Multimedia communications ECP 610

Omar A. Nasr

omaranasr@ieee.org

April, 2015

Speech coding (compression)

- A procedure to represent a digitized speech signal using as few bits as possible, maintaining at the same time a reasonable level of speech quality.
- The standard defines the compression algorithm, not the platform of implementation (DSP, GPP, FPGA, ASIC, .. etc)
- Uncoded speech: 8 kHz sampling x 16bits/sample = 128kbps

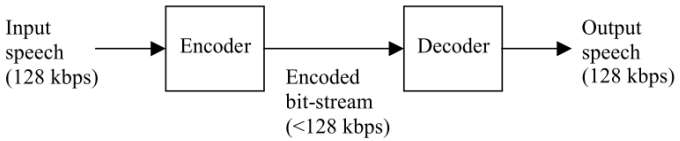


Figure 1.2 Block diagram of a speech coder.

- Issues: effects due to the channel errors

A good speech coder

- Low Bit rate
- High speech quality (intelligibility, naturalness, pleasantness, and speaker recognizability)
- Robustness across Different Speakers / Languages (males, females, adults, kids)
- Robustness in the Presence of Channel Errors
- Low Memory Size and Low Computational Complexity
- Low Coding Delay

Coder delay

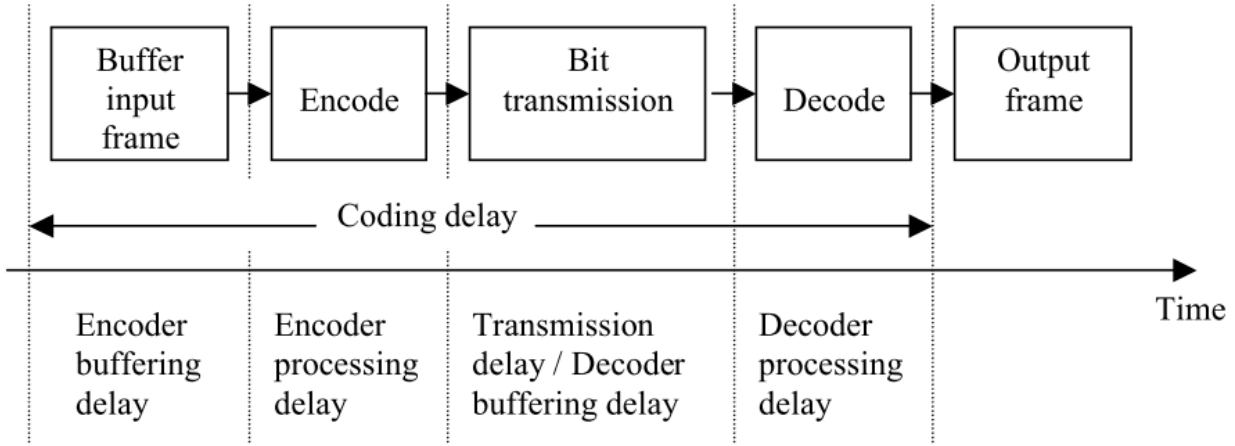


Figure 1.4 Illustration of the components of coding delay.

Classification of speech coders

TABLE 1.1 Classification of Speech Coders According to Bit-Rate

Category	Bit-Rate Range
High bit-rate	>15 kbps
Medium bit-rate	5 to 15 kbps
Low bit-rate	2 to 5 kbps
Very low bit-rate	<2 kbps

Classification by coding technique

- Waveform coders
  - preserve the original shape of the signal waveform, and hence the resultant coders can generally be applied to any signal source.
  - Data rates 24-64kbps
  - Can be measured by SNR
- Parametric coders
  - the speech signal is assumed to be generated from a model, which is controlled by some parameters
  - Does not preserve the shape of the signal
  - Low bit rates (can reach less than 2kbps)

Classification by coding technique

- Hybrid coders
  - Parametric + waveform
  - Assume a model, then add more parameters to reach a waveform that is close to the original waveform
  - Medium bit rate

Parametric speech coding

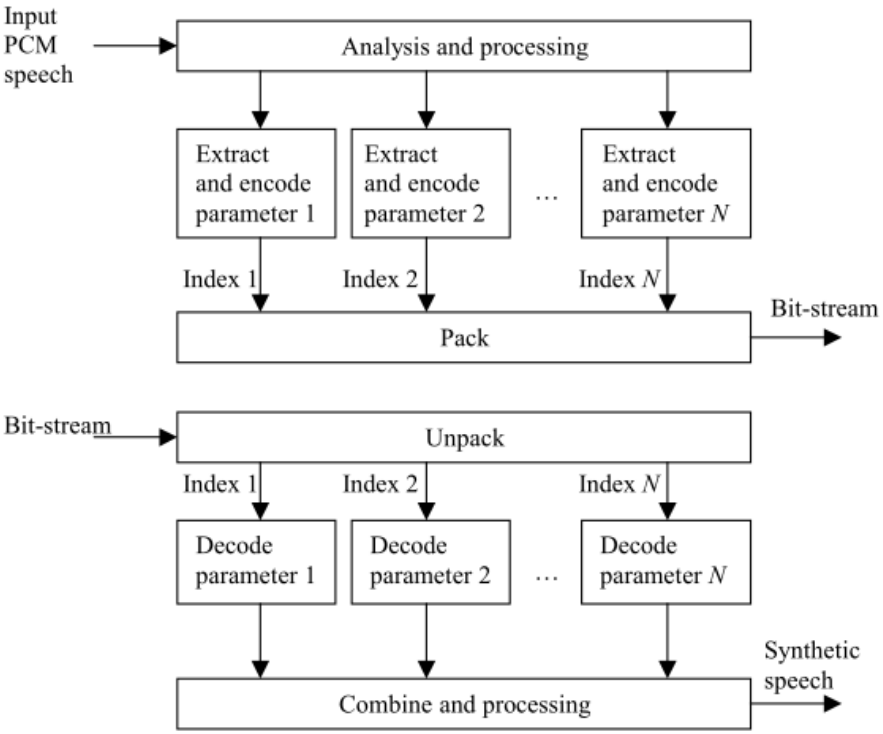


Figure 1.12 General structure of a speech coder. *Top*: Encoder. *Bottom*: Decoder.

Models

- Human auditory systems
- Speech production model
- Phase perception

Linear prediction

- Basic idea: approximate each speech sample as a linear combination of the past few samples
- Weights minimizes the mean square prediction error
- The resultant weights are the Linear Prediction Coefficients (LPCs)
- LPCs change from frame to frame
- Another interpretation of LP is as a spectrum estimation method
- By computing the LPCs of a signal frame, it is possible to generate another signal in such a way that the spectral contents are close to the original one

Linear prediction

- Prediction … redundancy removal
- The problem of linear prediction

$$\hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-i], \qquad e[n] = s[n] - \hat{s}[n].$$

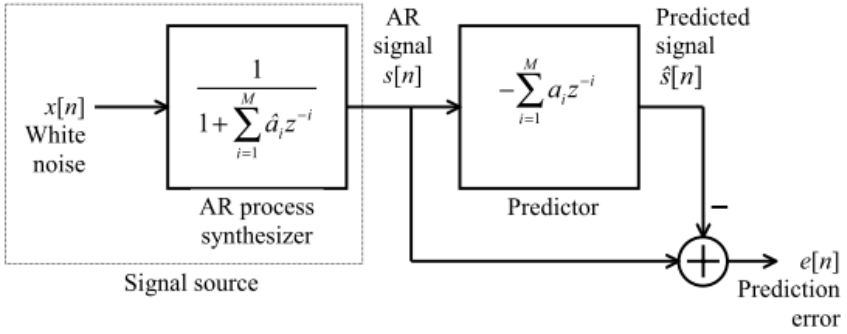


Figure 4.1 Linear prediction as system identification.

$$J = E\{e^2[n]\} = E\left\{\left(s[n] + \sum_{i=1}^{M} a_i s[n-i]\right)^2\right\}$$

Derivation of the LPCs

$$\frac{\partial J}{\partial a_k} = 2E\left\{\left(s[n] + \sum_{i=1}^{M} a_i s[n-i]\right)s[n-k]\right\} = 0$$

$$E\{s[n]s[n-k]\} + \sum_{i=1}^{M} a_i E\{s[n-i]s[n-k]\} = 0$$

$$\sum_{i=1}^{M} a_i R_s[i-k] = -R_s[k]$$

$$\mathbf{R_s a} = -\mathbf{r_s},$$

$$\mathbf{R_s} = \begin{pmatrix} R_s[0] & R_s[1] & \cdots & R_s[M-1] \\ R_s[1] & R_s[0] & \cdots & R_s[M-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_s[M-1] & R_s[M-2] & \cdots & R_s[0] \end{pmatrix},$$

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_M]^T,$$

$$\mathbf{r_s} = [R_s[1] \quad R_s[2] \quad \cdots \quad R_s[M]]^T.$$

$$\mathbf{a} = -\mathbf{R_s}^{-1}\mathbf{r_s}.$$

Prediction Gain

$$PG = 10 \log_{10}\left(\frac{\sigma_s^2}{\sigma_e^2}\right) = 10 \log_{10}\left(\frac{E\{s^2[n]\}}{E\{e^2[n]\}}\right)$$

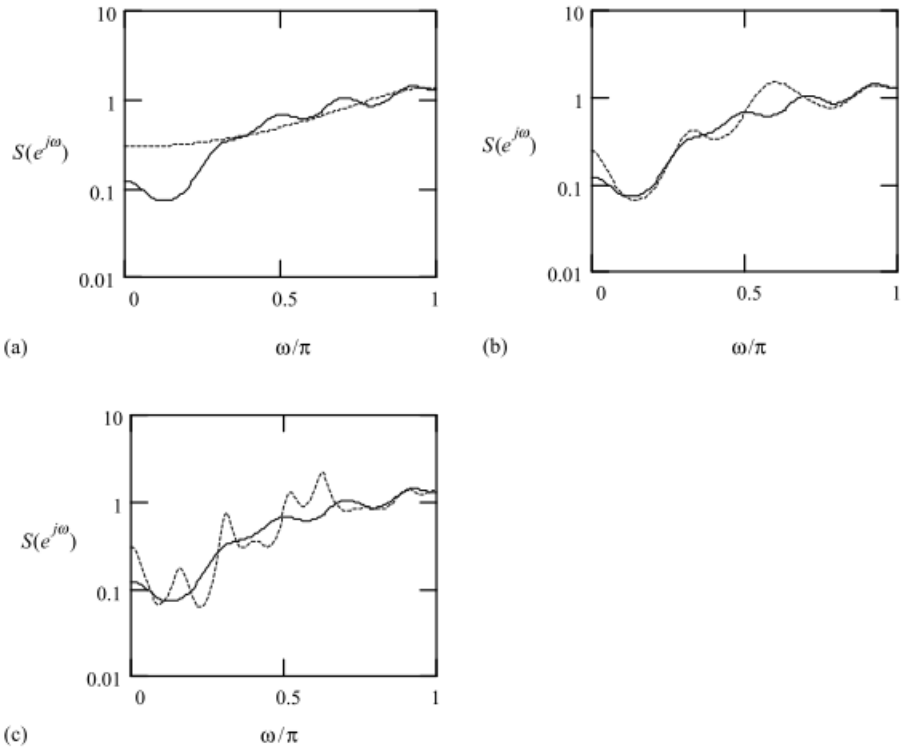


Figure 4.5 Plots of PSD (solid trace) together with several estimates (dot trace) using the LPC found with (a) $M = 2$, (b) $M = 10$, and (c) $M = 20$.

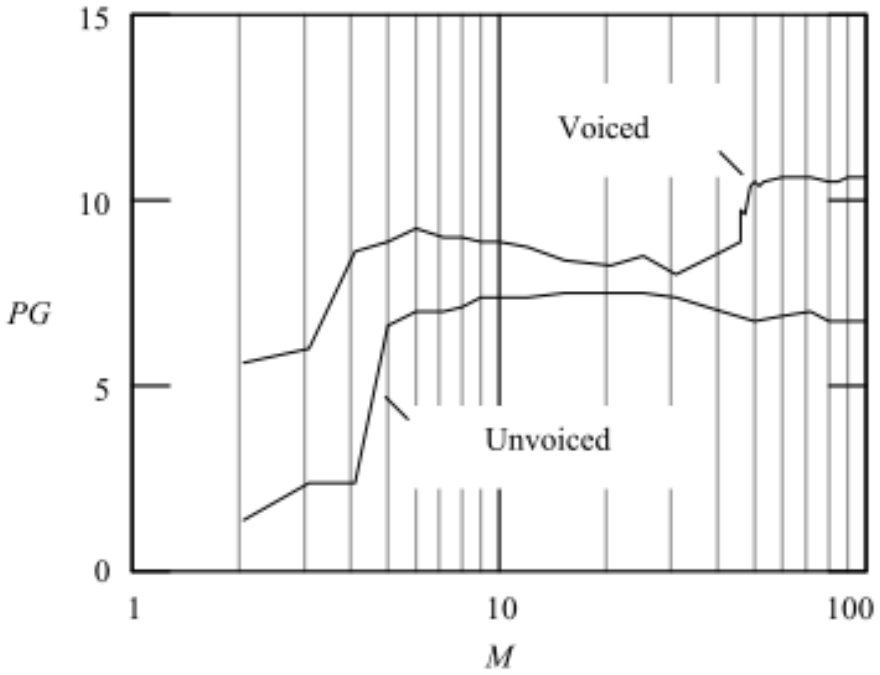


Figure 4.9 Plot of prediction gain (PG) as a function of the prediction order (M) for the signal frames in Figure 4.6.

For voiced frames, capture the envelop

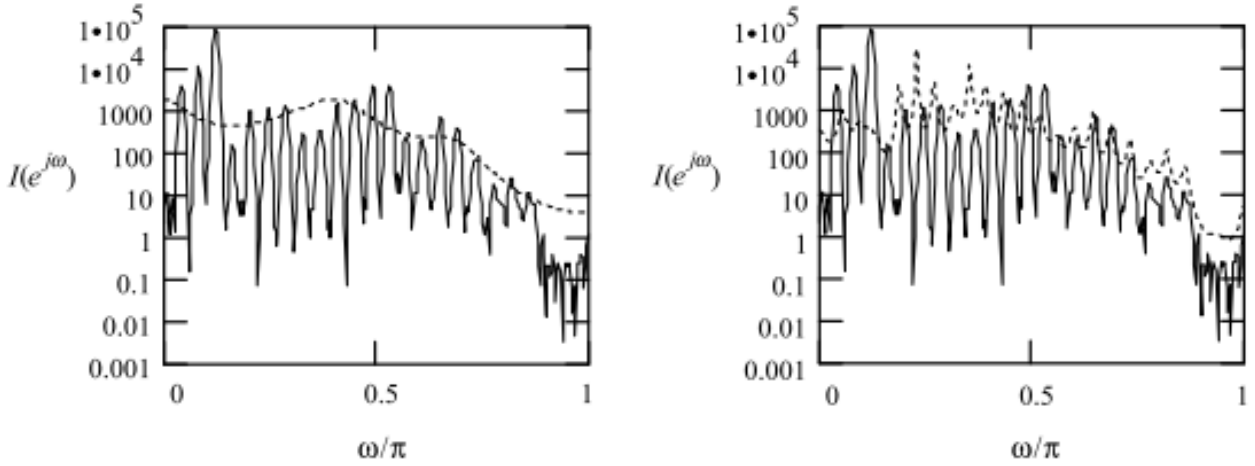
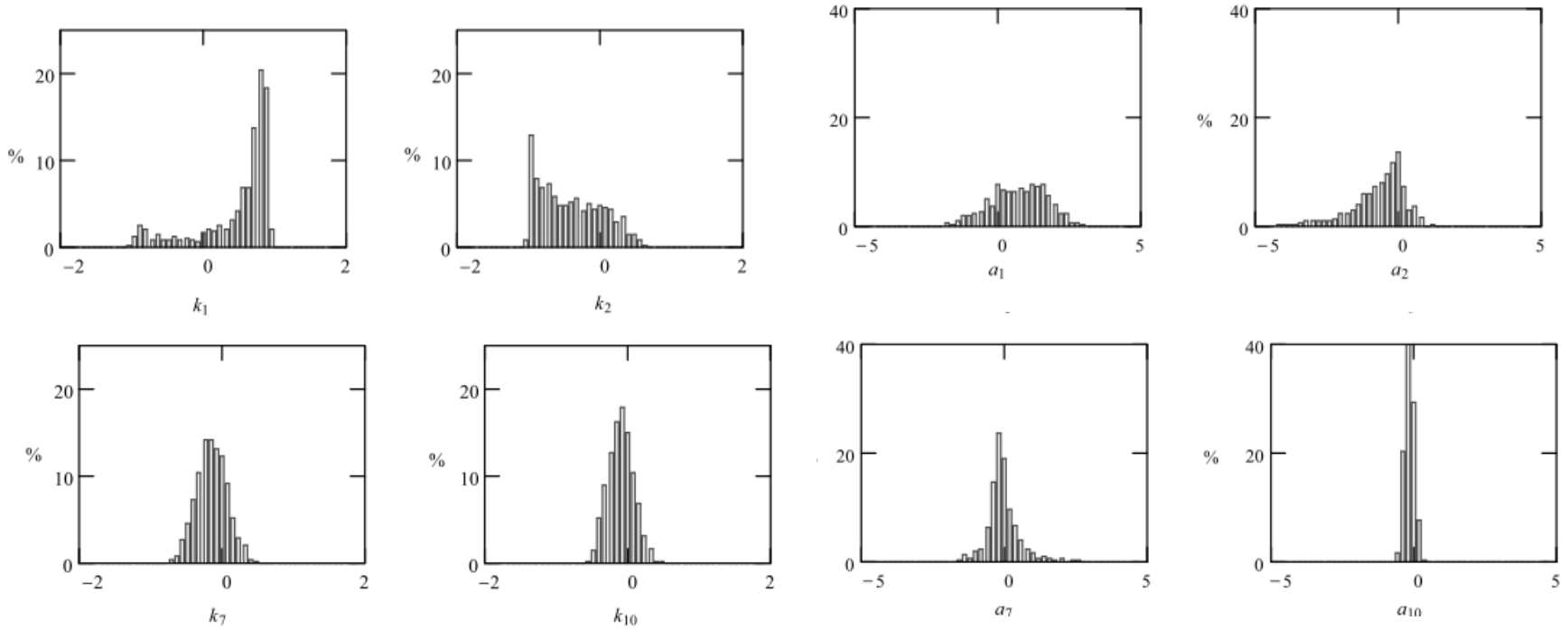


Figure 4.12 LPC-based spectrum estimate (dotted line) and periodogram (solid line) for a voiced speech frame. *Left*: $M = 10$. *Right*: $M = 50$.

Reflection coefficients

- There is a linear mapping between reflection coefficients and the linear prediction coefficients
- The effect of quantization of reflection coefficients is less than the quantization of the LPC coefficients



Long term linear prediction

- Prediction order should be > pitch period to accurately model voiced signals
- Problem: time varying + high bit rate (many LPCs)

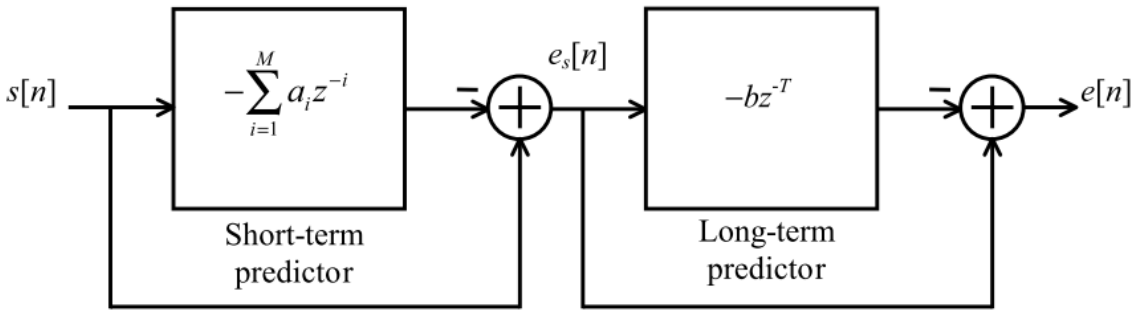


Figure 4.16 Short-term prediction-error filter connected in cascade to a long-term prediction-error filter.

Long Term Linear Prediction

LPCs. Experiments using an extensive amount of speech samples revealed that by shortening the time interval in which the long-term parameters were estimated from 20 to 5 ms, an increase in prediction gain of 2.2 dB was achievable [Ramachandran and Kabal, 1989].

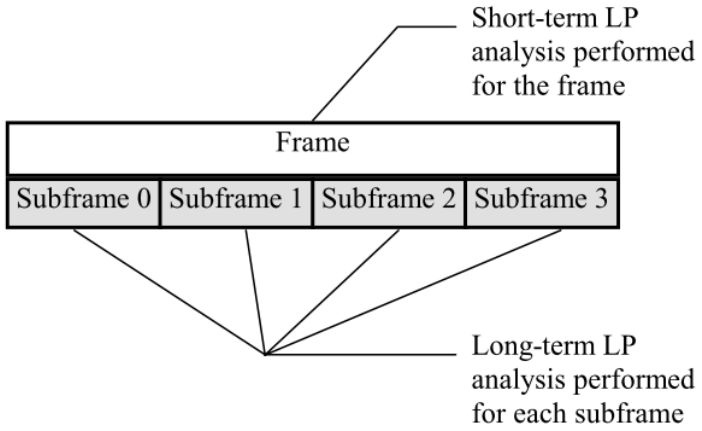


Figure 4.19 The frame/subframe structure.

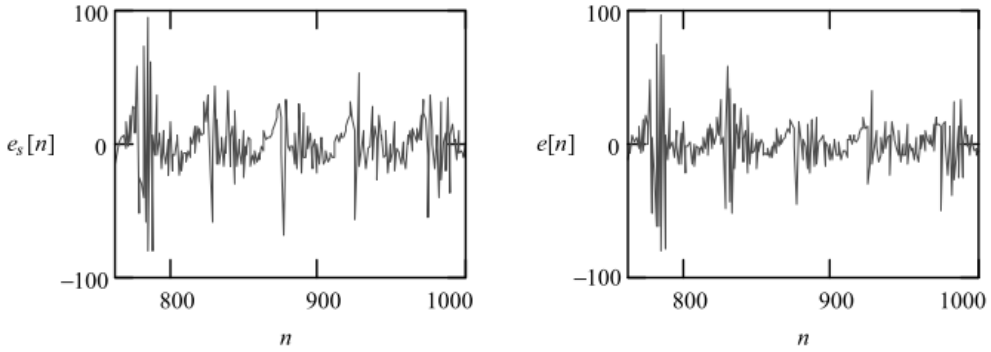


Figure 4.18 *Left*: Input to long-term prediction-error filter (short-term prediction error). *Right*: Output of the long-term prediction-error filter (overall prediction error).

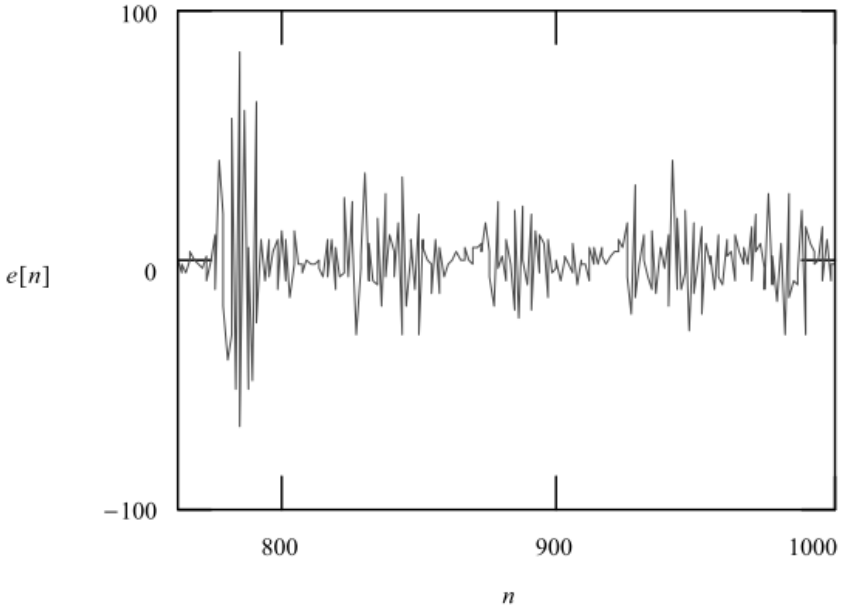


Figure 4.21 Example of long-term prediction-error filter's output.

Synthesis filters

$$H(z) = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}}$$

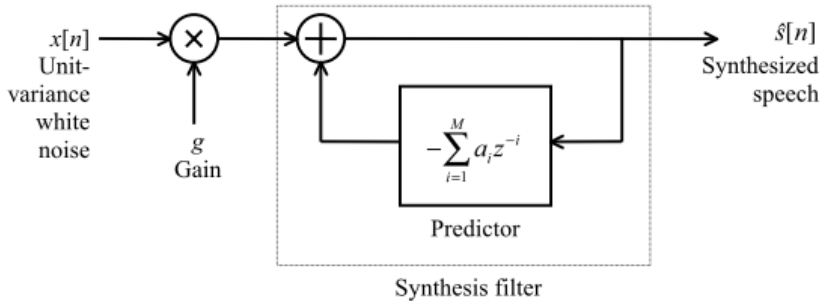


Figure 4.23 The synthesis filter.

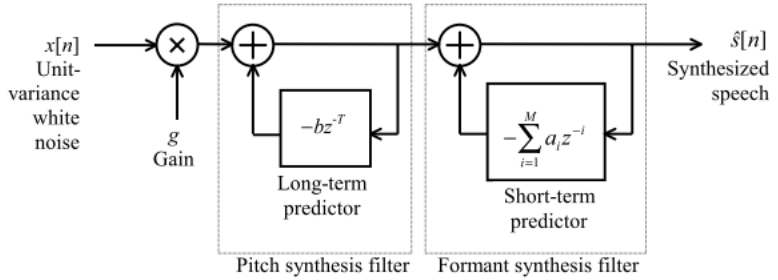


Figure 4.25 Long-term and short-term linear prediction model for speech production.

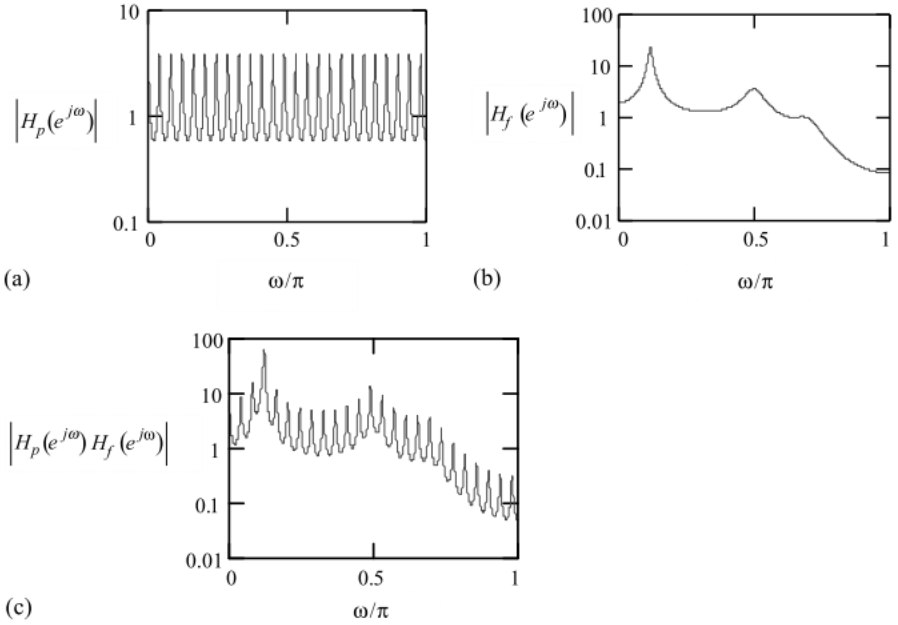


Figure 4.26 Magnitude plots of the transfer functions for (a) a pitch synthesis filter, (b) a formant synthesis filter, and (c) a cascade connection between pitch synthesis filter and formant synthesis filter.

Pre-emphasis of the speech waveform

- To compensate the roll off of the high frequencies in the spectrum

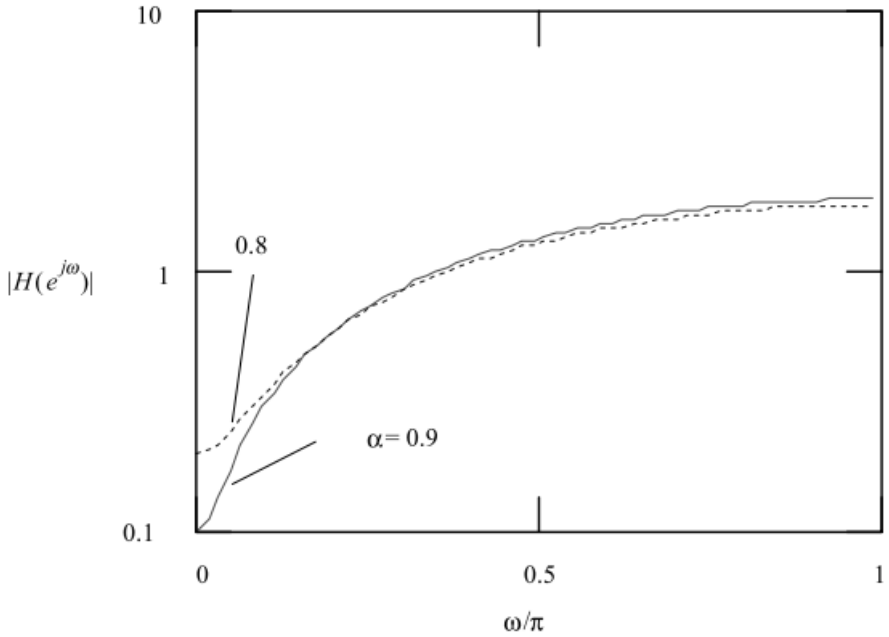


Figure 4.27 Magnitude plots of the transfer functions of the pre-emphasis filter.

Waveform CODECs

- G.711

$$D = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f_{\mathbf{x}}(x)\, dx,$$

$$\text{SNR} = 10 \log_{10}(\sigma^2 / D)$$

- Objective: minimize average distortion.
- You need to know the distribution of the input signal
- G.711 standard

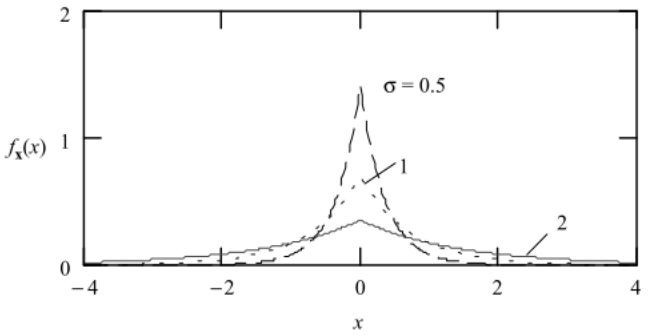


Figure 6.4 Plot of PDF for random variables with Laplacian distribution.

G.726

www.itu.int/rec/T-REC-G.726-199012-I/en

International Telecommunication Union

Home : ITU-T : Publications : Recommendations :
G Series : G.726 : G.726 (12/90)

Recently posted - Search Recommend

ITU Sectors | Newsroom | Events | Publications | Statistics | About ITU

G.726 : 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)

Recommendation G.726 (12/90)

Approved in 1990-12

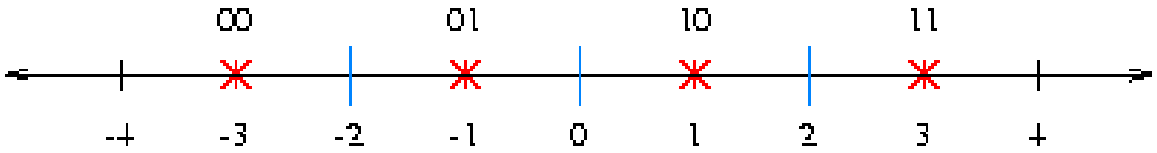
Status : In force

Corresponding ANSI-C code is available in the G.726 module of t
Software Tools Library

Access : Freely available items

Available languages and formats :

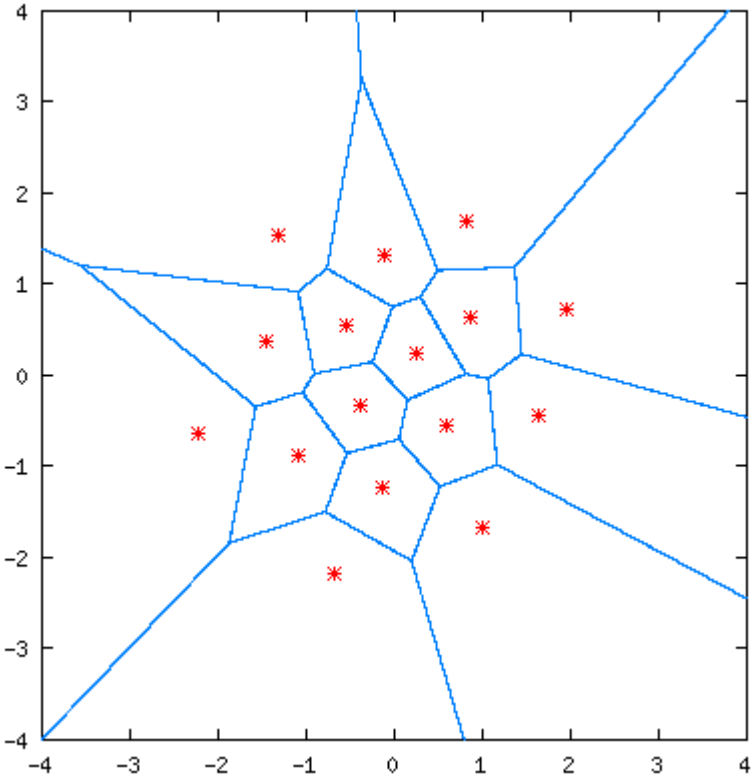
Vector quantization



- every pair of numbers falling

in a particular region are

approximated by a red star associated with that region



Definition 7.1: Vector Quantizer. A vector quantizer Q of dimension M and size N is a mapping from a vector $\mathbf{x}$ in M-dimensional Euclidean space $\boldsymbol{R}^M$ into a finite set $\boldsymbol{Y}$ containing N M-dimensional outputs or reproduction points, called codevectors or codewords. Thus,

$$Q: \boldsymbol{R}^M \rightarrow \boldsymbol{Y},$$

where

$$\mathbf{x} = [x_1, x_2, \ldots, x_M]^T,$$
$$(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N) \in \boldsymbol{Y},$$
$$\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iM}]^T; \quad i = 1, \ldots, N.$$

$\boldsymbol{Y}$ is known as the codebook of the quantizer. The mapping action is written as

$$Q(\mathbf{x}) = \mathbf{y}_i; \quad i = 1, \ldots, N. \tag{7.1}$$

$$\boldsymbol{R}_i = \{\mathbf{x} : d(\mathbf{x}, \mathbf{y}_i) \leq d(\mathbf{x}, \mathbf{y}_j);\ \forall\, j \in \boldsymbol{I}\}, \quad i \in \boldsymbol{I},$$

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{j=1}^{M} (x_j - \hat{x}_j)^2,$$

Linear Prediction Coding

- FS1015, 2.4kbps, 1982
- Originally for military applications. its synthetic output speech that often requires trained operators for reliable usage

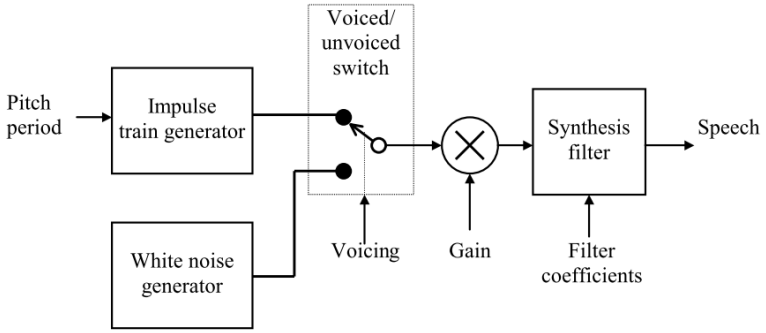
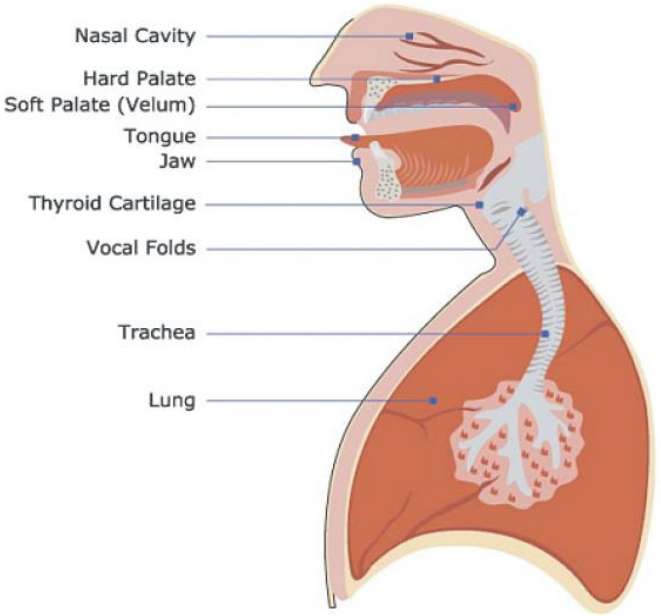


Figure 9.1 The LPC model of speech production.



- Each frame has parameters
- Encoder estimates paramters

Linear Prediction Coding

- Frame duration : 180 samples (22.5 ms)

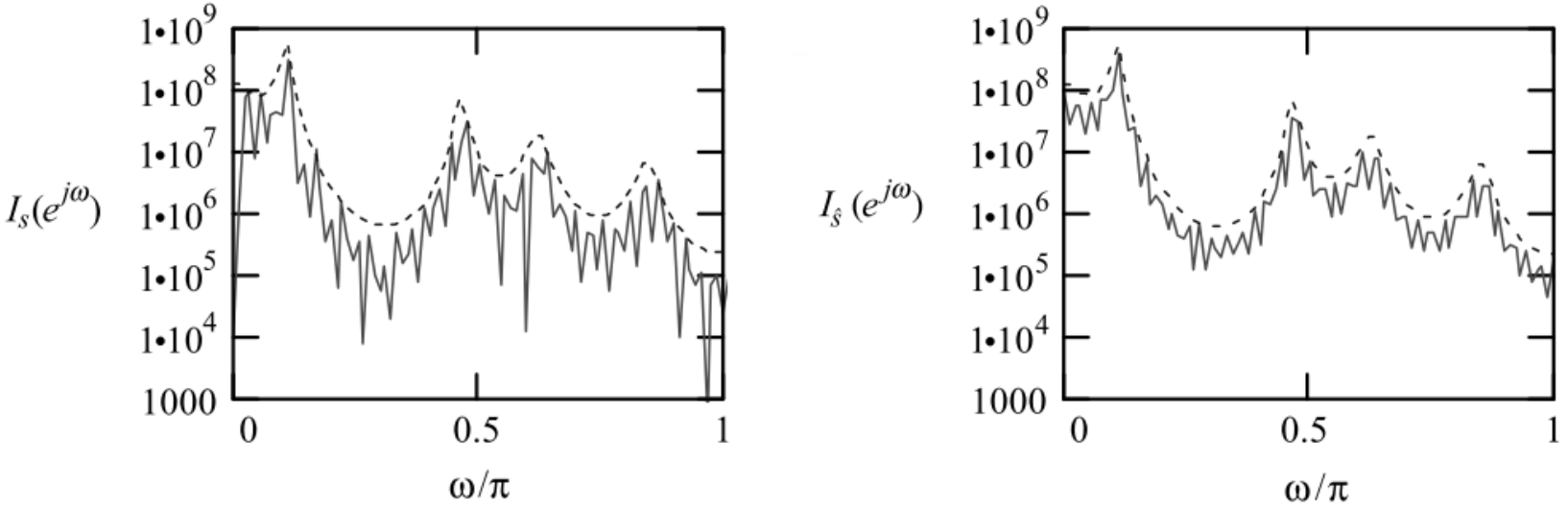


Figure 9.5 Plots of periodograms for a voiced frame. *Left*: Original; *right*: synthetic. The PSD using the estimated LPC is superimposed (dotted line).

FS1015 (LPC10)

- Input: 8kHz speech, PCM, 12 bits/sample
- Frame size: 180 samples = 22.5 ms
- Possible pitch periods = only 60 values
- 54 bits per frame. Hence bit rate = 54*8000/180 = 2400

TABLE 9.1 Bit Allocation for the FS1015 Coder[a]

	Resolution	
Parameter	Voiced	Unvoiced
Pitch period / voicing	7	7
Power	5	5
LPC	41	20
Synchronization	1	1
Error protection	—	21
Total	**54**	**54**

[a] Data from Tremain [1982], Figure 8.

Advantages and disadvantages

- Advantages:
  - Low bit rate
  - Very simple encoder and decoder
- Disadvantages:
  - Sometimes the speech frame cannot be classified as strictly voiced or unvoiced
  - The use of noise or impulse train is not a good modelling
  - Bad quality

Samples:

REGULAR-PULSE EXCITATION CODERS

- Multipulse excitation
  - Open loop

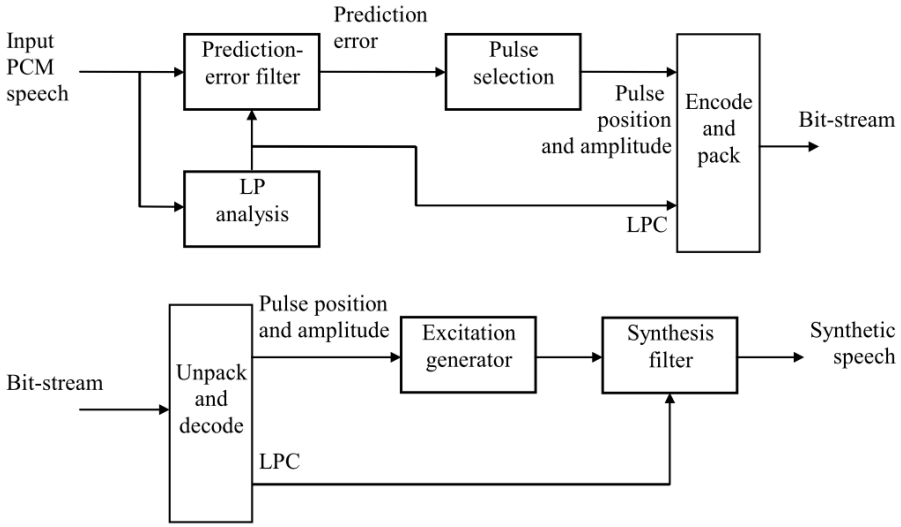
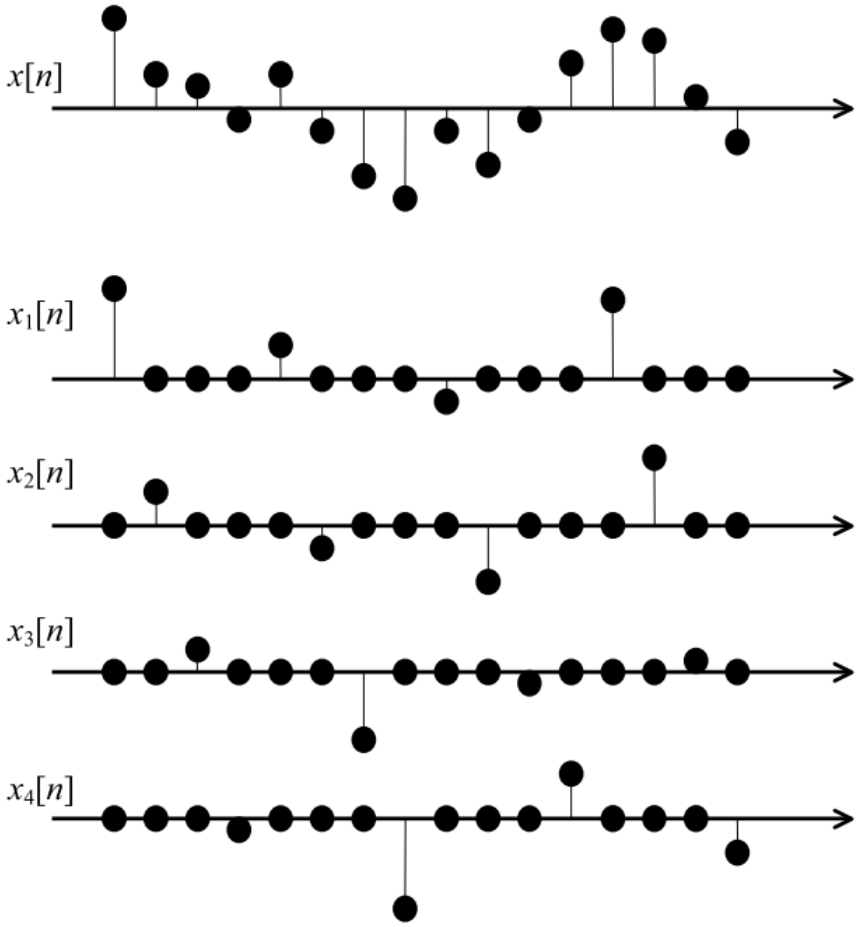


Figure 10.1 Encoder (*top*) and decoder (*bottom*) of an open-loop multipulse coder.

  - Use a certain criteria to select only few pulses of the prediction error

- Regular pulse excitation



- Closed loop (Analysis by Synthesis)

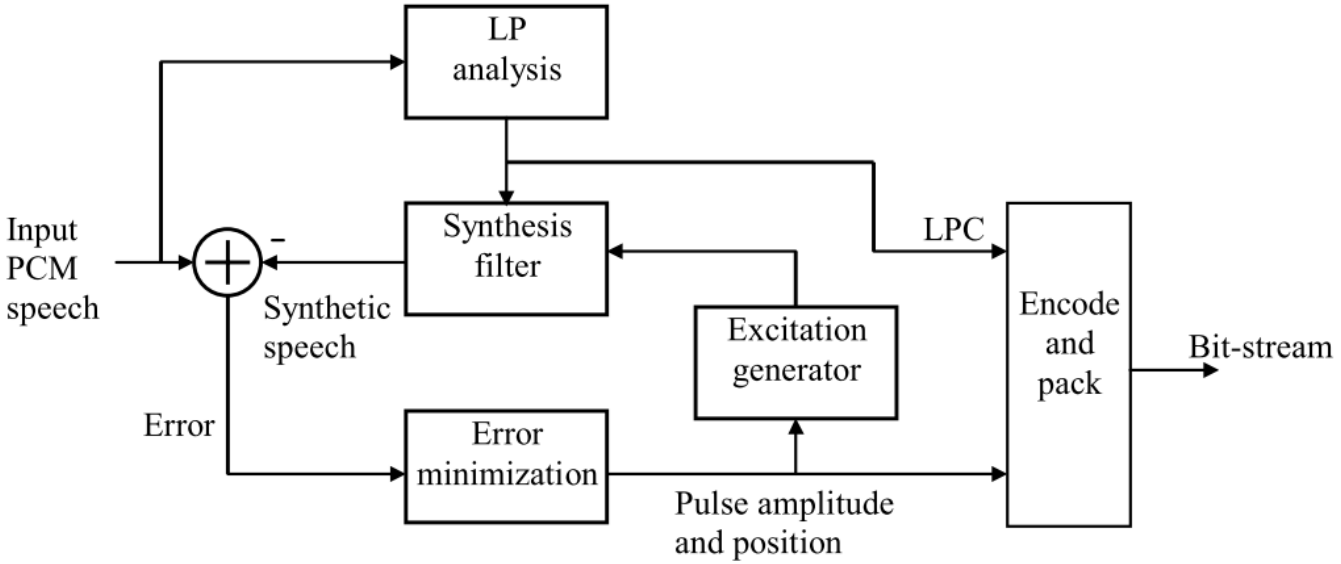


Figure 10.3 Encoder of a closed-loop multipulse coder.

GSM 6.10 (1988)

- Regular pulse excited Long Term prediction (RPE-LTP)
  - Low computational cost
  - High quality reproduction
  - Robustness against channel errors
  - Coding efficiency
- 8 reflection coefficients
- One LPC vector every 160 samples (20ms)
- Selects one of 4 subsampled error sequences at each subframe (40 samples)

GSM 6.10 (1988)

TABLE 10.1 Bit Allocation for the GSM 6.10 RPE-LTP Coder[a]

Parameter	Number per Frame	Resolution	Total Bits per Frame
LPC	8	6, 6, 5, 5, 4, 4, 3, 3	36
Pitch period	4	7	28
Long-term gain	4	2	8
Position	4	2	8
Peak magnitude	4	6	24
Sample amplitude	4·13	3	156
Total			**260**

[a]Data from ETSI [1992a], Table 1.1a.

Code Excited Linear Prediction (CELP)

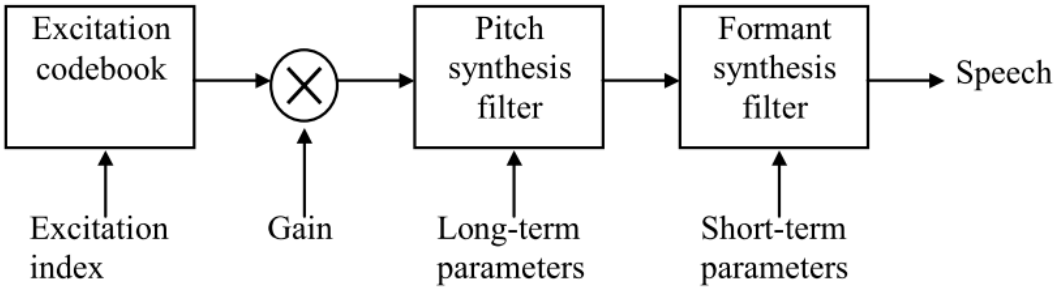


Figure 11.1 The CELP model of speech production.

- Excitation codebook can be fixed/adaptive , deterministic/random
- No strict (Voiced/unvoiced) classification

CELP

- Analysis by synthesis

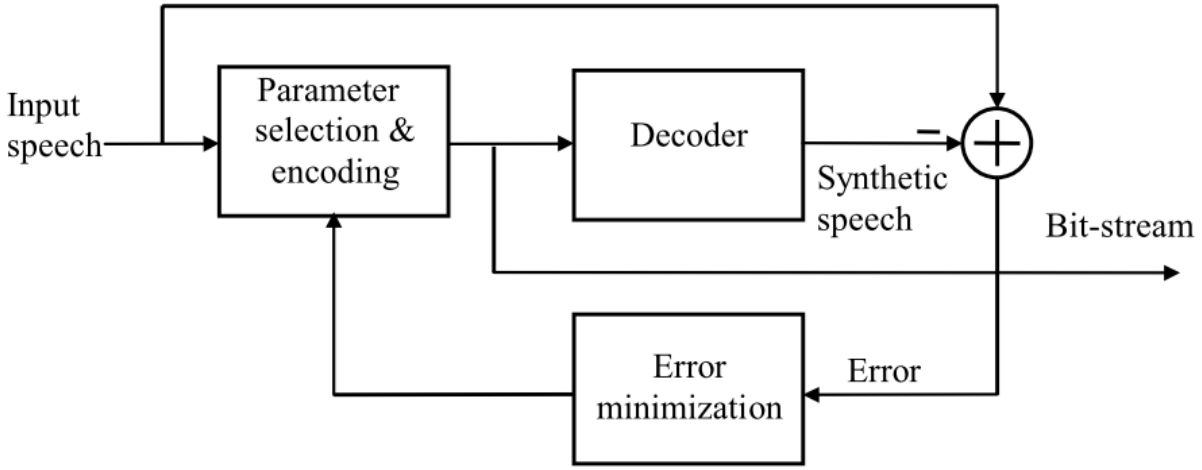


Figure 11.2 Block diagram showing an encoder based on the analysis-by-synthesis principle.

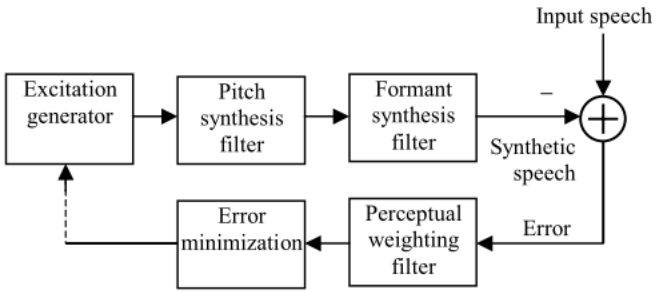


Figure 11.5 Analysis-by-synthesis loop of a CELP encoder with perceptual weighting.

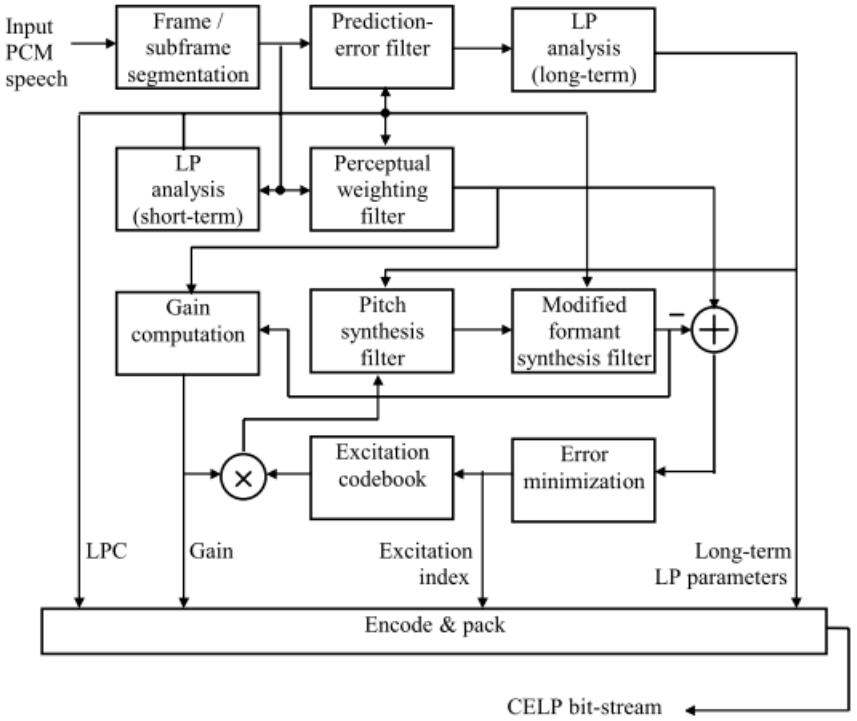


Figure 11.9 Block diagram of a generic CELP encoder.

CELP

- Advantages?
- Disadvantages?

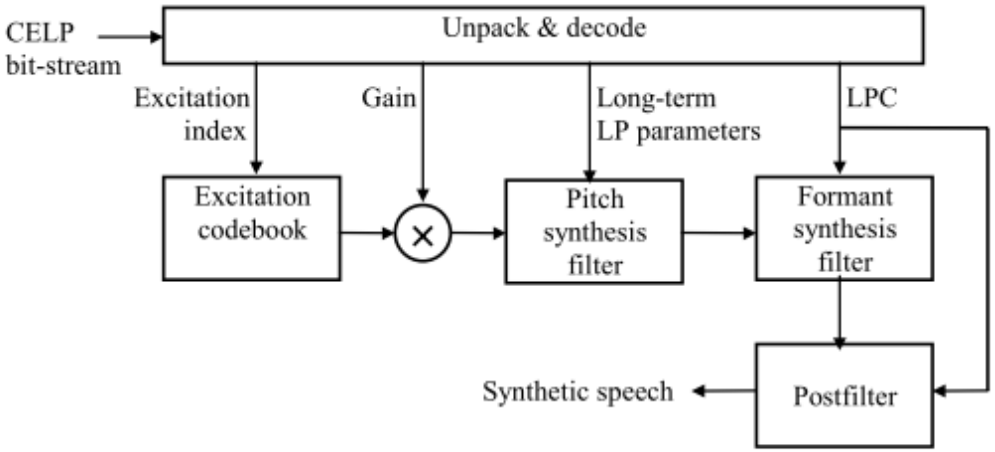


Figure 11.10 Block diagram of a generic CELP decoder.

G.728 (LD-CELP)

- 20 samples frames – Four 5 samples subframes
- Pitch period: first coarse estimate, then a fine estimate
  - Compared to previous pitch to check for halving or doubling
- Pitch: once per frame (obtained in decimated domain by a factor of 4, then normal domain)
- Bit rate: 16 kbps

Vector Sum Excited Linear Prediction (VSELP)

- A CELP coder with a particular codebook structure having reduced computational cost.
- IS54 (7.96kbps) , GSM 6.20 (5.6kbps) “Half Rate"
- Basic idea:
  - Form the codebook from some basis functions

$$\mathbf{v}1^{(I1)} = \sum_{i=0}^{6} \theta_i^{(I1)} \mathbf{a}1_i$$

- G.729 uses CELP

GSM EFR ACELP

- A-CELP based
- 12.2kbps bit rate + 10.6kbps channel coding = 22.8kbps

TABLE 16.6 Bit Allocation for the GSM EFR Coder[a]

Parameter	Number per Frame	Resolution	Total Bits per Frame
LPC index	1	38	38
Pitch period	4	9, 6, 9, 6	30
Adaptive codebook gain	4	4	16
Algebraic codebook index	4	35	140
Algebraic codebook gain	4	5	20
Total			**244**

[a] Data from Salami et al. [1997a], Table 1.

- ETSI AMR (Adaptive Multirate)
- All coders based no ACELP
- 12.2 (EFR), 10.2, 7.95, 7.40, 6.70, 5.90, 5.15, and 4.75 kbps.

MELP (Mixed Excited Linear Prediction)

- 2.4 kbps

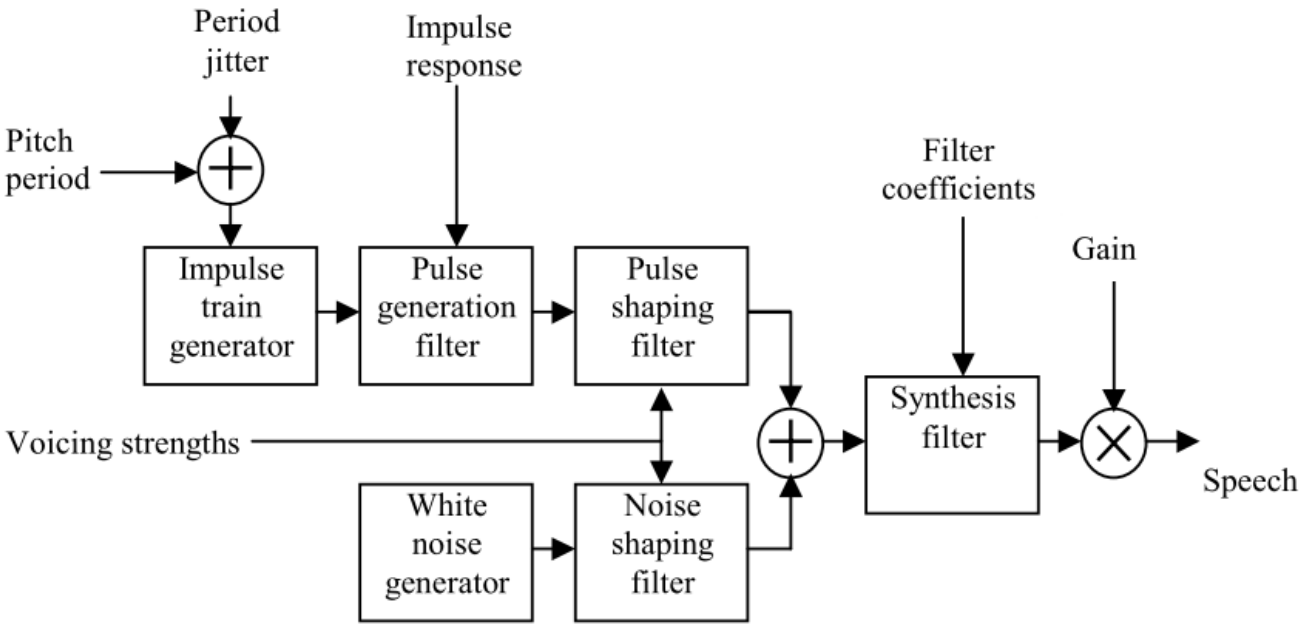


Figure 17.1 The MELP model of speech production.

Fourier Magnitudes

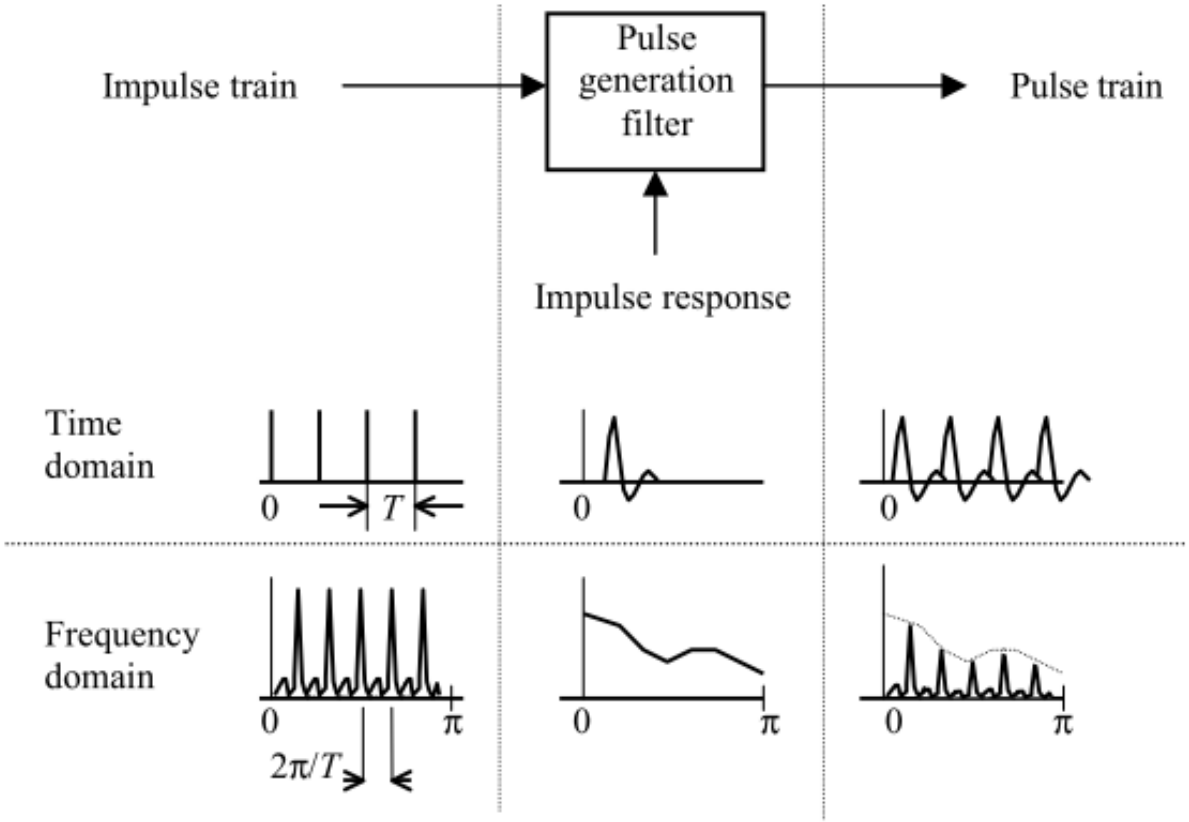


Figure 17.2 Illustration of signals associated with the pulse generation filter.

Shaping filters

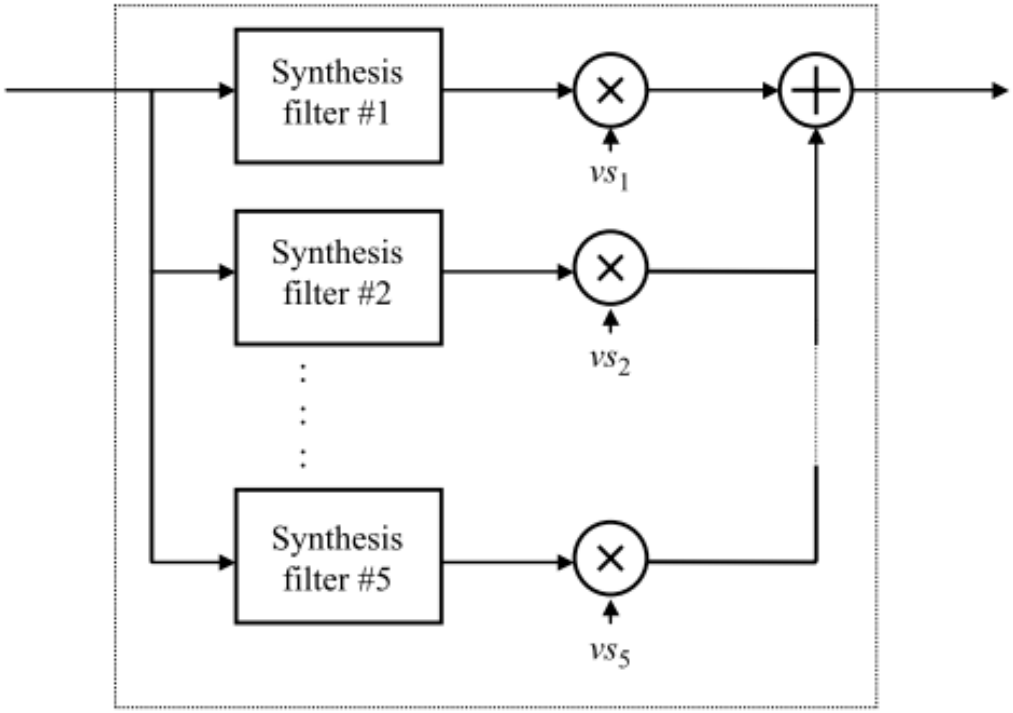


Figure 17.10 Block diagram of the pulse shaping filter.

MELP bit allocation

Parameter	Voiced	Unvoiced
LPC	25	25
Pitch period/low-band voicing strength	7	7
Bandpass voicing strength	4	—
First gain	3	3
Second gain	5	5
Aperiodic flag	1	—
Fourier magnitudes	8	—
Synchronization	1	1
Error protection	—	13
Total	**54**	**54**

[a] Data from McCree et al. [1997], Table 1.