

CECS401

Fundamentals of Spoken Language Processing

Note-9

Tuesday 9/28/99

E. Application of LPC in Speech Coding

Multipulse LPC vocoder

Analysis-by-synthesis method of determining excitation:

- LPC analyzer computes LPC parameters from speech samples s_n .
- LPC synthesizer produces synthetic speech samples \hat{s}_n in response to excitation v_n .
- The synthetic speech is compared with the original speech to produce an error sequence $e_n = s_n - \hat{s}_n$.
- Perceptually-weighted mean-squared error ϵ is computed for every 5 msec.
- The locations and amplitudes of the excitation pulses v_n are chosen to minimize the error ϵ .

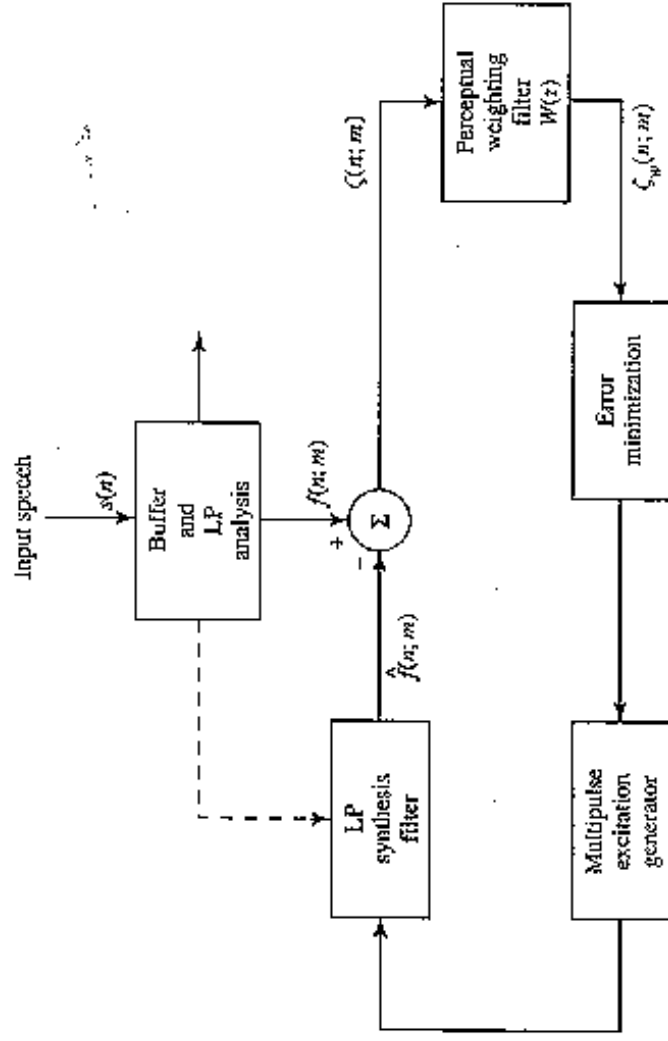


FIGURE 7.38. Analysis-by-synthesis method for obtaining the multipulse excitation.

Perceptual error weighting:

Perceptual error weighting is based on auditory masking, a phenomenon linking to hearing perception of neighboring signal components.

Masking

If a strong signal A (the masker) masks a weaker signal B (the maskee), then B is not heard, even though it is present:

- *simultaneous masking*: A and B occur at the same time.
- *temporal masking*: B either precedes or succeeds A.

Critical band

Human ear has nonlinear spectral resolution, finer at low frequencies and coarser at high frequencies.

Frequency components of sounds are integrated into critical bands.

Critical band rate z (CBR measure in unit of Bark):

CBR represents the center frequency locations of these subbands and is measured in Bark.

- Below 1000 Hz, CBR is approximately linear with frequency (Hz).
- Beyond 1000 Hz, CBR is approximately linear with log frequency (Hz).

Critical bandwidth (CBW):

CBW represents the bandwidths of these subbands.

- If the center frequency is below 500 Hz, CBW is about 100 Hz
- If the center frequency is beyond 500 Hz, CBW is about 20% of the center frequency value.

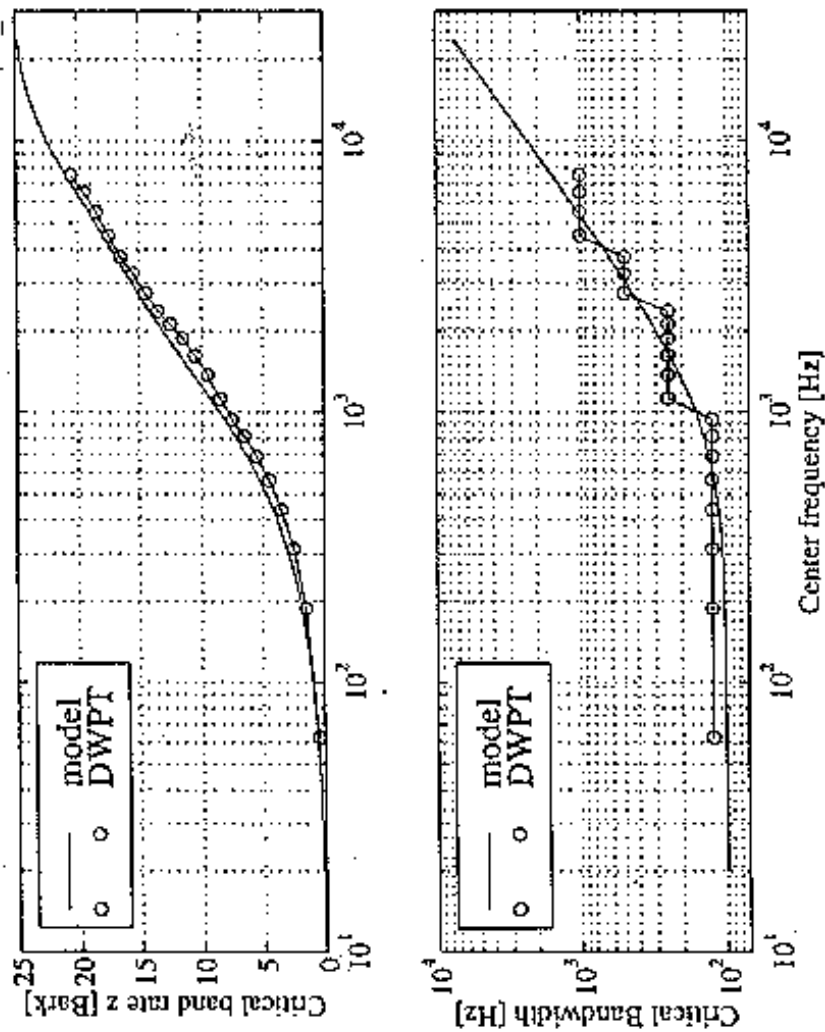


Fig. 5. Critical band rate (top) and critical bandwidth (bottom) as a function of center frequency.

Tonality of masker:

Defined by the spectral content of the masker within one critical band.

- A tonal masker consists of one or few tones (pure sinusoids)
- A narrowband noise-like masker consists of many neighboring tones with random phases.

Masking threshold:

A maskee is not heard if it is below the masking threshold. The threshold depends on the sound pressure level, the frequency of the masker, and the tonalities of the masker and maskee.

Noise is a better masker than tones.

In speech coding, the maskee is the quantization noise, which is generally assumed as uniform white noise.

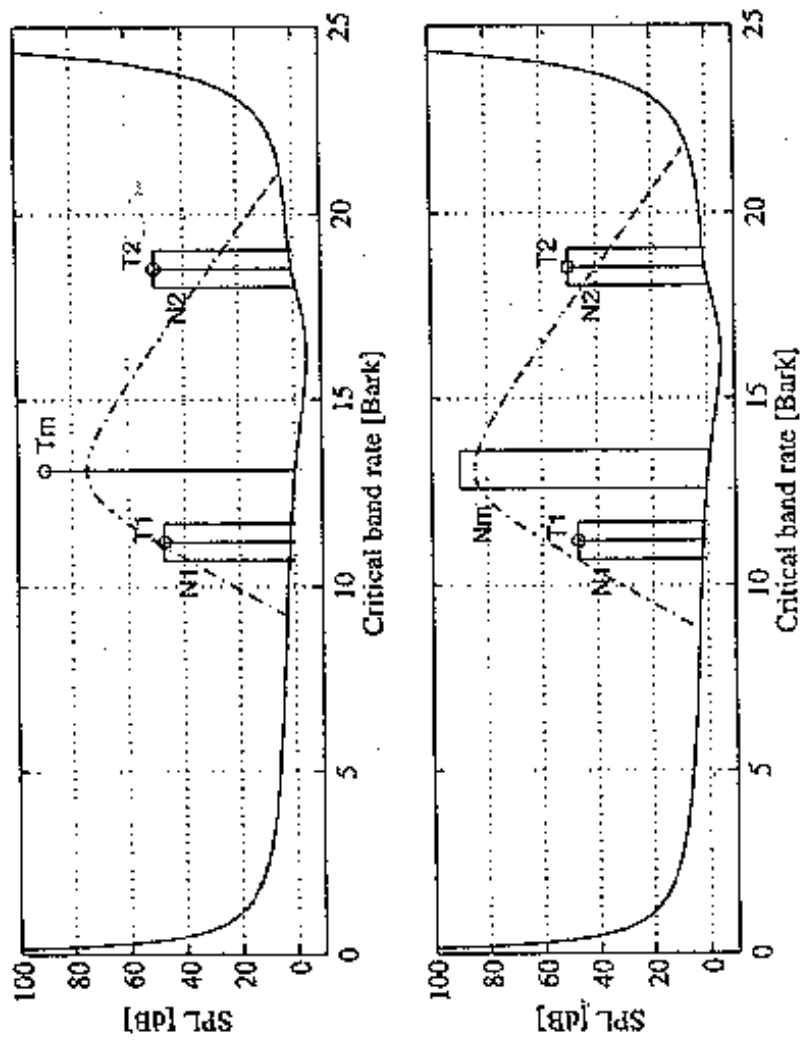


Fig. 7. Simultaneous masking. Top: Tone masker. Bottom: Noise-like masker (one Bark wide).

Perceptual weighting criterion:

Deemphasis error spectrum in the formant regions since strong energy of speech formants can mask quantization noise.

Frequency-weighted error:

$$\epsilon = \int_0^{f_B} |S(f) - \hat{S}(f)|^2 W(f) df$$

f_B is the bandwidth of the signal.

$$W(z) = \frac{\hat{A}(z)}{\hat{A}(z/\gamma)}$$

$\hat{A}(z)$ is the inverse filter computed from estimated LPC parameters, $0 \leq \gamma \leq 1$ adjusts the error weighting in formant regions.

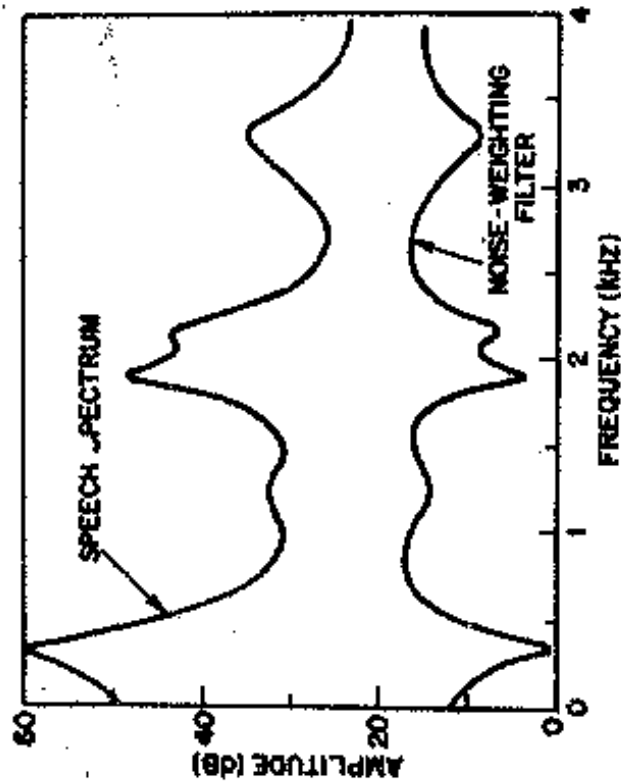


Fig. 5. An example of the speech spectrum and the frequency response of the corresponding error-weighting filter.

Error minimization procedure:

For a given discrete time interval $[n', n'']$, the location and amplitude of excitation pulses are determined one at a time.

Consider placing the first pulse of amplitude $\alpha_1(k_1)$ at location $n = k_1$.

$$\hat{S}(z) = \frac{1}{\hat{A}(z)} \alpha_1(k_1) z^{-k_1}$$

$$\hat{S}(z)W(z) = \frac{1}{\hat{A}(z/\gamma)} \alpha_1(k_1) z^{-k_1} \triangleq \Theta_W(z) \alpha_1(k_1) z^{-k_1}$$

$$Z^{-1}(\hat{S}(z)W(z)) = \alpha_1(k_1) \theta_W(n - k_1)$$

$$Z^{-1}(S(z)W(z)) = \zeta_W^0(n)$$

$$\zeta_W^1(n) = \zeta_W^0(n) - \alpha_1(k_1) \theta_W(n - k_1)$$

$$\epsilon(k_1) = \sum_{n=n'}^{n''} (\zeta_W^1(n))^2 = \sum_{n=n'}^{n''} (\zeta_W^0(n) - \alpha_1(k_1) \theta_W(n - k_1))^2$$

Minimization of $\epsilon(k_1)$ w.r.t. $\alpha_1(k_1)$ yields:

$$\hat{\alpha}_1(k_1) = \frac{\sum_{n=n'}^{n''} \zeta_W^0(n) \theta_W(n - k_1)}{\sum_{n=n'}^{n''} \theta_W^2(n - k_1)} \triangleq \frac{\rho_{\zeta\theta}(k_1)}{\rho_{\theta\theta}(k_1)}$$

$$\min_{\alpha_1(k_1)} \epsilon(k_1) = \hat{\epsilon}(k_1) = \sum_{n=n'}^{n''} (\zeta_W^0(n))^2 - \frac{\rho_{\zeta\theta}^2(k_1)}{\rho_{\theta\theta}(k_1)}$$

$$\hat{k}_1 = \arg \min_{n' \leq k_1 \leq n''} \hat{\epsilon}(k_1)$$

To place a second pulse of amplitude $\alpha_2(k_2)$ at location $n = k_2$, minimizes

$$\epsilon(k_2) = \sum_{n=n'}^{n''} (\zeta_W^2(n))^2 = \sum_{n=n'}^{n''} (\zeta_W^1(n) - \alpha_2(k_2) \theta_W(n - k_2))^2$$

Repeat the procedure until

- the perceptual weighted error is reduced below a specified threshold, or
- the number of pulses reaches the limit according to a specified bit rate.

Example

Speech duration = 0.1 sec., LPC order = 16, window size = 20 msec, shift = 10 ms. The excitation pulse locations and amplitudes were determined by minimizing the errors over successive 5 ms intervals.

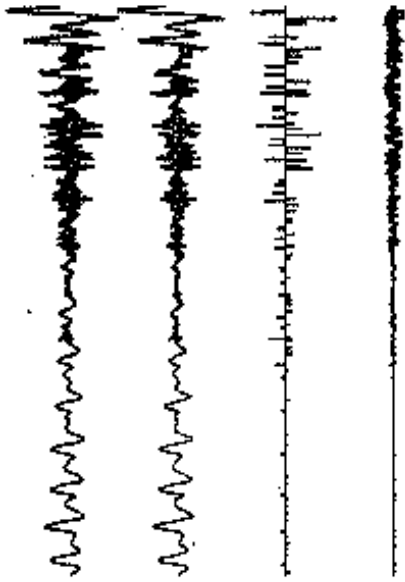


Fig. 7. Waveforms of the original speech, the synthetic speech, the excitation, and the error signals.

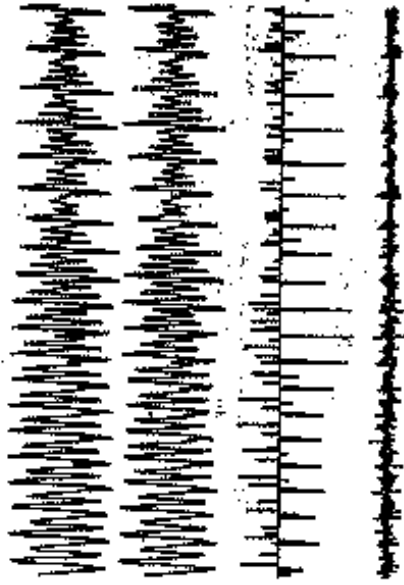


Fig. 8. Another example of the waveforms of the original speech, the synthetic speech, the excitation signal, and the error.

Incorporating pitch predictor into multipulse LPC vocoder:
For voiced speech, the multipulse LPC excitation sequence shows a significant correlation from one pitch period to the next. The perceptual error can be further reduced by including a pitch predictor:

$$v_n = b_1 v_{n-\hat{p}} + \Theta_p v_n$$

with the system function

$$\Theta_p(z) = \frac{\Theta_p}{1 - bz^{-\hat{p}}}$$

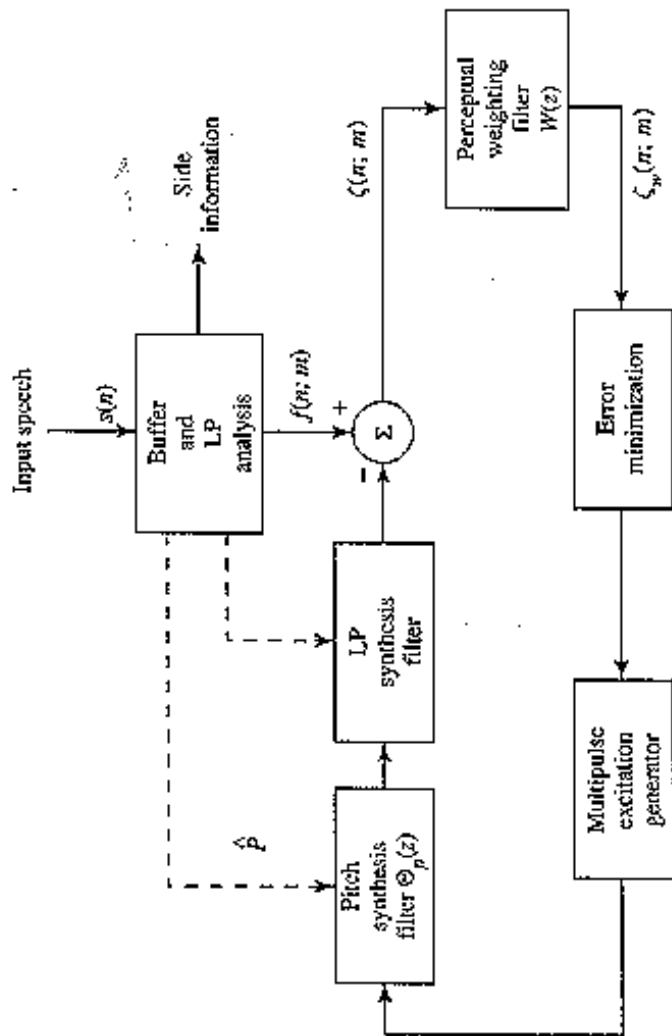


FIGURE 7.39. Analysis-by-synthesis method for obtaining the multipulse excitation with a pitch synthesis filter.

Multipulse LPC has been implemented for commercial use. The transmitted information includes

Excitation parameters updated every 5 ms:

- the location of pulses (differential coding)
- log amplitude of the largest pulse (6 bits)
- the pulse amplitudes relative to the overall largest amplitude (4 bits each pulse)

LPC vocal tract parameters and pitch updated every 20 ms.

Good-quality speech has been achieved at 9600 bps. The codec has been used for airborne mobile satellite telephone services.

Code-excited LPC vocoder (Schroeder & Atal, 1985)

CELP is an analysis-by-synthesis coding method that selects the excitation sequence from a codebook of zero-mean Gaussian noise sequences.

Excitation model:

The excitation code words are generated by

$$v_n = \sum_{k=0}^{N-1} c_k \cos \left(\frac{\pi k n}{N} + \phi_k \right)$$

where ϕ_k is uniformly distributed over $[0, 2\pi]$, c_k is Rayleigh distributed, or

- vector quantization from residues of speech.

Example

For 8 KHz sampled speech

- codewords are 40 samples long (5 ms)
- codebook of 1024 codewords is sufficient (10 bits per codeword)
- bit rate of excitation signal is 2,000 bps, i.e., 0.25 bit per sample.

Determining excitation codeword:

- compute LPC parameters for each analysis frame
- a codeword in the excitation codebook is selected and scaled to excite the synthesis system consisting of a cascade of pitch predictor and LPC predictor
- perceptually weighted errors between the original speech and synthesized speech is computed for each excitation subframe
- by exhaustively searching through the excitation codebook, the scale or gain and the excitation codeword are determined to minimize the error.

CELP achieves good speech quality at 4,800 bps.

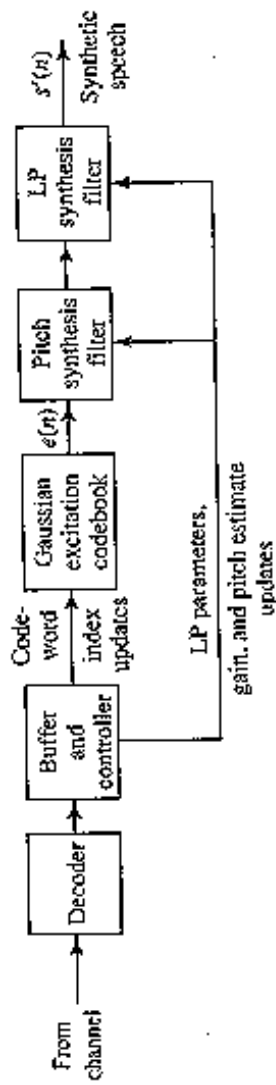


FIGURE 7.40. CELP synthesizer.

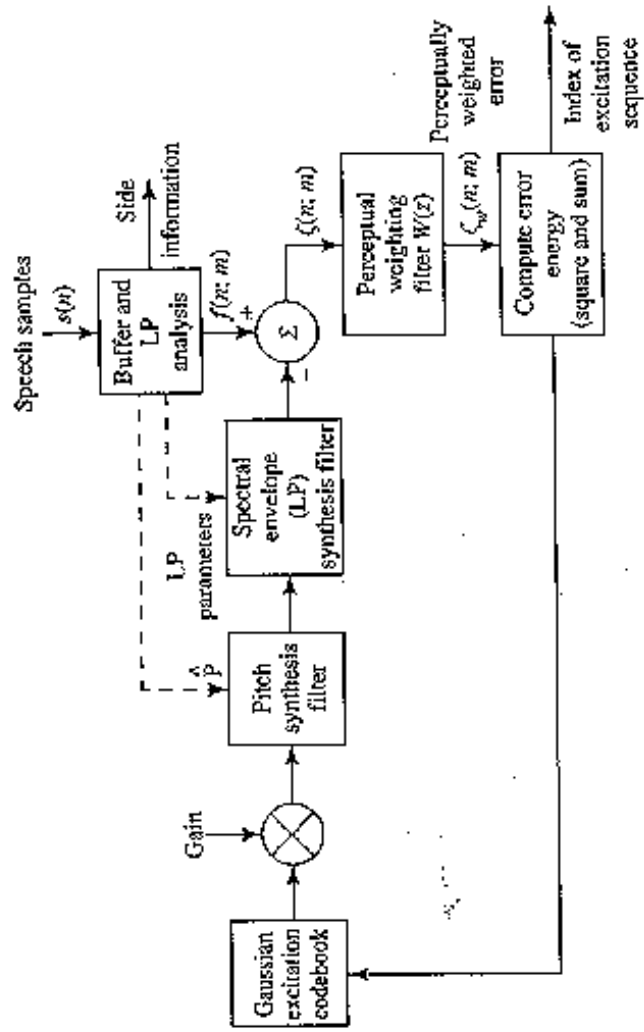


FIGURE 7.41. CELP analysis-by-synthesis coder.