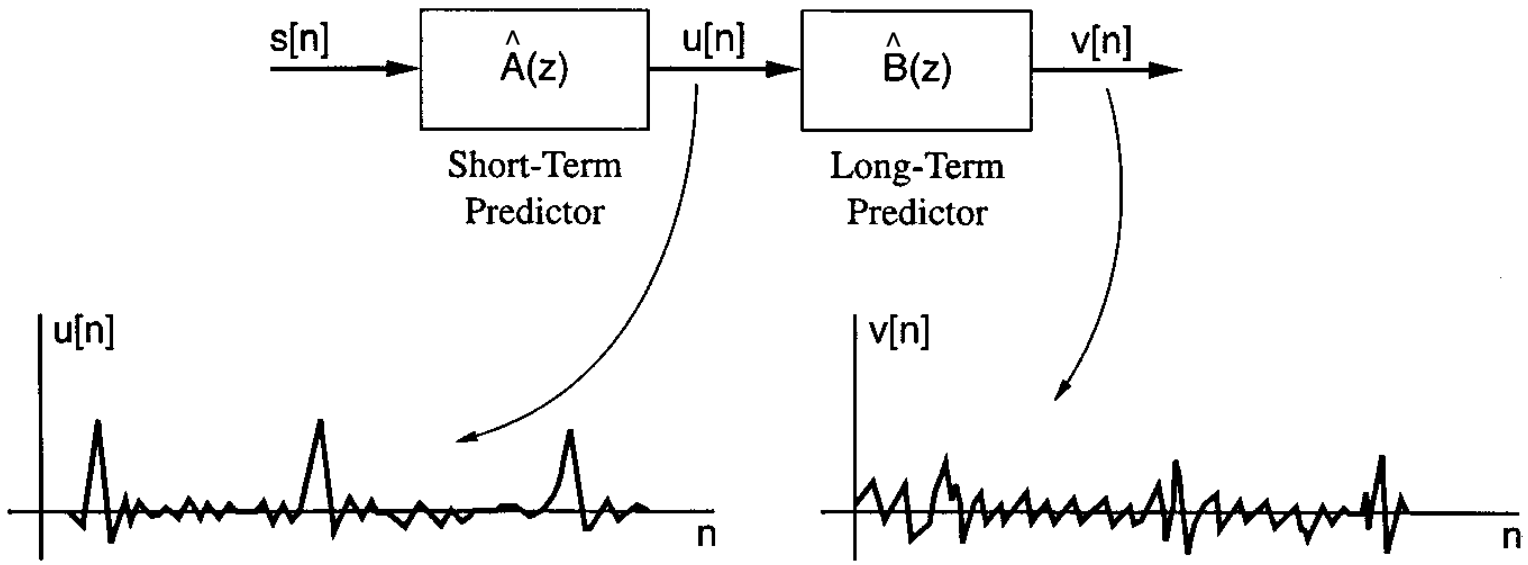
$$u(n) = \sum_{k=1}^{l} b(k)u(n-D-k) + v(n), \quad D : \text{pitch period}$$

- Perceptual weighting filter

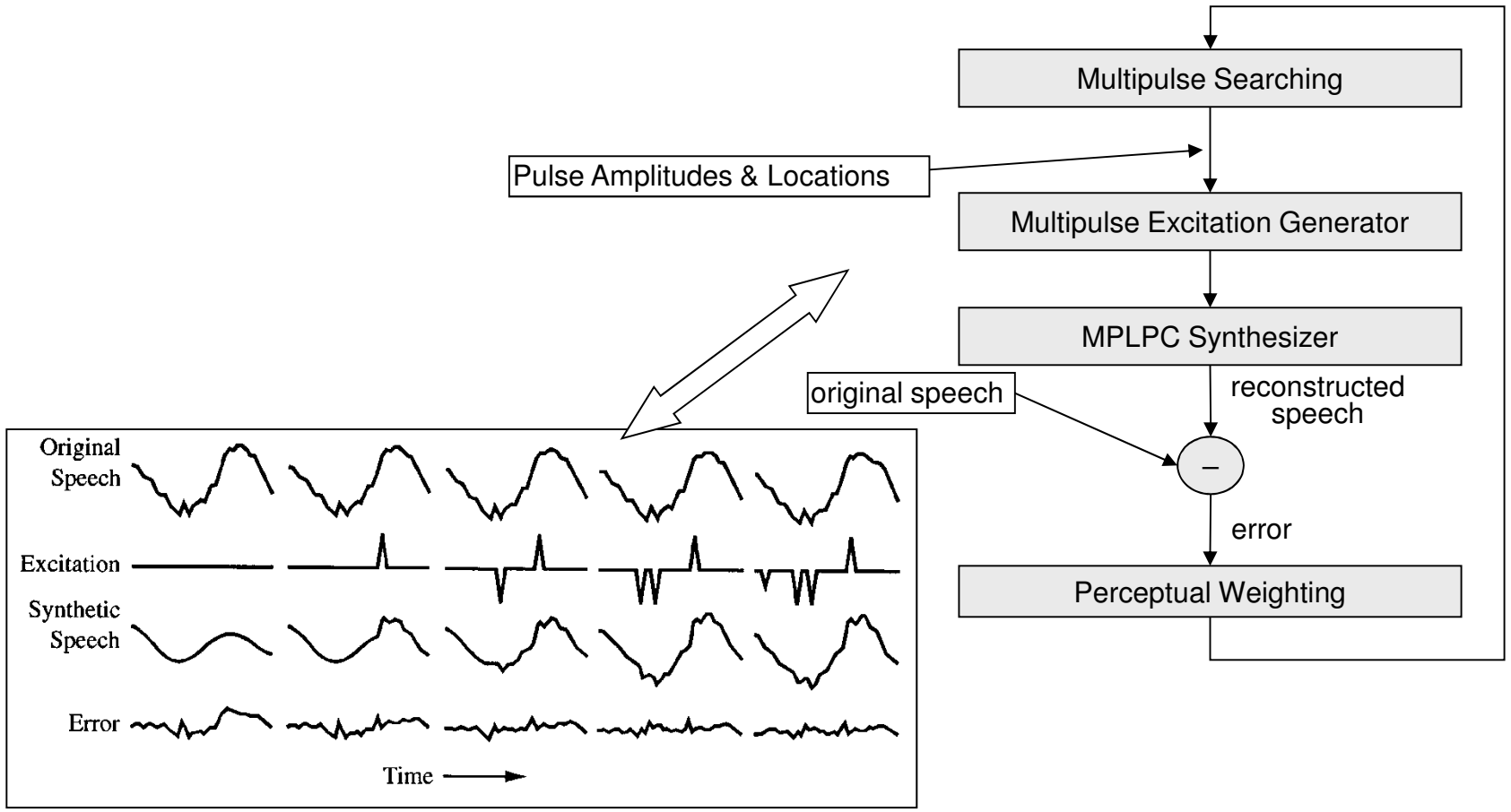
$$W(z) = \frac{\hat{A}(z)}{\hat{A}(z/\gamma)}, \quad \gamma \sim [0.8, 0.9].$$

Short- and Long-term Predictions

- The effect of short-term and long-term prediction on an input signal is as follows:



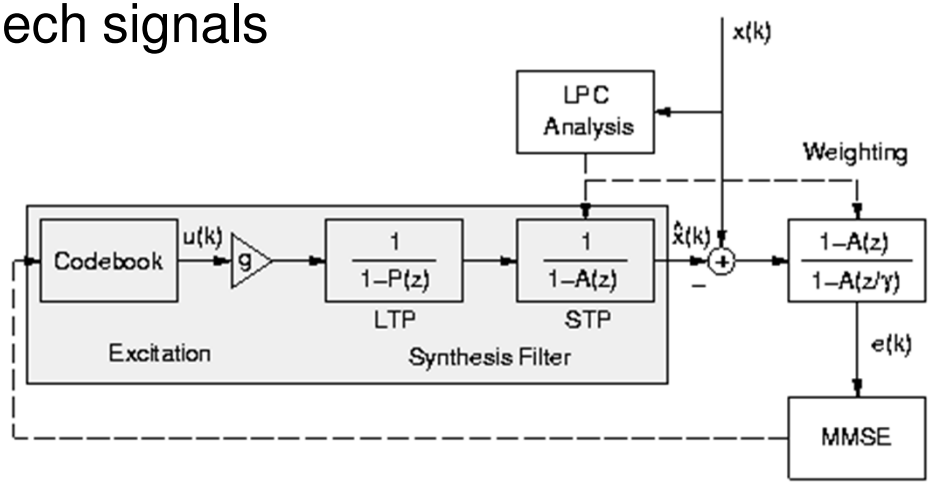
MPLPC Pulse Search Loop†



† T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002

Code Excited LPC (CELP)

- Most popular narrow band speech coding method
- An extension to the analysis-by-synthesis approach used in MPLPC
- Principle:
  - After short- and long-term prediction, the error residuals can be modeled by an i.i.d. random variable (called excitation sequence)
  - Only finite number of excitation sequences are required for representing speech signals



CELP Procedure

1. Partition speech signal into 20-30ms frames
2. Perform short-term LP analysis $\rightarrow$ LPC coeffs.
3. Perform long-term LP analysis $\rightarrow$ pitch period, scaling factor
4. Select the best excitation codeword from a code book using the synthesis filters derived in step 2 & 3

Algebraic Codebook for CELP

- Fixed codebook with fixed amplitude
- Only signs and positions needs to be transferred
- GSM-EFR example
  - Each frame (20 ms) is divided into 4 sub-frame of 5ms (40 samples)
  - Each sub-frame is divided into 5 tracks of 8 interlaced positions
  - Two pulse positions (and signs) are coded and transmitted for each track

Track	Pulses	Amplitudes	Positions
1	0, 5	±1, ±1	0,5,10,15,20,25,30,35
2	1, 6	±1, ±1	1,6,11,16,21,26,31,36
3	2, 7	±1, ±1	2,7,12,17,22,27,32,37
4	3, 8	±1, ±1	3,8,14,18,23,28,33,38
5	4, 9	±1, ±1	4,9,15,19,24,29,34,39

Algebraic Codebook used in GSM-EFR (06.60)

Subjective Speech Quality Criterions

- Mean Opinion Score (MOS) are often used to evaluate subjective quality of multimedia codecs
- The scale of MOS are often from 1~5 for audio quality tests:
  - 5 Excellent Imperceptible
  - 4 Good Perceptible but not annoying
  - 3 Fair (Perceptible and) Slightly annoying
  - 2 Poor Annoying (but not objectionable)
  - 1 Bad Very annoying (objectionable)

Speech Coding Performance

- Mean Opinion Score (MOS) Performance

Standards	Typical rates (year)	Quality: MOS (1-5) / Network
G.711 (PCM)	64 kbits/s (1972)	4.4 (PSTN)
G.721 (ADPCM)	32 kbits/s (1984)	4.1 (PSTN)
GSM 06.10 (RPE-LTP)	13 kbits/s (1991)	3.6 (Cellular)
G.729 (ACELP)	8 kbits/s (1995)	~4.2 (Internet VOIP)
G.723.1 (MP-MLQ/ACELP)	6.3, 5.3 kbits/s (1995)	~4.0 (Internet VOIP)
GSM-AMR (ACELP)	4.75-12.2 kbits/s (1999)	~3.9 (3GPP)
iLBC (Block-indep. LPC)	13.33, 15.2 kbits/s (2000)	>4.0 (Internet VOIP)
MOS: 5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad		

Internet Low Bitrate Codec (iLBC)

- iLBC is a narrow-band speech codec designed for robust voice communication over IP
  - Developed by Global IP Sound in 2000, became IETF RFC-3951 in 2002
  - Used in Skype as the speech codec
- The iLBC codec supports two basic frame lengths
  - 13.3 kbps with an encoding frame length of 30 ms
  - 15.2 kbps with an encoding frame length of 20 ms
- iLBC codes full 4kHz speech frequency band (G.723.1 codes only 300 Hz ~ 3400Hz)

Highlights of iLBC

- A block-based LPC codec, the residual signal is coded using an adaptive codebook
  - Unlike CELP-based codecs, each speech frame in iLBC is independently coded
- For 30 ms mode, use two LPC analysis windows to operate on the 240-sample frame
  - The residuals after LPC is divided into 6 sub-frames, two sub-frames with highest residuals are identified
  - 57 samples of these two sub-frames are encoded as the start state, the remaining samples are coded using the adaptive codebook applied both forward and backward in time, starting from the start state vector

Packet Loss Performance

- iLBC is more robust for VOIP applications