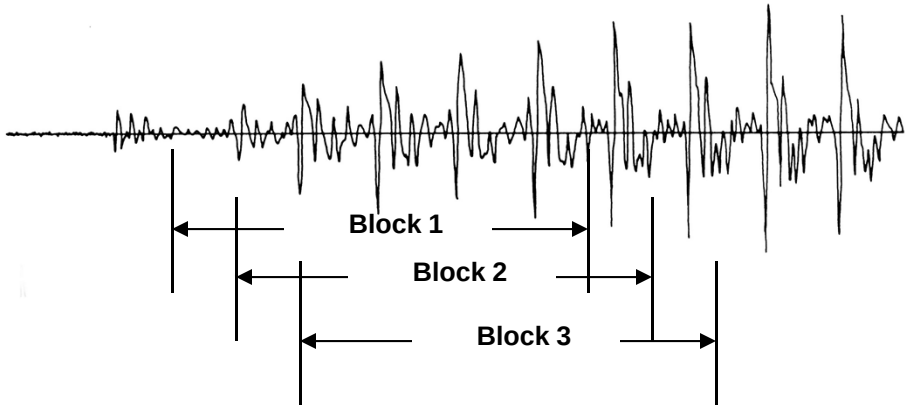
Digital Speech Processing—Lecture 17

Speech Coding Methods Based on Speech Models

Waveform Coding versus Block Processing

- Waveform coding
  - sample-by-sample matching of waveforms
  - coding quality measured using *SNR*
- Source modeling (block processing)
  - block processing of signal => vector of outputs every block
  - overlapped blocks



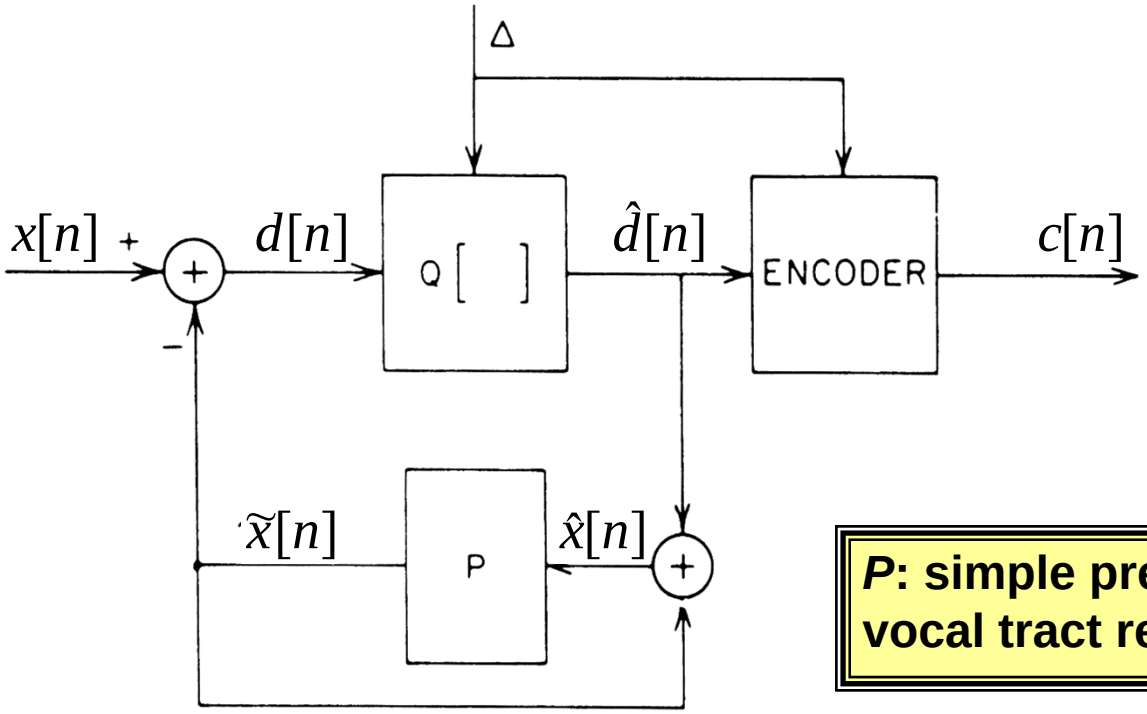
Model-Based Speech Coding

- we've carried waveform coding based on optimizing and maximizing *SNR* about as far as possible
  - achieved bit rate reductions on the order of 4:1 (i.e., from 128 Kbps PCM to 32 Kbps ADPCM) at the same time achieving toll quality *SNR* for telephone-bandwidth speech
- to lower bit rate further without reducing speech quality, we need to exploit features of the speech production model, including:
  - source modeling
  - spectrum modeling
  - use of codebook methods for coding efficiency
- we also need a new way of comparing performance of different waveform and model-based coding methods
  - an objective measure, like *SNR*, isn't an appropriate measure for model-based coders since they operate on blocks of speech and don't follow the waveform on a sample-by-sample basis
  - new subjective measures need to be used that measure user-perceived quality, intelligibility, and robustness to multiple factors

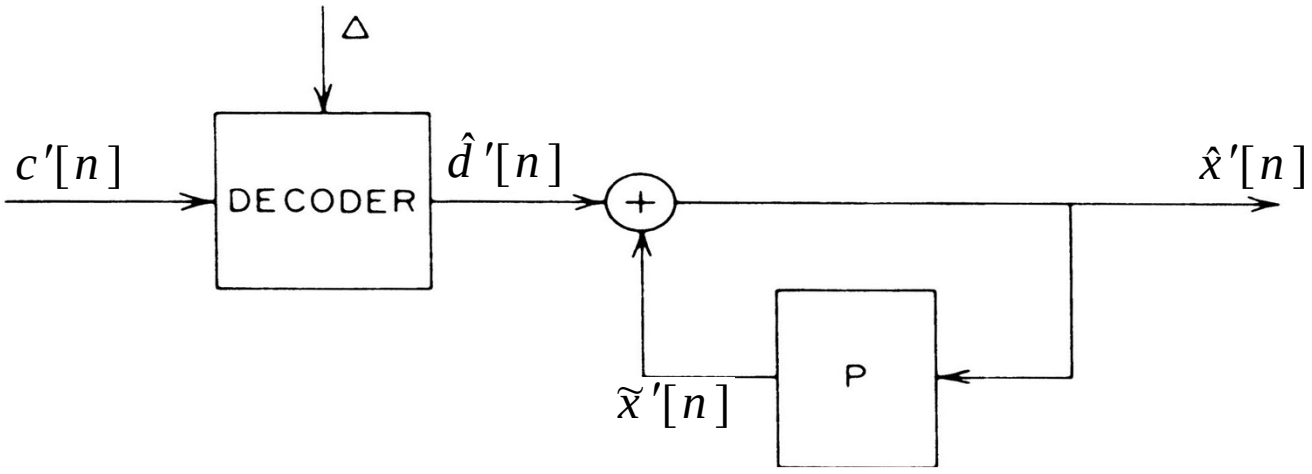
Topics Covered in this Lecture

- Enhancements for ADPCM Coders
  - pitch prediction
  - noise shaping
- Analysis-by-Synthesis Speech Coders
  - multipulse linear prediction coder (MPLPC)
  - code-excited linear prediction (CELP)
- Open-Loop Speech Coders
  - two-state excitation model
  - LPC vocoder
  - residual-excited linear predictive coder
  - mixed excitation systems
- speech coding quality measures - MOS
- speech coding standards

Differential Quantization



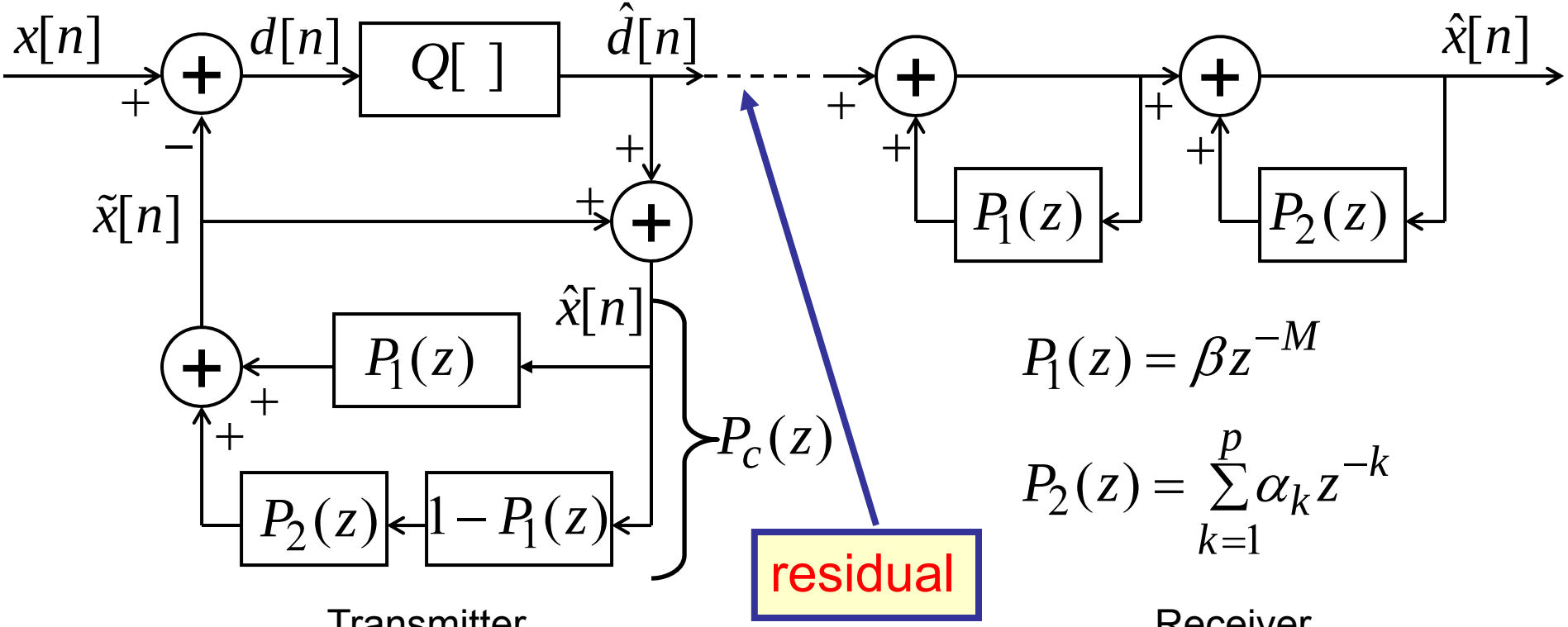
***P*: simple predictor of vocal tract response**



Issues with Differential Quantization

- difference signal retains the character of the excitation signal
  - switches back and forth between quasi-periodic and noise-like signals
- prediction duration (even when using p=20) is order of 2.5 msec (for sampling rate of 8 kHz)
  - predictor is predicting vocal tract response – not the excitation period (for voiced sounds)
- Solution – incorporate two stages of prediction, namely a short-time predictor for the vocal tract response and a long-time predictor for pitch period

Pitch Prediction



$$P_1(z) = \beta z^{-M}$$

$$P_2(z) = \sum_{k=1}^{p} \alpha_k z^{-k}$$

- first stage pitch predictor:

$$P_1(z) = \beta \cdot z^{-M}$$

- second stage linear predictor (vocal tract predictor):

$$P_2(z) = \sum_{k=1}^{p} \alpha_k z^{-k}$$

Pitch Prediction

- first stage pitch predictor:

$$P_1(z) = \beta \cdot z^{-M}$$

- this predictor model assumes that the pitch period, *M*, is an integer number of samples and β is a gain constant allowing for variations in pitch period over time (for unvoiced or background frames, values of *M* and β are irrelevant)
- an alternative (somewhat more complicated) pitch predictor is of the form:

$$P_1(z) = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1} = \sum_{k=-1}^{1} \beta_k z^{-M-k}$$

- this more advanced form provides a way to handle non-integer pitch period through interpolation around the nearest integer pitch period value, *M*

Combined Prediction

- The combined inverse system is the cascade in the decoder system:

$$H_c(z) = \left(\frac{1}{1-P_1(z)}\right)\left(\frac{1}{1-P_2(z)}\right) = \left(\frac{1}{1-P_c(z)}\right)$$

- with 2-stage prediction error filter of the form:

$$1-P_c(z) = \left[1-P_1(z)\right]\left[1-P_2(z)\right] = 1-\left[1-P_1(z)\right]P_2(z)-P_1(z)$$

$$P_c(z) = \left[1-P_1(z)\right]P_2(z)-P_1(z)$$

- which is implemented as a parallel combination of two predictors:

$\left[1-P_1(z)\right]P_2(z)$ and $P_1(z)$

- The prediction signal, $\tilde{x}[n]$ can be expressed as

$$\tilde{x}[n] = \beta\hat{x}[n-M] + \sum_{k=1}^{p}\alpha_k(\hat{x}[n-k]-\beta\hat{x}[n-k-M])$$

Combined Prediction Error

- The combined prediction error can be defined as:

$$d_c[n] = x[n] - \tilde{x}[n]$$

$$= v[n] - \sum_{k=1}^{p} \alpha_k v[n-k]$$

- where

$$v[n] = x[n] - \beta x[n-M]$$

- is the prediction error of the pitch predictor.
- The optimal values of β, M and $\{\alpha_k\}$ are obtained, in theory, by minimizing the variance of $d_c[n]$. In practice a sub-optimum solution is obtained by first minimizing the variance of $v[n]$ and then minimizing the variance of $d_c[n]$ subject to the chosen values of β and M

Solution for Combined Predictor

- Mean-squared prediction error for pitch predictor is:

$$E_1 = \left\langle (v[n])^2 \right\rangle = \left\langle \left(x[n] - \beta x[n-M] \right)^2 \right\rangle$$

- where $\langle \ \rangle$ denotes averaging over a finite frame of speech samples.
- We use the covariance-type of averaging to eliminate windowing effects, giving the solution:

$$(\beta)_{opt} = \frac{\langle x[n]x[n-M] \rangle}{\left\langle \left(x[n-M] \right)^2 \right\rangle}$$

- Using this value of β, we solve for $(E_1)_{opt}$ as

$$(E_1)_{opt} = \left\langle \left(x[n] \right)^2 \right\rangle \left(1 - \frac{(\langle x[n]x[n-M] \rangle)^2}{\left\langle \left(x[n] \right)^2 \right\rangle \left\langle \left(x[n-M] \right)^2 \right\rangle} \right)$$

- with minimum normalized covariance:

$$\rho[M] = \frac{\langle x[n]x[n-M] \rangle}{\left(\left\langle \left(x[n] \right)^2 \right\rangle \left\langle \left(x[n-M] \right)^2 \right\rangle \right)^{1/2}}$$

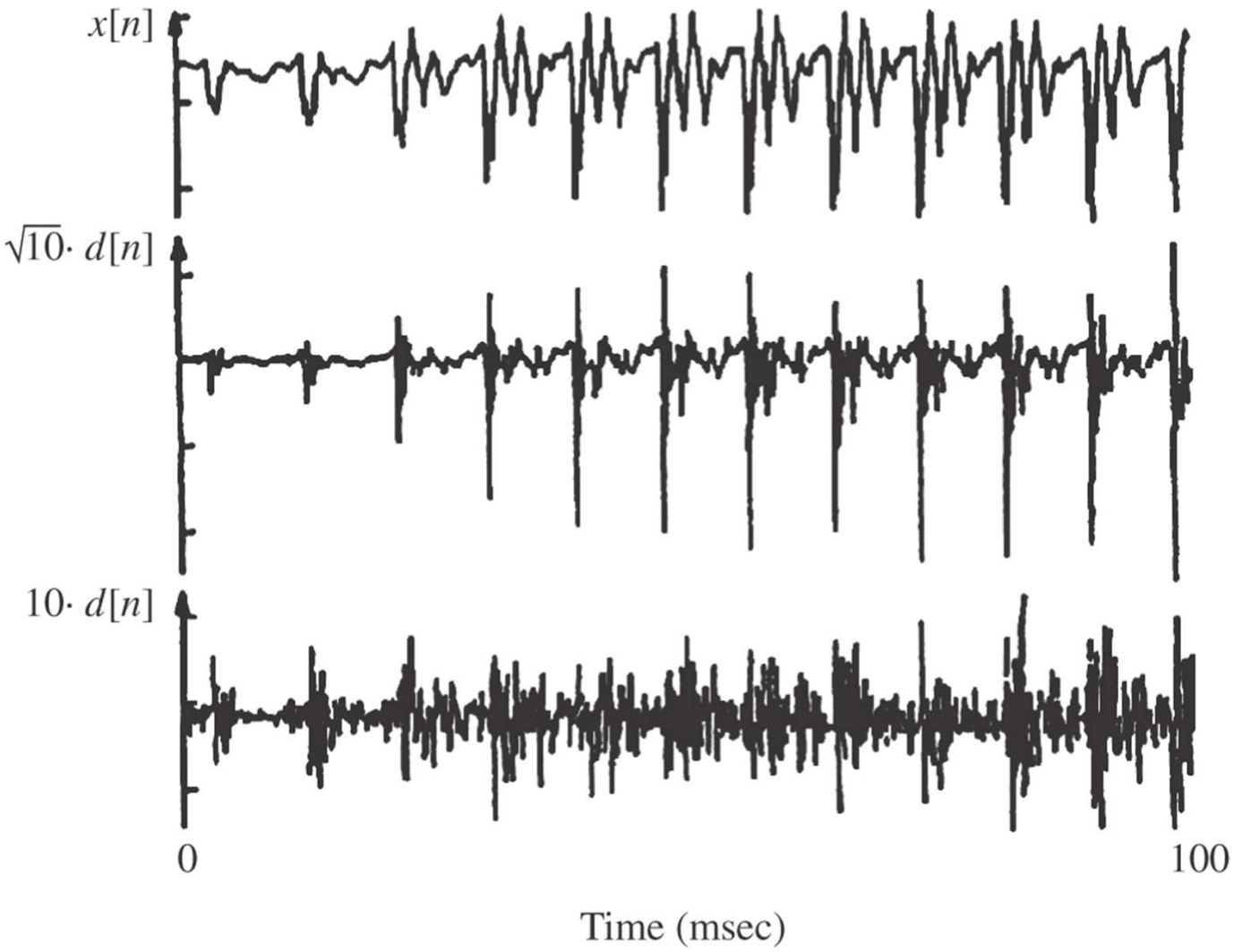
Solution for Combined Predictor

- Steps in solution:
  - first search for *M* that maximizes $\rho[M]$
  - compute β_{opt}
- Solve for more accurate pitch predictor by minimizing the variance of the expanded pitch predictor
- Solve for optimum vocal tract predictor coefficients, α_k, $k=1,2,\ldots,p$

Pitch Prediction

Vocal tract prediction

Pitch and vocal tract prediction



Noise Shaping in DPCM Systems

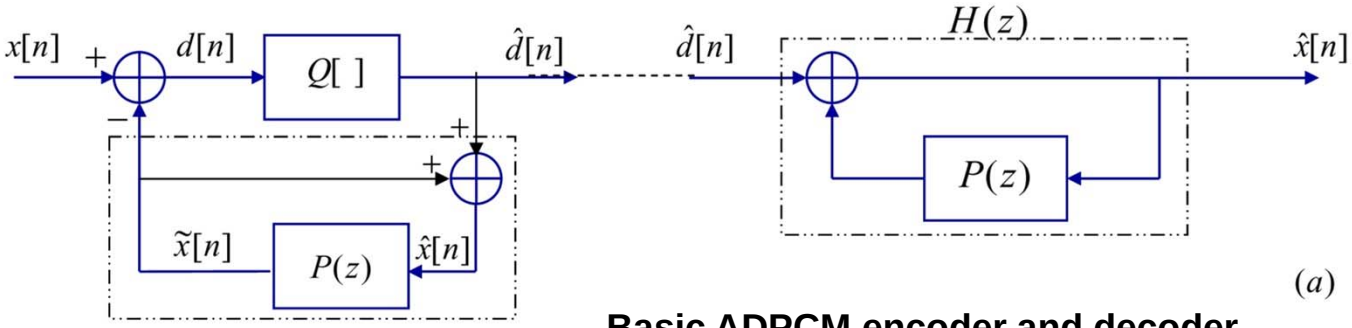
Noise Shaping Fundamentals

- The output of an ADPCN encoder/decoder is:

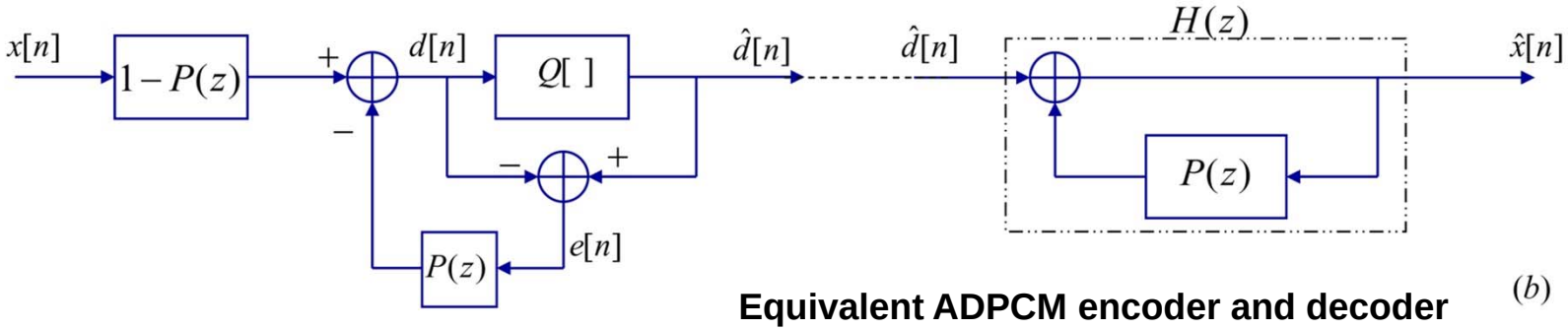
$$\hat{x}[n] = x[n] + e[n]$$

- where $e[n]$ is the quantization noise. It is easy to show that $e[n]$ generally has a flat spectrum and thus is especially audible in spectral regions of low intensity, i.e., between formants.
- This has led to methods of shaping the quantization noise to match the speech spectrum and take advantage of spectral masking concepts

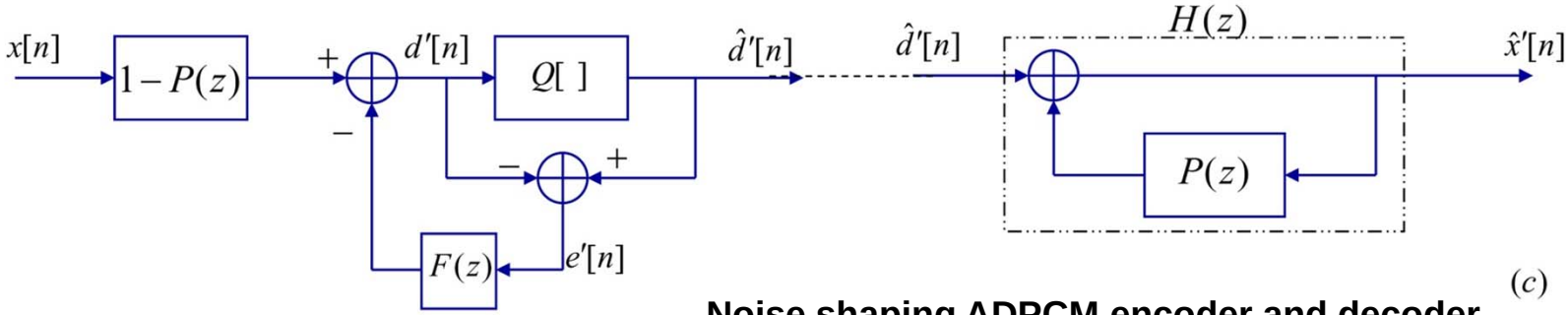
Noise Shaping



Basic ADPCM encoder and decoder



Equivalent ADPCM encoder and decoder



Noise shaping ADPCM encoder and decoder

Noise Shaping

- The equivalence of parts (b) and (a) is shown by the following:

$$\hat{x}[n] = x[n] + e[n] \leftrightarrow \hat{X}(z) = X(z) + E(z)$$

- From part (a) we see that:

$$D(z) = X(z) - P(z)\hat{X}(z) = [1 - P(z)]X(z) - P(z)E(z)$$

- with

$$E(z) = \hat{D}(z) - D(z)$$

- showing the equivalence of parts (b) and (a). Further, since

$$\hat{D}(z) = D(z) + E(z) = [1 - P(z)]X(z) + [1 - P(z)]E(z)$$

$$\hat{X}(z) = H(z)\hat{D}(z) = \left(\frac{1}{1 - P(z)}\right)\hat{D}(z)$$

$$= \left(\frac{1}{1 - P(z)}\right)([1 - P(z)]X(z) + [1 - P(z)]E(z))$$

$$= X(z) + E(z)$$

- Feeding back the quantization error through the predictor, $P(z)$ ensures that the reconstructed signal, $\hat{x}[n]$, differs from $x[n]$, by the quantization error, $e[n]$, incurred in quantizing the difference signal, $d[n]$.

Shaping the Quantization Noise

- To shape the quantization noise we simply replace $P(z)$ by a different system function, $F(z)$, giving the reconstructed signal as:

$$\hat{X}'(z) = H(z)\hat{D}'(z) = \left(\frac{1}{1-P(z)}\right)\hat{D}'(z)$$

$$= \left(\frac{1}{1-P(z)}\right)\left(\left[1-P(z)\right]X(z) + \left[1-F(z)\right]E'(z)\right)$$

$$= X(z) + \left(\frac{1-F(z)}{1-P(z)}\right)E'(z)$$

- Thus if $x[n]$ is coded by the encoder, the z-transform of the reconstructed signal at the receiver is:

$$\hat{X}'(z) = X(z) + \tilde{E}'(z)$$

$$\tilde{E}'(z) = \left(\frac{1-F(z)}{1-P(z)}\right)E'(z) = \Gamma(z)E'(z)$$

- where $\Gamma(z) = \left(\frac{1-F(z)}{1-P(z)}\right)$ is the effective noise shaping filter

Noise Shaping Filter Options

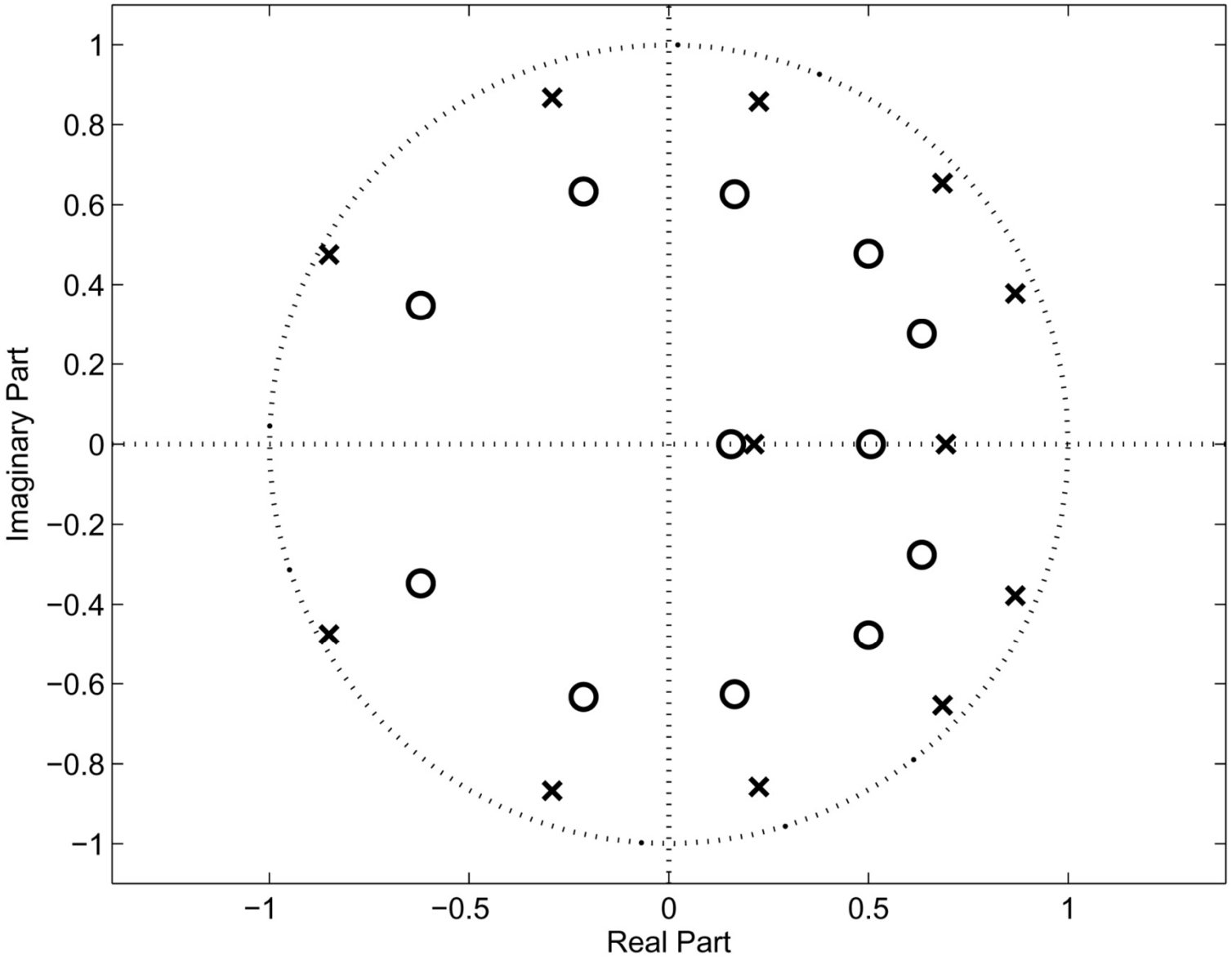
- Noise shaping filter options:

1. $F(z) = 0$ and we assume noise has a flat spectrum, then the noise and speech spectrum have the same shape

2. $F(z) = P(z)$ then the equivalent system is the standard DPCM system where $\tilde{E}'(z) = E'(z) = E(z)$ with flat noise spectrum

3. $F(z) = P(\gamma^{-1}z) = \sum_{k=1}^{p} \alpha_k \gamma^k z^{-k}$ and we "shape" the noise spectrum to "hide" the noise beneath the spectral peaks of the speech signal; each zero of $[1 - P(z)]$ is paired with a zero of $[1 - F(z)]$ where the paired zeros have the same angles in the z-plane, but with a radius that is divided by γ.

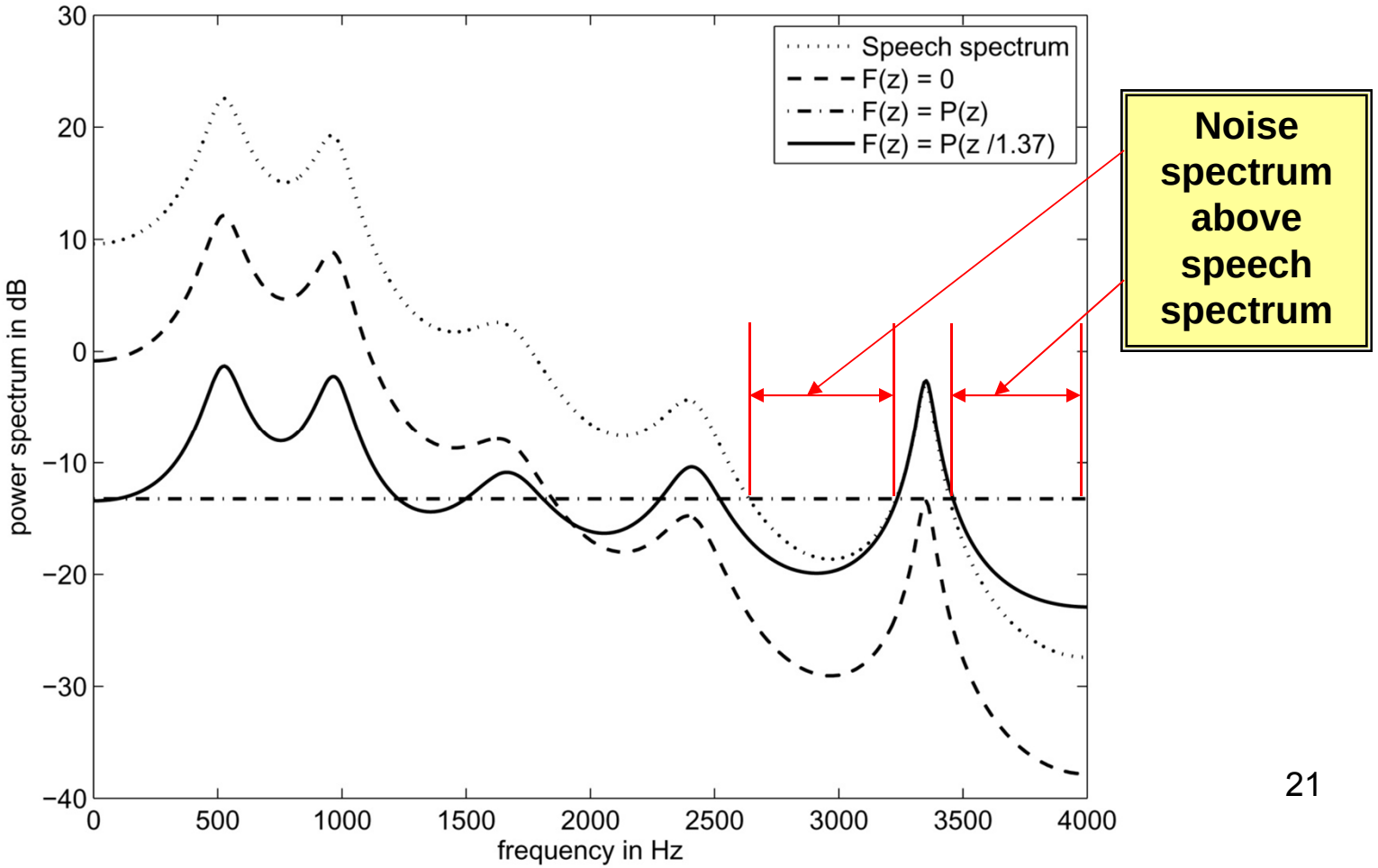
Noise Shaping Filter



Noise Shaping Filter

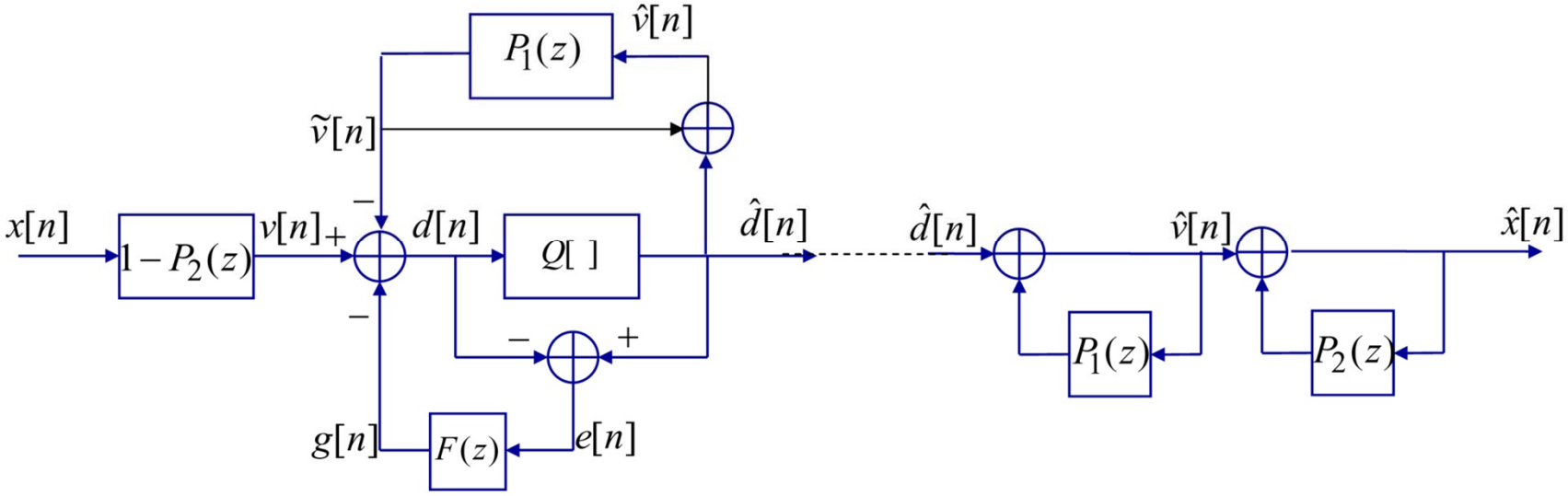
If we assume that the quantization noise has a flat spectrum with noise power of $\sigma_{e'}^2$, then the power spectrum of the shaped noise is of the form:

$$P_{e'}(e^{j2\pi F/F_S}) = \left|\frac{1-F(e^{j2\pi F/F_S})}{1-P(e^{j2\pi F/F_S})}\right|\sigma_{e'}^2$$



Fully Quantized Adaptive Predictive Coder

Full ADPCM Coder



- Input is $x[n]$
- $P_2(z)$ is the short-term (vocal tract) predictor
- Signal $v[n]$ is the short-term prediction error
- Goal of encoder is to obtain a quantized representation of this excitation signal, from which the original signal can be reconstructed.

Quantized ADPCM Coder

- Total bit rate for ADPCM coder:

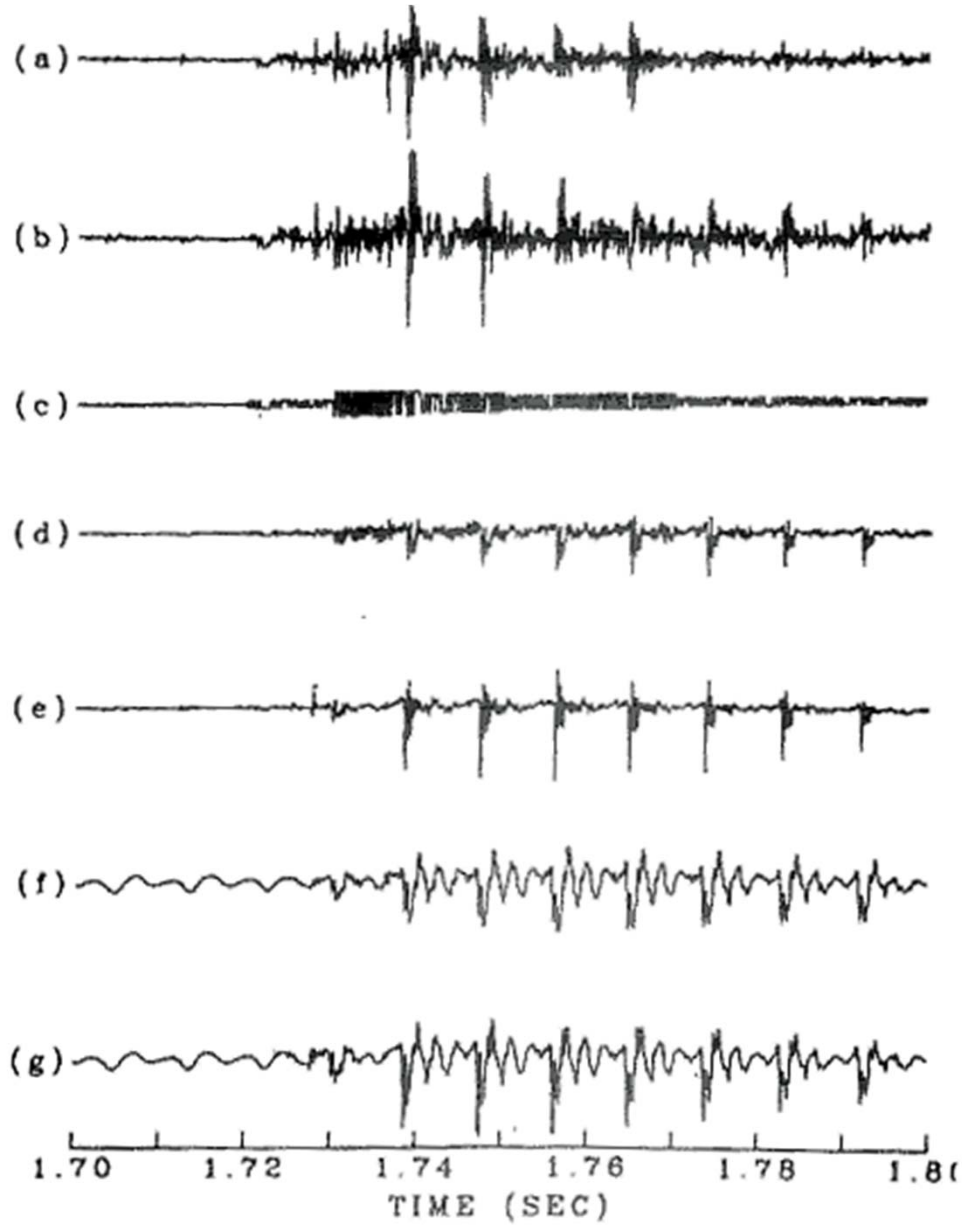
$$I_{ADPCM} = BF_S + B_\Delta F_\Delta + B_P F_P$$

- where B is the number of bits for the quantization of the difference signal, B_Δ is the number of bits for encoding the step size at frame rate F_Δ, and B_P is the total number of bits allocated to the predictor coefficients (both long and short-term) with frame rate F_P
- Typically $F_S = 8000$ and even with $B \approx 1-4$ bits, we need between 8000 and 3200 bps for quantization of difference signal
- Typically we need about 3000-4000 bps for the side information (step size and predictor coefficients)
- Overall we need between 11,000 and 36,000 bps for a fully quantized system

Bit Rate for LP Coding

- speech and residual sampling rate: F_s=8 kHz
- LP analysis frame rate: F_Δ=F_P = 50-100 frames/sec
- quantizer stepsize: 6 bits/frame
- predictor parameters:
  - *M* (pitch period): 7 bits/frame
  - pitch predictor coefficients: 13 bits/frame
  - vocal tract predictor coefficients: PARCORs 16-20, 46-50 bits/frame
- prediction residual: 1-3 bits/sample
- total bit rate:
  - $BR = 72*F_P + F_s$ (minimum)

Two-Level (B=1 bit) Quantizer



Prediction residual

Quantizer input

Quantizer output

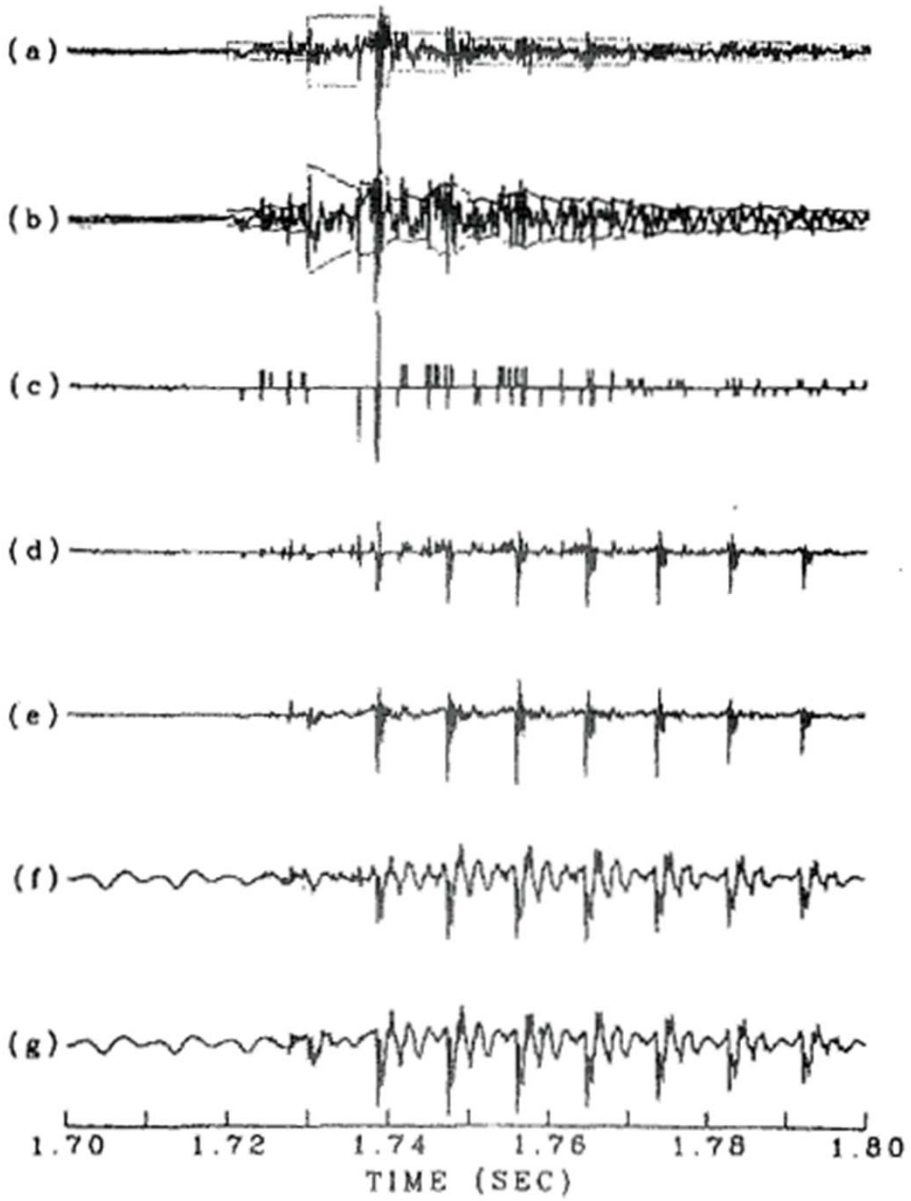
Reconstructed pitch

Original pitch residual

Reconstructed speech

Original speech

Three-Level Center-Clipped Quantizer



Prediction residual

Quantizer input

Quantizer output

Reconstructed pitch

Original pitch residual

Reconstructed speech

Original speech

Summary of Using LP in Speech Coding

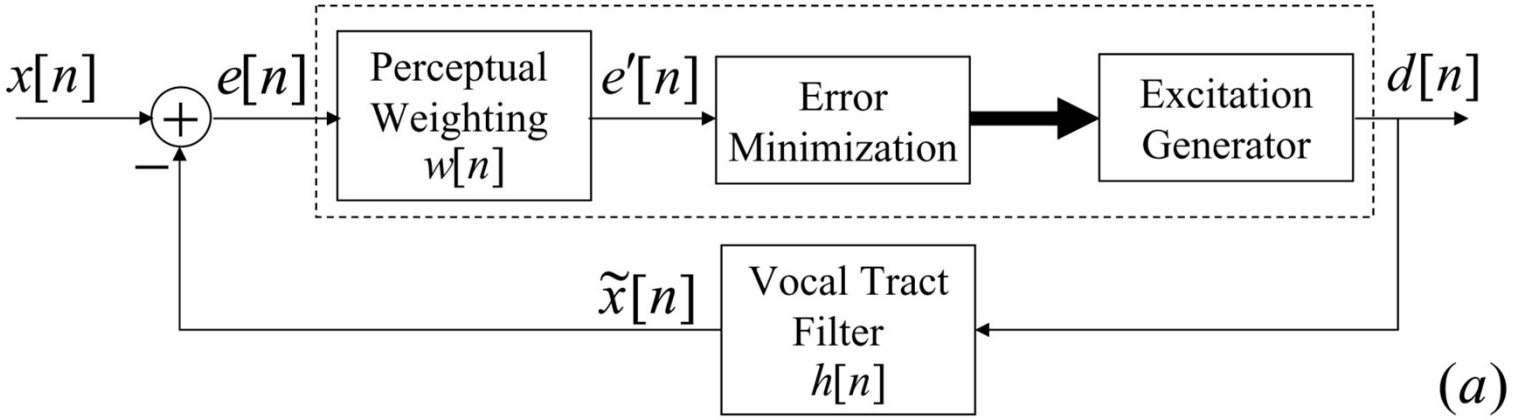
- the predictor can be more sophisticated than a vocal tract response predictor—can utilize periodicity (for voiced speech frames)
- the quantization noise spectrum can be shaped by noise feedback
  - key concept is to hide the quantization noise under the formant peaks in the speech, thereby utilizing the perceptual masking power of the human auditory system
- we now move on to more advanced LP coding of speech using Analysis-by-Synthesis methods

Analysis-by-Synthesis Speech Coders

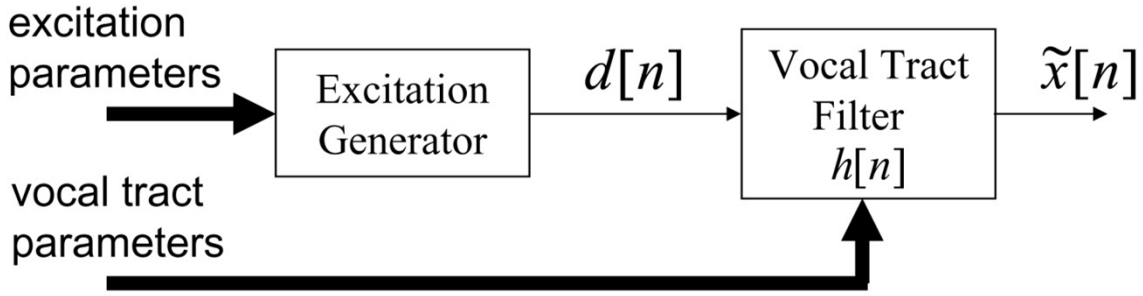
A-b-S Speech Coding

- The key to reducing the data rate of a closed-loop adaptive predictive coder was to force the coded difference signal (the input/excitation to the vocal tract model) to be more easily represented at low data rates while maintaining very high quality at the output of the decoder synthesizer

A-b-S Speech Coding



(a)



(b)

Replace quantizer for generating excitation signal with an optimization process (denoted as Error Minimization above) whereby the excitation signal, $d[n]$ is constructed based on minimization of the mean-squared value of the synthesis error, $d[n]=x[n]-\tilde{x}[n]$; utilizes Perceptual Weighting filter.

A-b-S Speech Coding

- Basic operation of each loop of closed-loop A-b-S system:
  1. at the beginning of each loop (and only once each loop), the speech signal, $x[n]$, is used to generate an optimum p^{th} order LPC filter of the form:

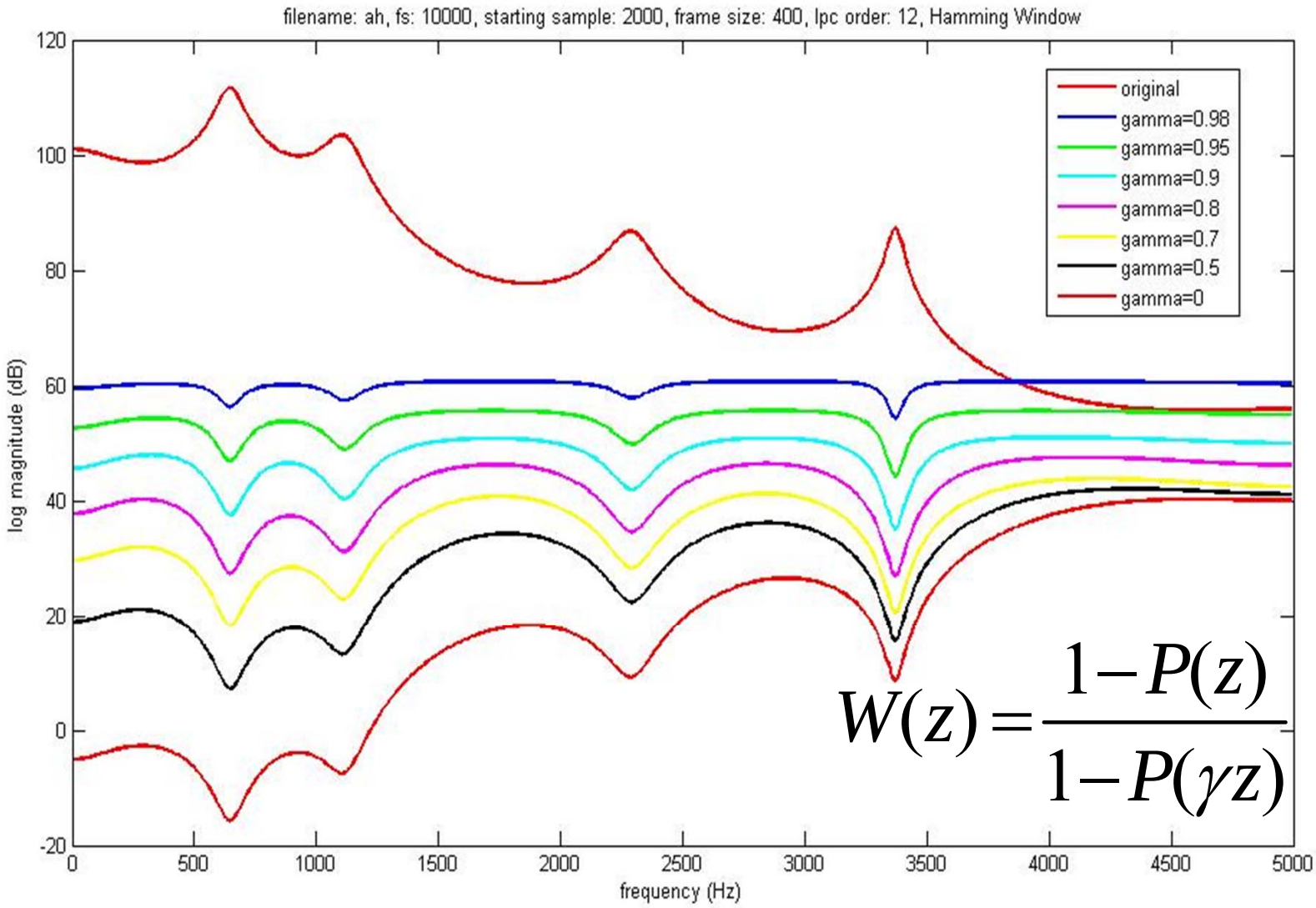
$$H(z) = \frac{1}{1 - P(z)} = \frac{1}{\sum_{i=1}^{p} \alpha_i z^{-1}}$$

  2. the difference signal, $d[n] = x[n] - \tilde{x}[n]$, based on an initial estimate of the speech signal, $\tilde{x}[n]$, is perceptually weighted by a speech-adaptive filter of the form:

$$W(z) = \frac{1 - P(z)}{1 - P(\gamma z)}$$ (see next vugraph)

  3. the error minimization box and the excitation generator create a sequence of error signals that iteratively (once per loop) improve the match to the weighted error signal
  4. the resulting excitation signal, $d[n]$, which is an improved estimate of the actual LPC prediction error signal for each loop iteration, is used to excite the LPC filter and the loop processing is iterated until the resulting error signal meets some criterion for stopping the closed-loop iterations.

Perceptual Weighting Function



As γ approaches 1, weighting is flat; as γ approaches 0, weighting becomes inverse frequency response of vocal tract.

Perceptual Weighting

- Perceptual weighting filter often modified to form:

$$W(z) = \frac{1 - P(\gamma_1 z)}{1 - P(\gamma_2 z)}, \quad 0 \le \gamma_1 \le \gamma_2 \le 1$$

- so as to make the perceptual weighting be less sensitive to the detailed frequency response of the vocal tract filter

Implementation of A-B-S Speech Coding

- **Goal**: find a representation of the excitation for the vocal tract filter that produces high quality synthetic output, while maintaining a structured representation that makes it easy to code the excitation at low data rates
- **Solution**: use a set of basis functions which allow you to iteratively build up an optimal excitation function is stages, by adding a new basis function at each iteration in the A-b-S process

Implementation of A-B-S Speech Coding

- Assume we are given a set of Q basis functions of the form:

$$\Im_\gamma = \{f_1[n], f_2[n], ..., f_Q[n]\}, \quad 0 \le n \le L-1$$

and each basis function is 0 outside the defining interval.

- At each iteration of the A-b-S loop, we select the basis function from $\Im_\gamma$ that maximally reduces the perceptually weighted mean-squared error, E :

$$E = \sum_{n=0}^{L-1} \Big[\big(x[n] - d[n] * h[n] \big) * w[n] \Big]^2$$

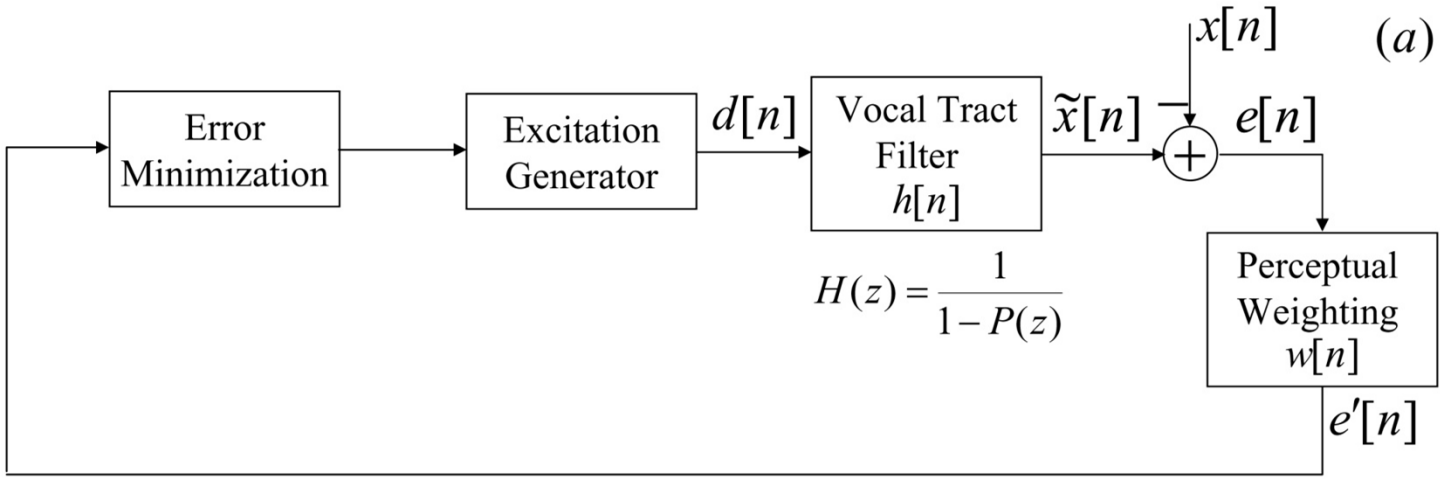
where $h[n]$ and $w[n]$ are the VT and perceptual weighting filters.

- We denote the optimal basis function at the k^{th} iteration as $f_{\gamma_k}[n]$, giving the excitation signal $d_k[n] = \beta_k \ f_{\gamma_k}[n]$ where β_k is the optimal weighting coefficient for basis function $f_{\gamma_k}[n]$.

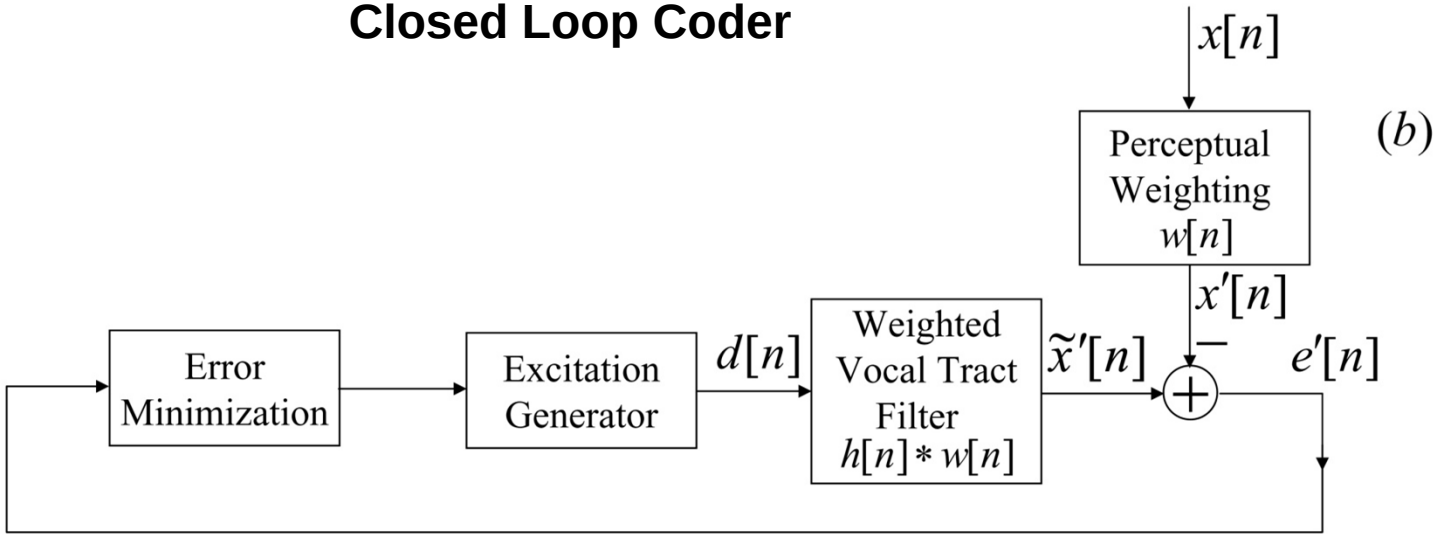
- The A-b-S iteration continues until the perceptually weighted error falls below some desired threshold, or until a maximum number of iterations, N, is reached, giving the final excitation signal, $d[n]$, as:

$$d[n] = \sum_{k=1}^{N} \beta_k \ f_{\gamma_k}[n]$$

Implementation of A-B-S Speech Coding



Closed Loop Coder



Reformulated Closed Loop Coder

Implementation of A-B-S Speech Coding

- Assume that $d[n]$ is known up to current frame ($n = 0$ for simplicity)
- Initialize 0^{th} estimate of the excitation, $d_0[n]$ as:

$$d_0[n] = \begin{cases} d[n] & n < 0 \\ 0 & 0 \le n \le L-1 \end{cases}$$

- Form the initial estimate of the speech signal as:

$$y_0[n] = \tilde{x}_0[n] = d_0[n] * h[n]$$

- since $d_0[n] = 0$ in the frame $0 \le n \le L-1$, $y_0[n]$ consists of the decaying signal from the previous frame(s). The initial (0^{th}) iteration is completed by forming the perceptually weighted difference signal as:

$$e'_0[n] = (x[n] - y_0[n]) * w[n]$$
$$= x'[n] - y'_0[n] = x'[n] - d_0[n] * h'[n]$$
$$x'[n] = x[n] * w[n]; \quad h'[n] = h[n] * w[n]$$

Implementation of A-B-S Speech Coding

- We now begin the k^{th} iteration of the A-b-S loop, $k = 1, 2, ..., N$
- We optimally select one of the $\Im_\gamma$ basis set (call this $f_{\gamma_k}[n]$) and determine the amplitude β_k giving:

$$d_k[n] = \beta_k \cdot f_{\gamma_k}[n], \quad k = 1, 2, ..., N$$

- We then form the new perceptually weighted error as:

$$e'_k[n] = e'_{k-1}[n] - \beta_k f_{\gamma_k}[n] * h'[n]$$
$$= e'_{k-1}[n] - \beta_k y'_k[n]$$

- We next define the mean-squared residual error for the k^{th} iteration as:

$$E_k = \sum_{n=0}^{L-1} \left(e'_k[n]\right)^2 = \sum_{n=0}^{L-1} \left(e'_{k-1}[n] - \beta_k y'_k[n]\right)^2$$

Implementation of A-B-S Speech Coding

- Since we assume we know γ_k we can find the optimum value of β_k by differentiating E_k with respect to β_k, giving:

$$\frac{\partial E_k}{\partial \beta_k} = -2\sum_{n=0}^{L-1}\left(e'_{k-1}[n] - \beta_k y'_k[n]\right)\cdot y'_k[n] = 0$$

letting us solve for β_k as:

$$\beta_k^{opt} = \frac{\sum_{n=0}^{L-1} e'_{k-1}[n]\cdot y'_k[n]}{\sum_{n=0}^{L-1}\left(y'_k[n]\right)^2}$$

leading to the expression of the minimum mean-squared error as:

$$E_k^{opt} = \sum_{n=0}^{L-1}\left(e'_{k-1}[n]\right)^2 - \left(\beta_k^{opt}\right)^2 \sum_{n=0}^{L-1}\left(y'_k[n]\right)^2$$

- Finally we find the optimum basis function by searching through all possible basis functions and picking the one that maximizes $\sum_{n=0}^{L-1}\left(y'_k[n]\right)^2$

Implementation of A-B-S Speech Coding

- Our final results are the relations:

$$\tilde{x}'[n] = \sum_{k=1}^{N} \beta_k f_{\gamma_k}[n] * h'[n] = \sum_{k=1}^{N} \beta_k \cdot y'_k[n]$$

$$E = \sum_{n=0}^{L-1} \left(x'[n] - \tilde{x}'[n] \right)^2 = \sum_{n=0}^{L-1} \left(x'[n] - \sum_{k=1}^{N} \beta_k \cdot y'_k[n] \right)^2$$

$$\frac{\partial E}{\partial \beta_j} = -2 \sum_{n=0}^{L-1} \left(x'[n] - \sum_{k=1}^{N} \beta_k \cdot y'_k[n] \right) \cdot y'_j[n]$$

where the re-optimized β_k's satisfy the relation:

$$\sum_{n=0}^{L-1} x'[n] y'_j[n] = \sum_{k=1}^{N} \beta_k \left(\sum_{n=0}^{L-1} y'_k[n] \cdot y'_j[n] \right), \quad j = 1, 2, ..., N$$

- At receiver use set of $f_{\gamma_k}[n]$ along with β_k to create excitation:

$$\tilde{x}[n] = \sum_{k=1}^{N} \beta_k f_{\gamma_k}[n] * h[n]$$

Analysis-by-Synthesis Coding

- Multipulse linear predictive coding (MPLPC)

$$f_\gamma[n] = \delta[n-\gamma] \qquad 0 \le \gamma \le Q-1 = L-1$$

B. S. Atal and J. R. Remde, "A new model of LPC excitation…," *Proc. IEEE Conf. Acoustics, Speech and Signal Proc.*, 1982.

- Code-excited linear predictive coding (CELP)

$f_\gamma[n]$ = vector of white Gaussian noise, $1 \le \gamma \le Q = 2^M$

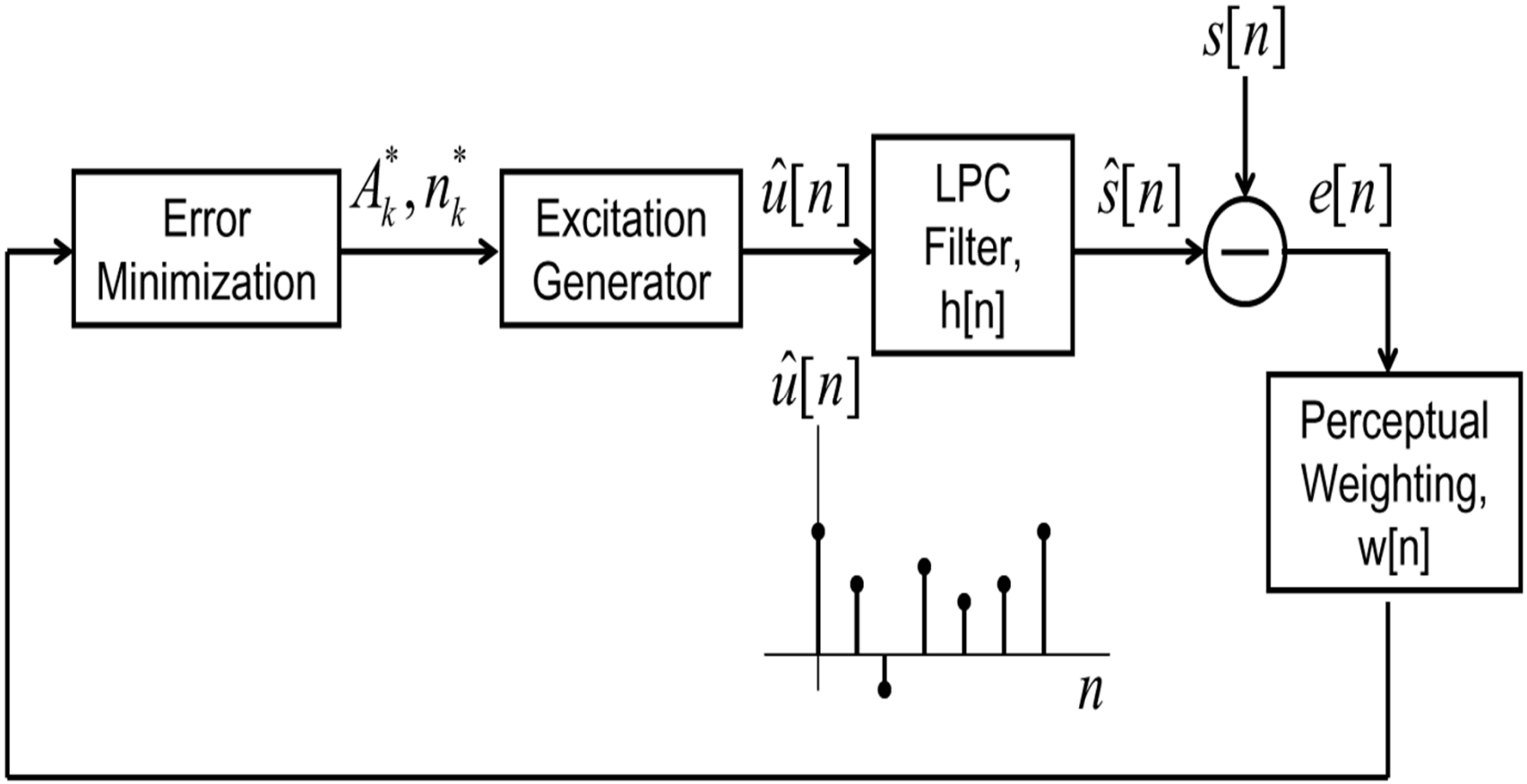
M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP)," *Proc. IEEE Conf. Acoustics, Speech and Signal Proc.*, 1985.

- Self-excited linear predictive vocoder (SEV)

$f_\gamma[n] = d[n-\gamma]$, $\Gamma_1 \le \gamma \le \Gamma_2$ – shifted versions of previous excitation source

R. C. Rose and T. P. Barnwell, "The self-excited vocoder," *Proc. IEEE Conf. Acoustics, Speech and Signal Proc.*, 1986.

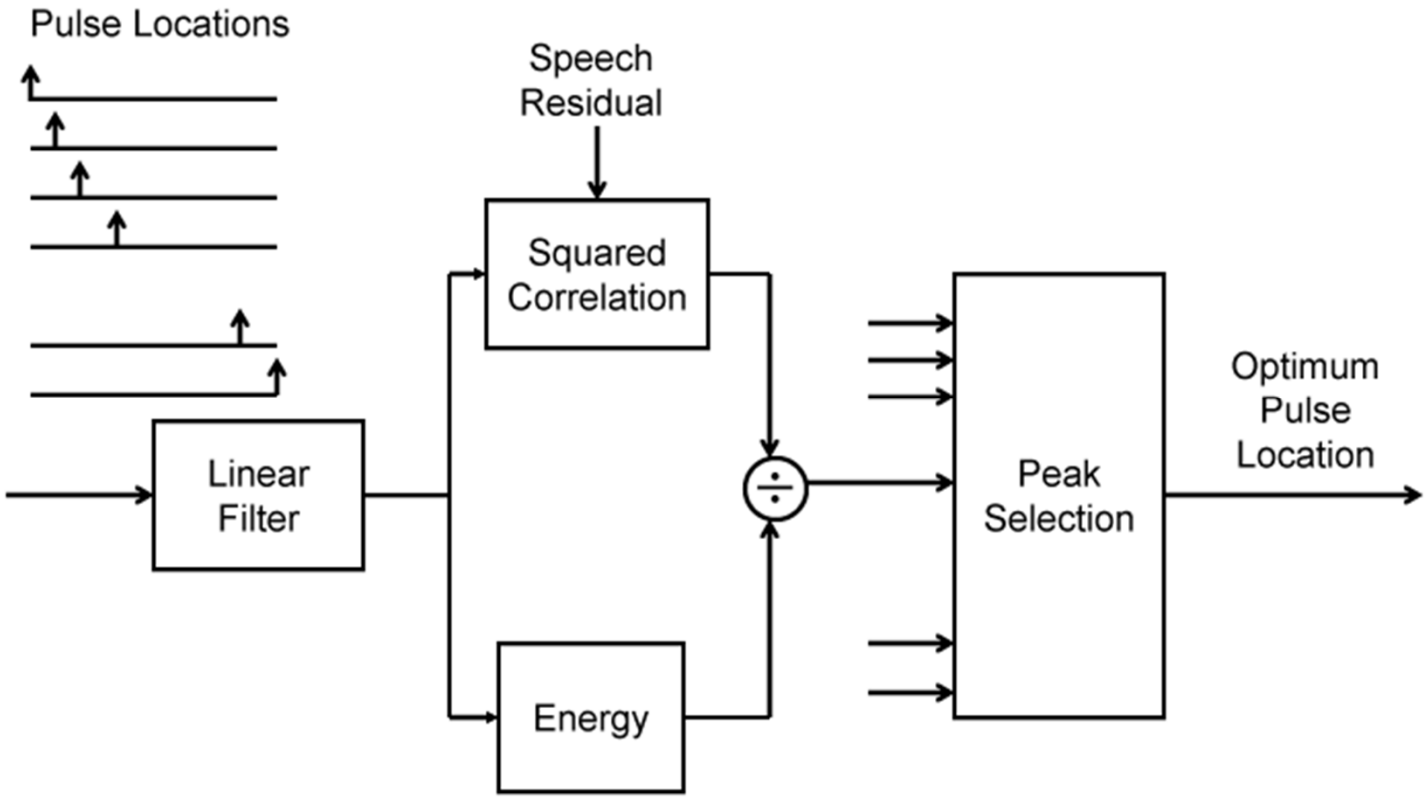
Multipulse Coder



Multipulse LP Coder

- Multipulse uses impulses as the basis functions; thus the basic error minimization reduces to:

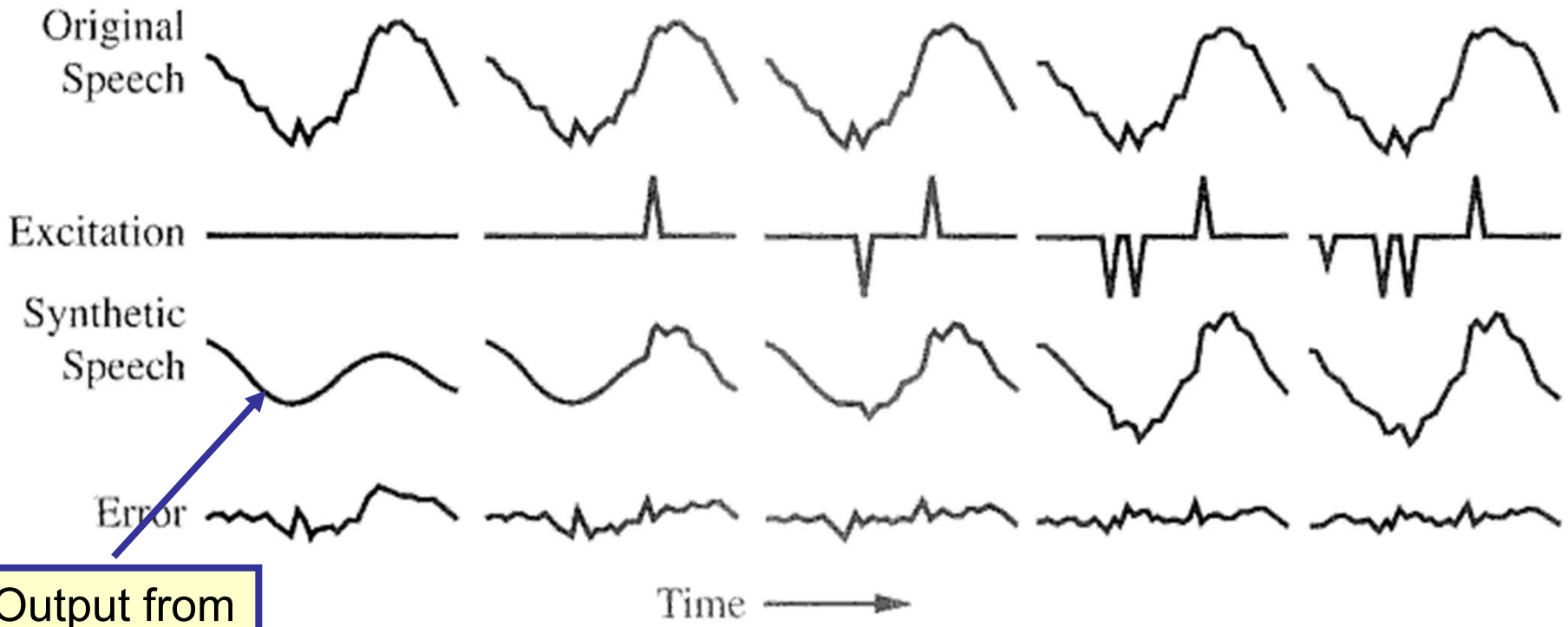
$$E = \sum_{n=0}^{L-1} \left(x[n] - \sum_{k=1}^{N} \beta_k h[n-\gamma_k] \right)^2$$



Iterative Solution for Multipulse

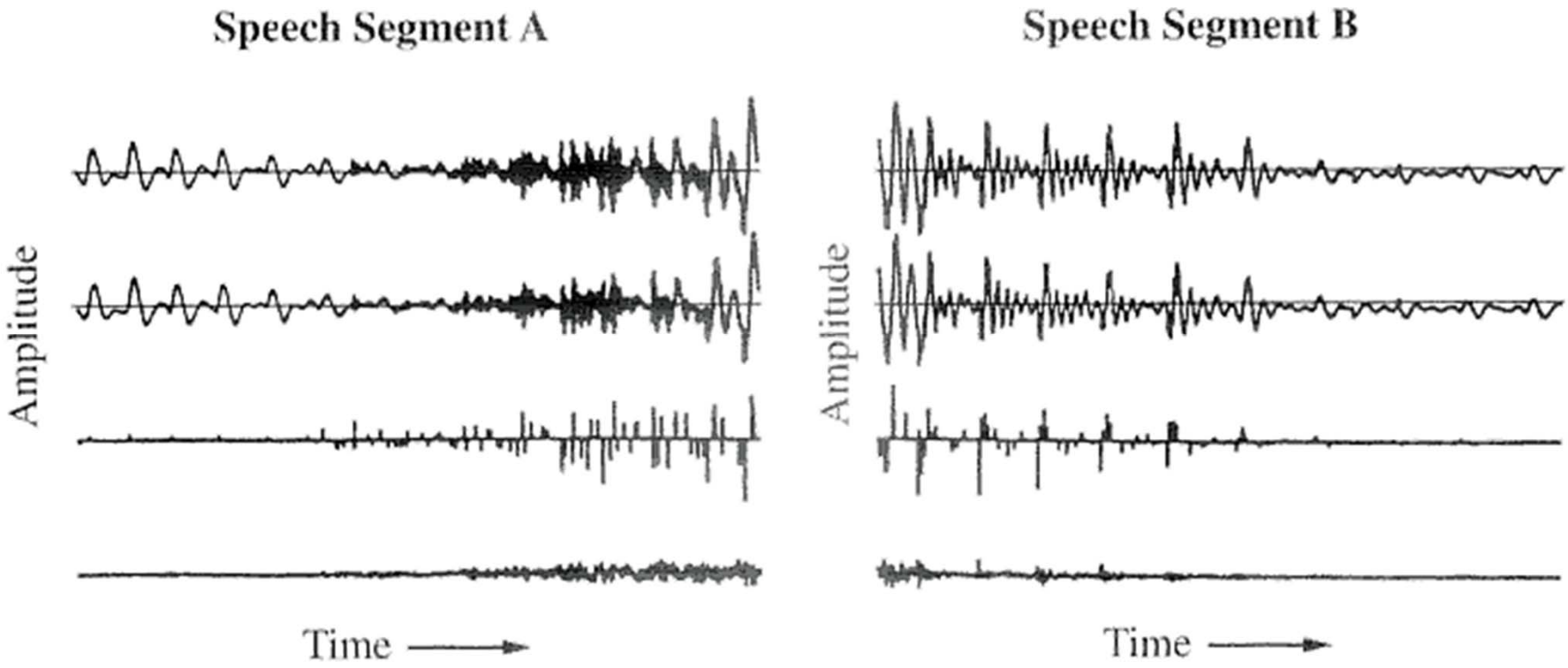
1. find best β_1 and γ_1 for single pulse solution
2. subtract out the effect of this impulse from the speech waveform and repeat the process
3. do this until desired minimum error is obtained
   - 8 impulses each 10 msec gives synthetic speech that is perceptually close to the original

Multipulse Analysis



B. S. Atal and J. R. Remde, "A new model of LPC excitation producing natural-sounding speech at low bit rates," *Proc. IEEE Conf. Acoustics, Speech and Signal Proc.*, 1982.

Examples of Multipulse LPC



B. S. Atal and J. R. Remde, "A new model of LPC Excitation Producing natural-sounding speech at low bit rates," *Proc. IEEE Conf. Acoustics, Speech and Signal Proc*., 1982.

Coding of MP-LPC

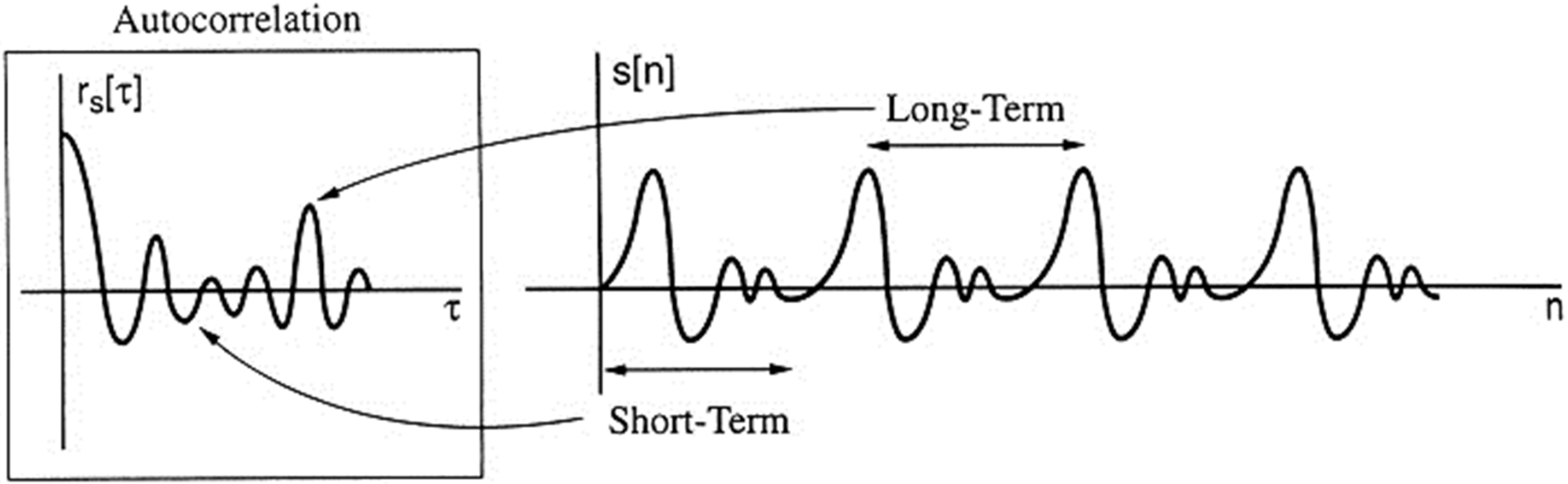
- 8 impulses per 10 msec => 800 impulses/sec X 9 bits/impulse => 7200 bps
- need 2400 bps for *A(z)* => total bit rate of 9600 bps
- code pulse locations differentially ($\Delta_i = N_i - N_{i-1}$) to reduce range of variable
- amplitudes normalized to reduce dynamic range

MPLPC with LT Prediction

- basic idea is that primary pitch pulses are correlated and predictable over consecutive pitch periods, i.e.,

$$s[n] \approx s[n-M]$$

- break correlation of speech into short term component (used to provide spectral estimates) and long term component (used to provide pitch pulse estimates)
- first remove short-term correlation by short-term prediction, followed by removing long-term correlation by long-term predictions



Short Term Prediction Error Filter

- prediction error filter

$$\hat{A}(z) = 1 - P(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}$$

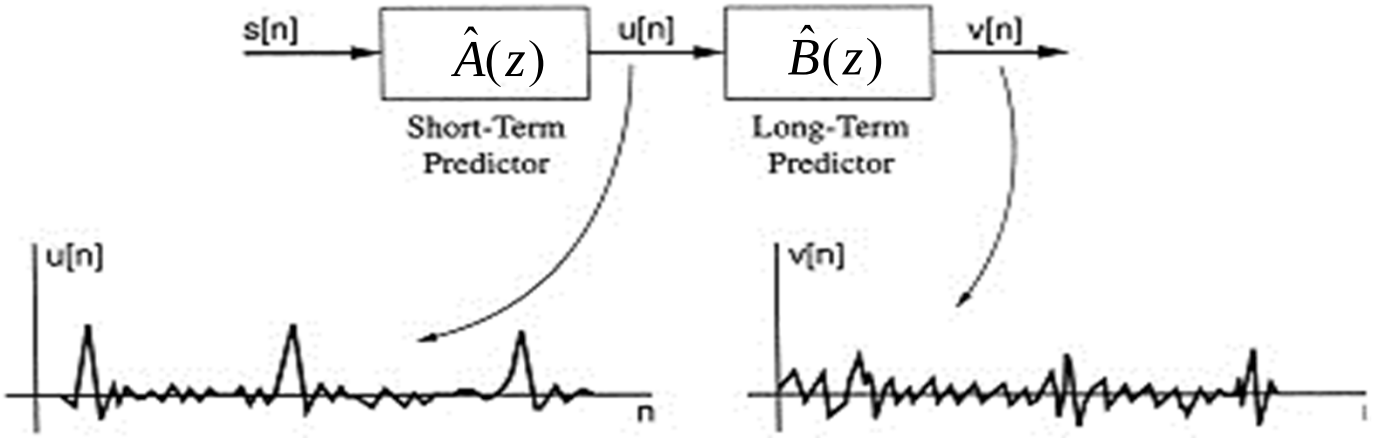
- short term residual, $u(n)$, includes primary pitch pulses that can be removed by long-term predictor of the form

$$\hat{B}(z) = 1 - bz^{-M}$$

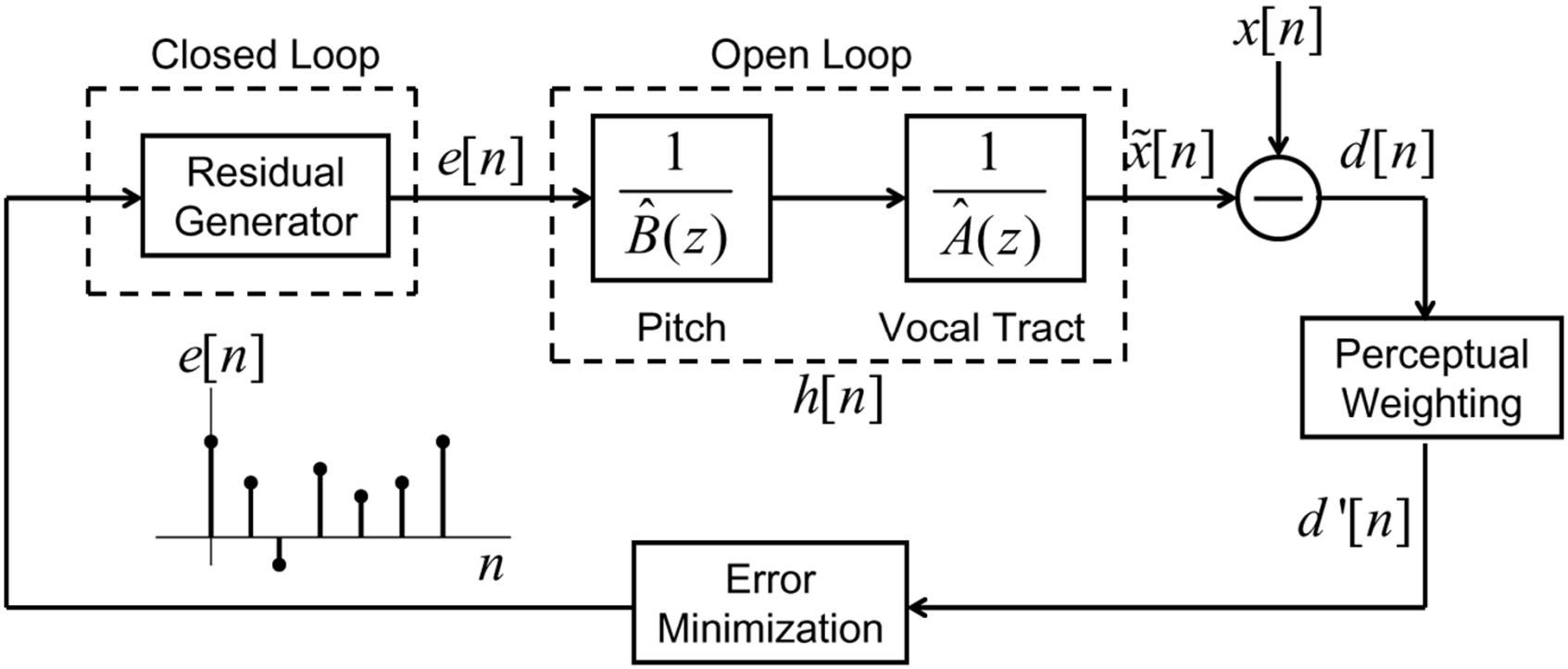
- giving

$$v(n) = u(n) - bu(n - M)$$

- with fewer large pulses to code than in $u(n)$



Analysis-by-Synthesis



- impulses selected to represent the output of the long term predictor, rather than the output of the short term predictor
- most impulses still come in the vicinity of the primary pitch pulse

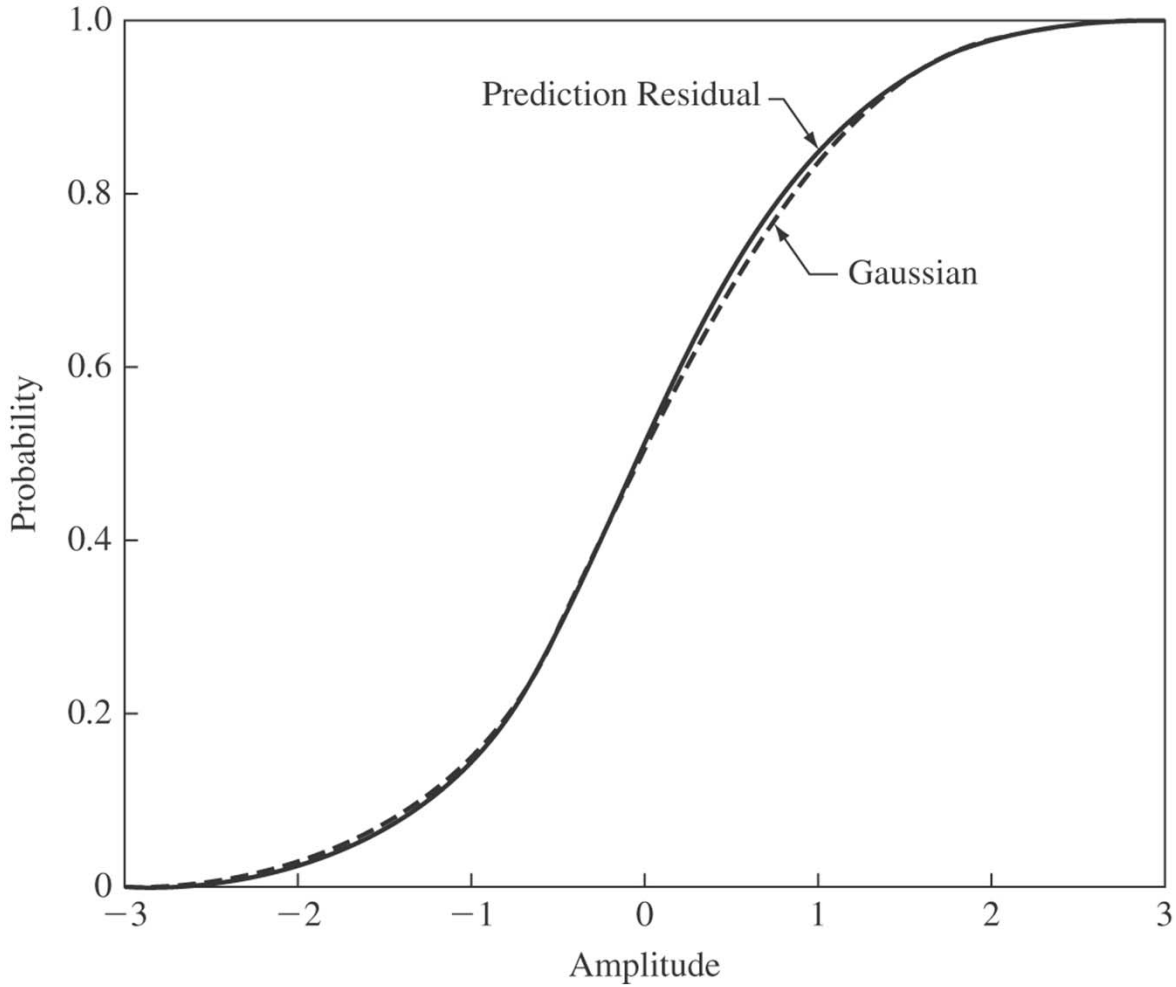
=> result is high quality speech coding at 8-9.6 Kbps

Code Excited Linear Prediction (CELP)

Code Excited LP

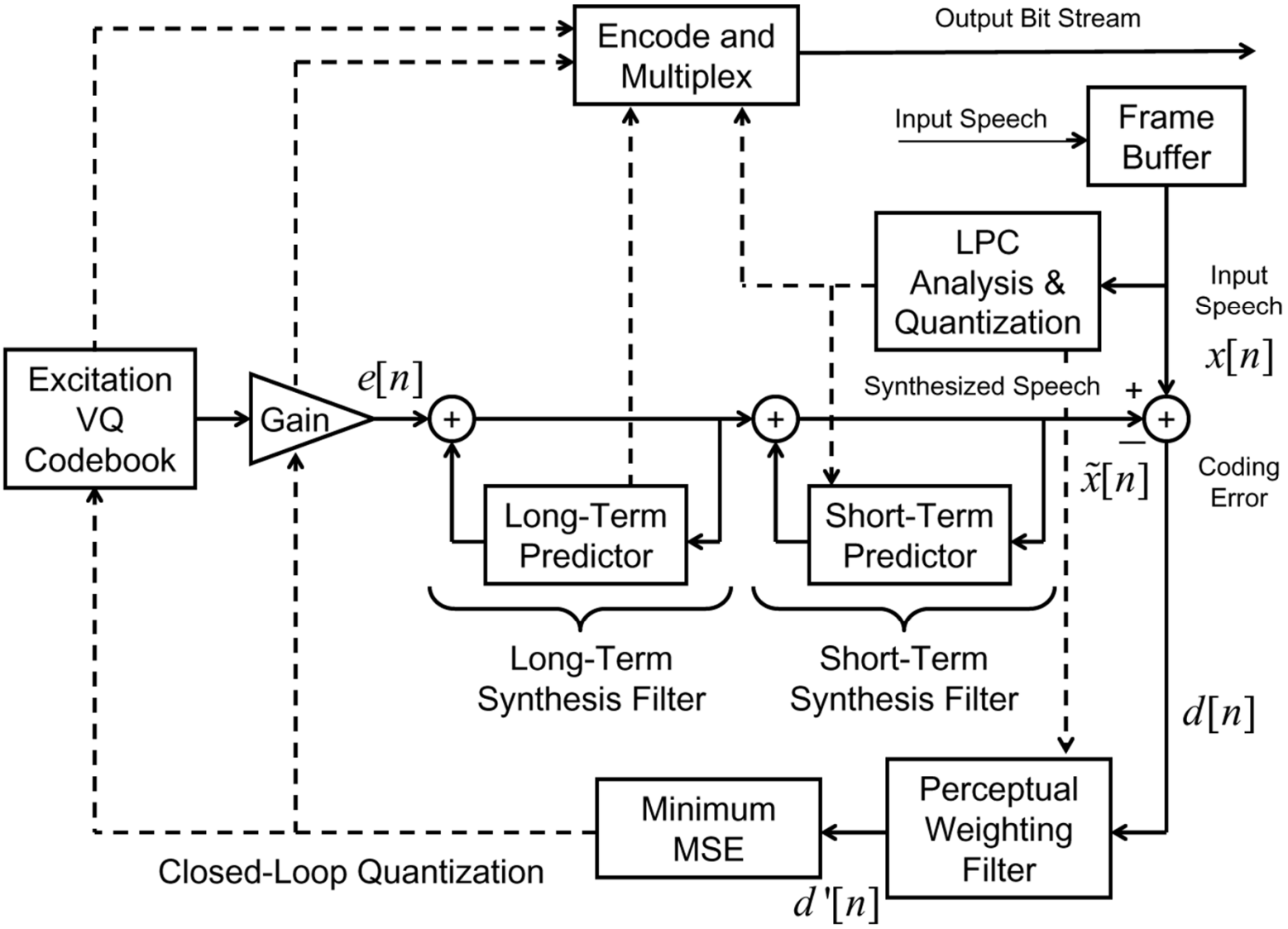
- basic idea is to represent the residual after long-term (pitch period) and short-term (vocal tract) prediction on each frame by codewords from a VQ-generated codebook, rather than by multiple pulses
- replaced residual generator in previous design by a codeword generator—40 sample codewords for a 5 msec frame at 8 kHz sampling rate
- can use either “deterministic” or “stochastic” codebook—10 bit codebooks are common
- deterministic codebooks are derived from a training set of vectors => problems with channel mismatch conditions
- stochastic codebooks motivated by observation that the histogram of the residual from the long-term predictor roughly is Gaussian pdf => construct codebook from white Gaussian random numbers with unit variance
- CELP used in STU-3 at 4800 bps, cellular coders at 800 bps

Code Excited LP



Stochastic codebooks motivated by the observation that the cumulative amplitude distribution of the residual from the long-term pitch predictor output is roughly identical to a Gaussian distribution with the same mean and variance.

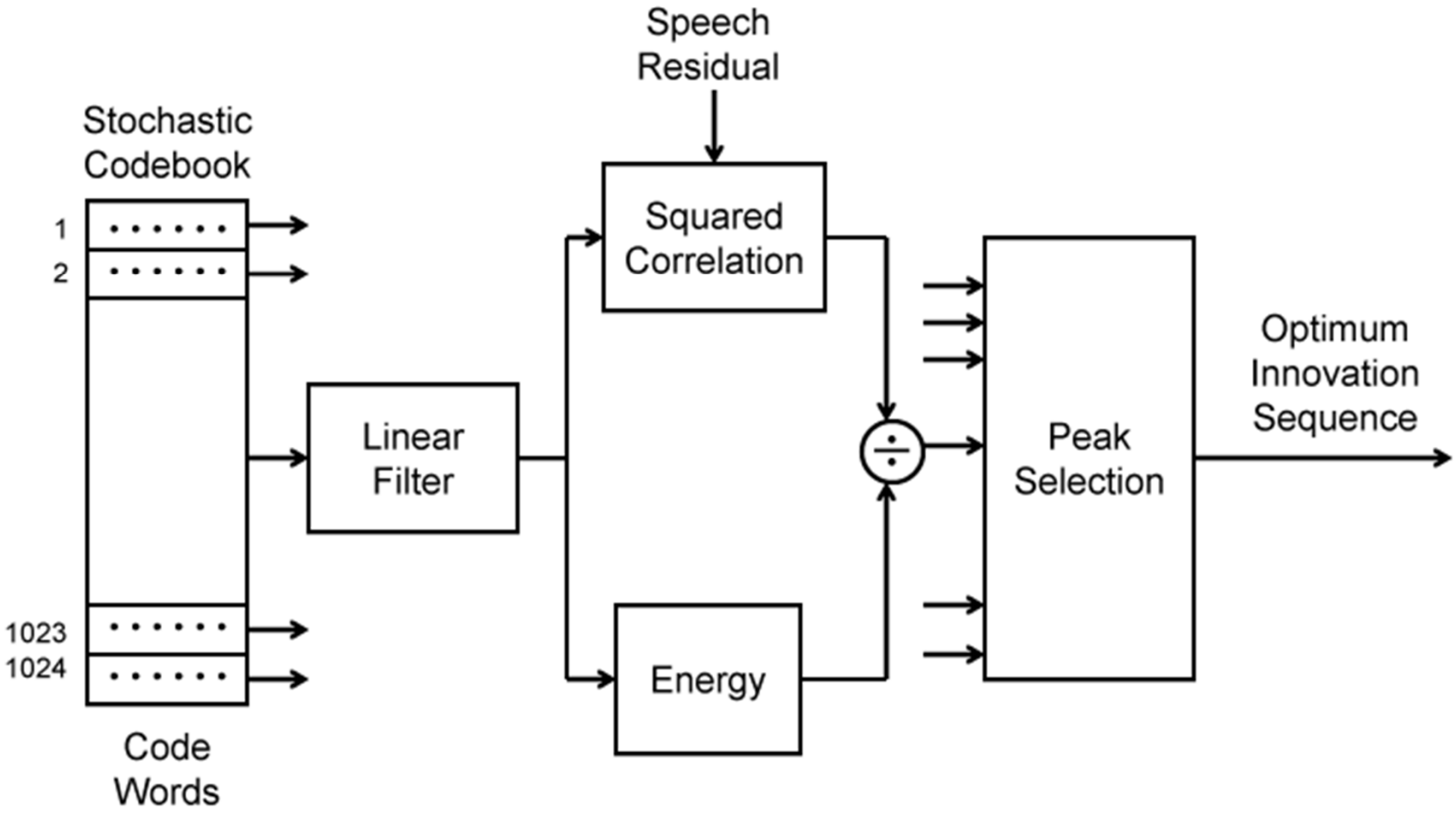
CELP Encoder



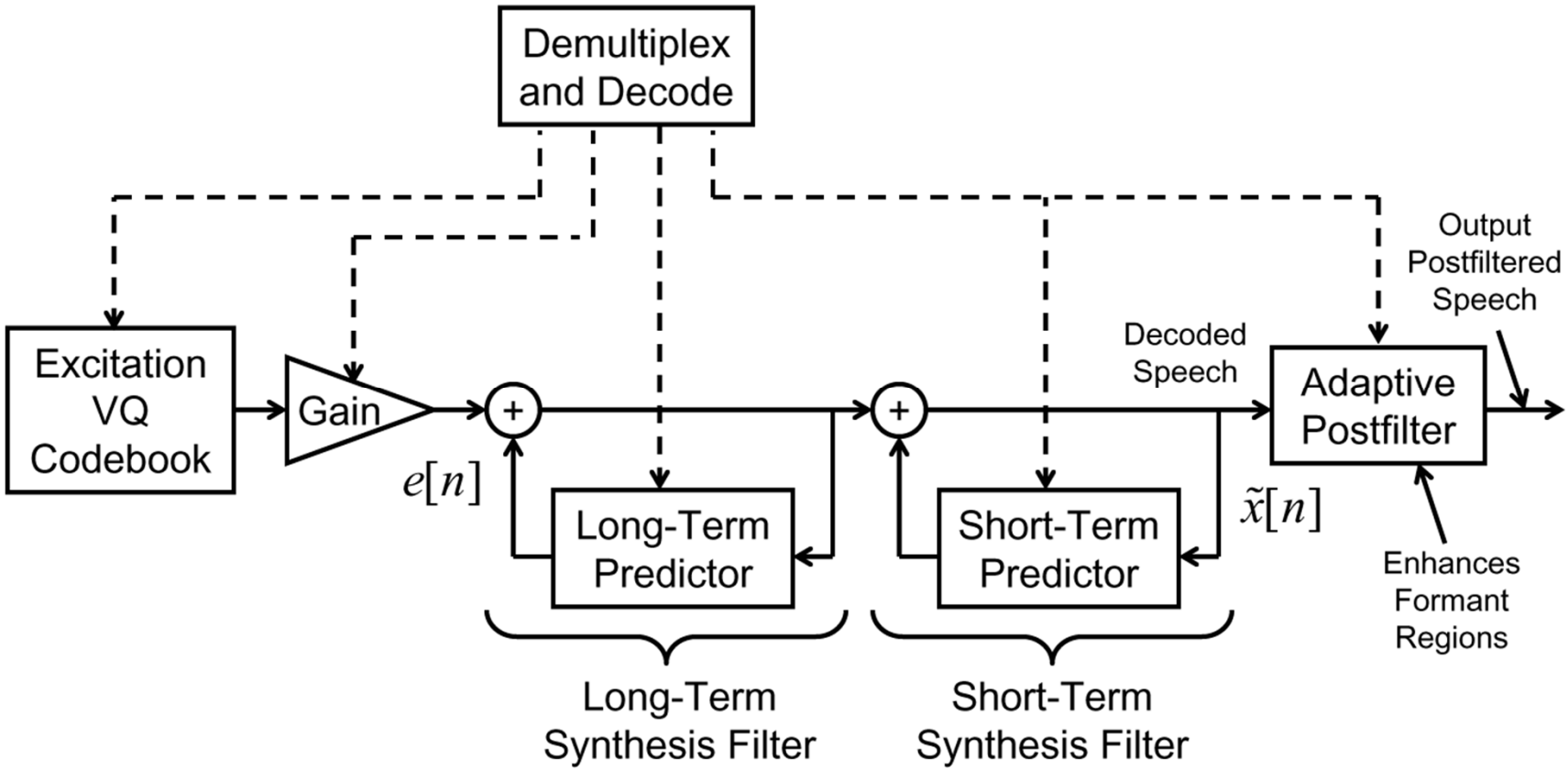
CELP Encoder

- For each of the excitation VQ codebook vectors, the following operations occur:
  - the codebook vector is scaled by the LPC gain estimate, yielding the error signal, $e[n]$
  - the error signal, $e[n]$, is used to excite the long-term pitch predictor, yielding the estimate of the speech signal, $\tilde{x}[n]$, for the current codebook vector
  - the signal, $d[n]$, is generated as the difference between the speech signal, $x[n]$, and the estimated speech signal, $\tilde{x}[n]$
  - the difference signal is perceptually weighted and the resulting mean-squared error is calculated

Stochastic Code (CELP) Excitation Analysis



CELP Decoder



CELP Decoder

- The signal processing operations of the CELP decoder consist of the following steps (for each 5 msec frame of speech):
  - select the appropriate codeword for the current frame from a matching excitation VQ codebook (which exists at both the encoder and the decoder)
  - scale the codeword sequence by the gain of the frame, thereby generating the excitation signal, $e[n]$
  - process $e[n]$ by the long-term synthesis filter (the pitch predictor) and the short-term vocal tract filter, giving the estimated speech signal, $\tilde{x}[n]$
  - process the estimated speech signal by an adaptive postfilter whose function is to enhance the formant regions of the speech signal, and thus to improve the overall quality of the synthetic speech from the CELP system

Adaptive Postfilter

- Goal is to suppress noise below the masking threshold at all frequencies, using a filter of the form:

$$H_p(z) = (1 - \mu z^{-1}) \frac{\left[1 - \sum_{k=1}^{p} \gamma_1^k \alpha_k z^{-k}\right]}{\left[1 - \sum_{k=1}^{p} \gamma_2^k \alpha_k z^{-k}\right]}$$

- where the typical ranges of the parameters are:

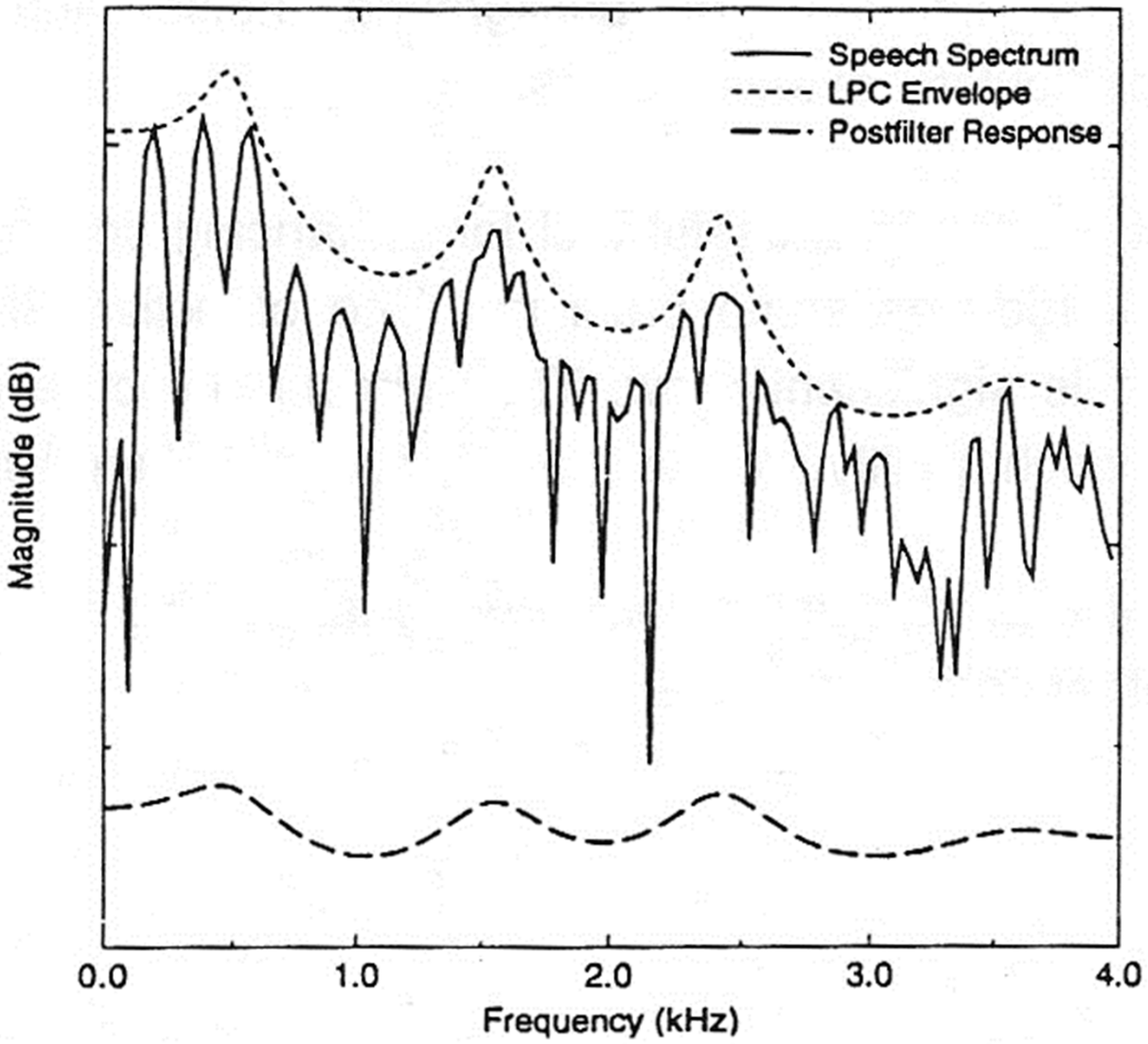
$$0.2 \le \mu \le 0.4$$

$$0.5 \le \gamma_1 \le 0.7$$

$$0.8 \le \gamma_2 \le 0.9$$

- The postfilter tends to attenuate the spectral components in the valleys without distorting the speech.

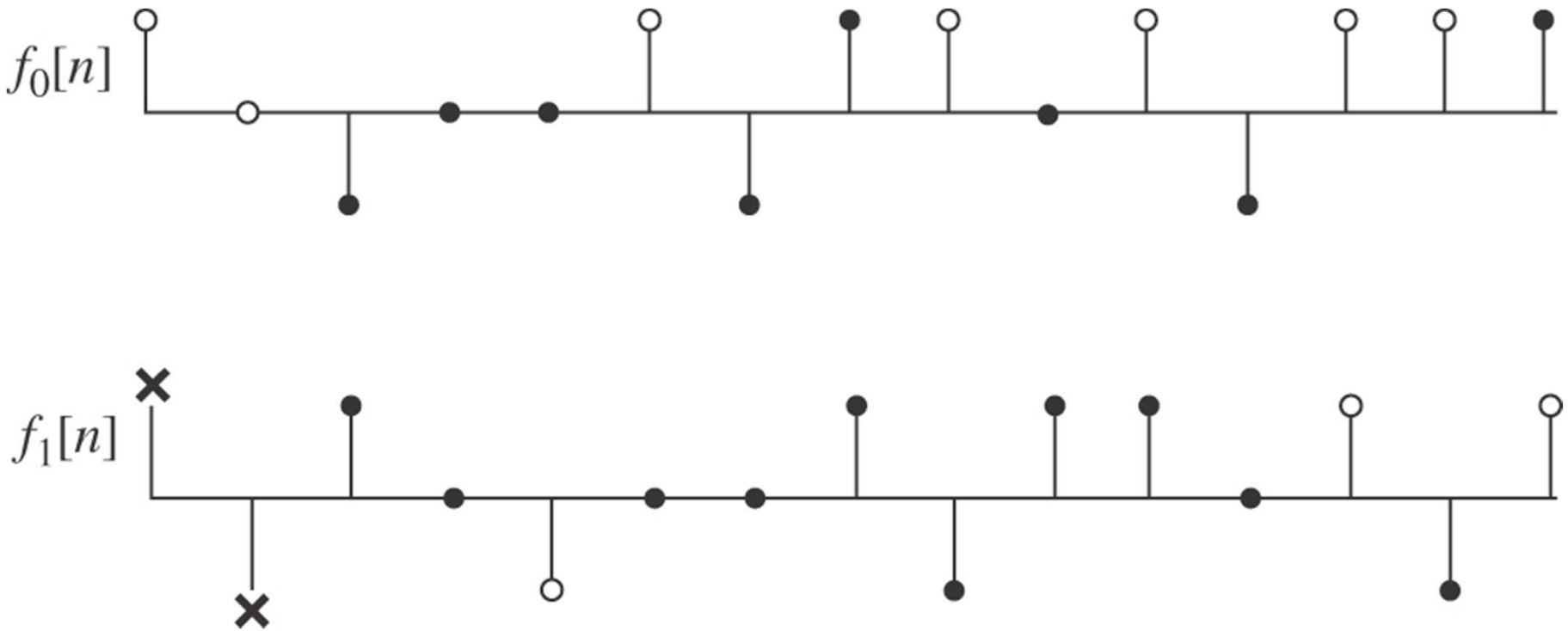
Adaptive Postfilter



CELP Codebooks

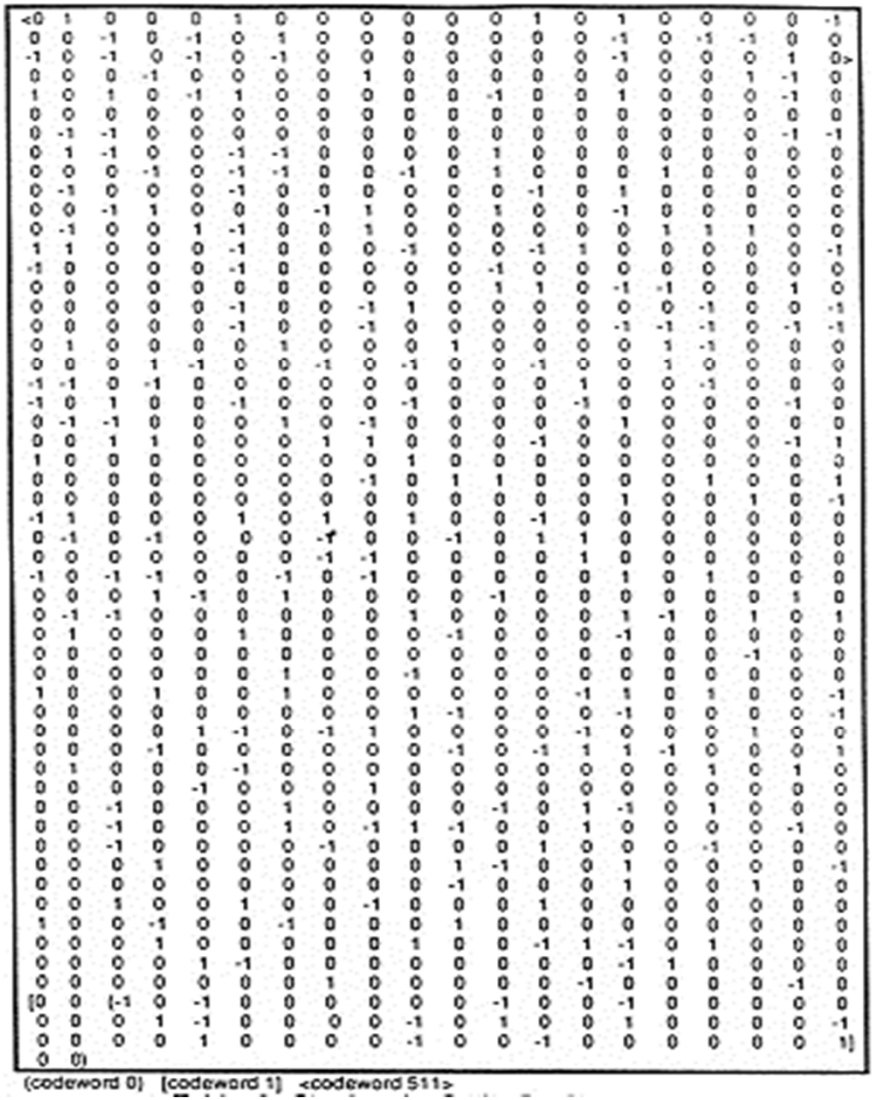
- Populate codebook from a one-dimensional array of Gaussian random numbers, where most of the samples between adjacent codewords were identical
- Such overlapping codebooks typically use shifts of one or two samples, and provide large complexity reductions for storage and computation of optimal codebook vectors for a given frame

Overlapped Stochastic Codebook

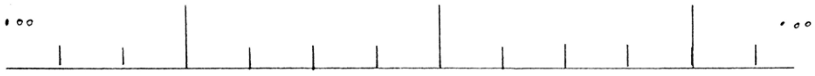


Two codewords which are identical except for a shift of two samples

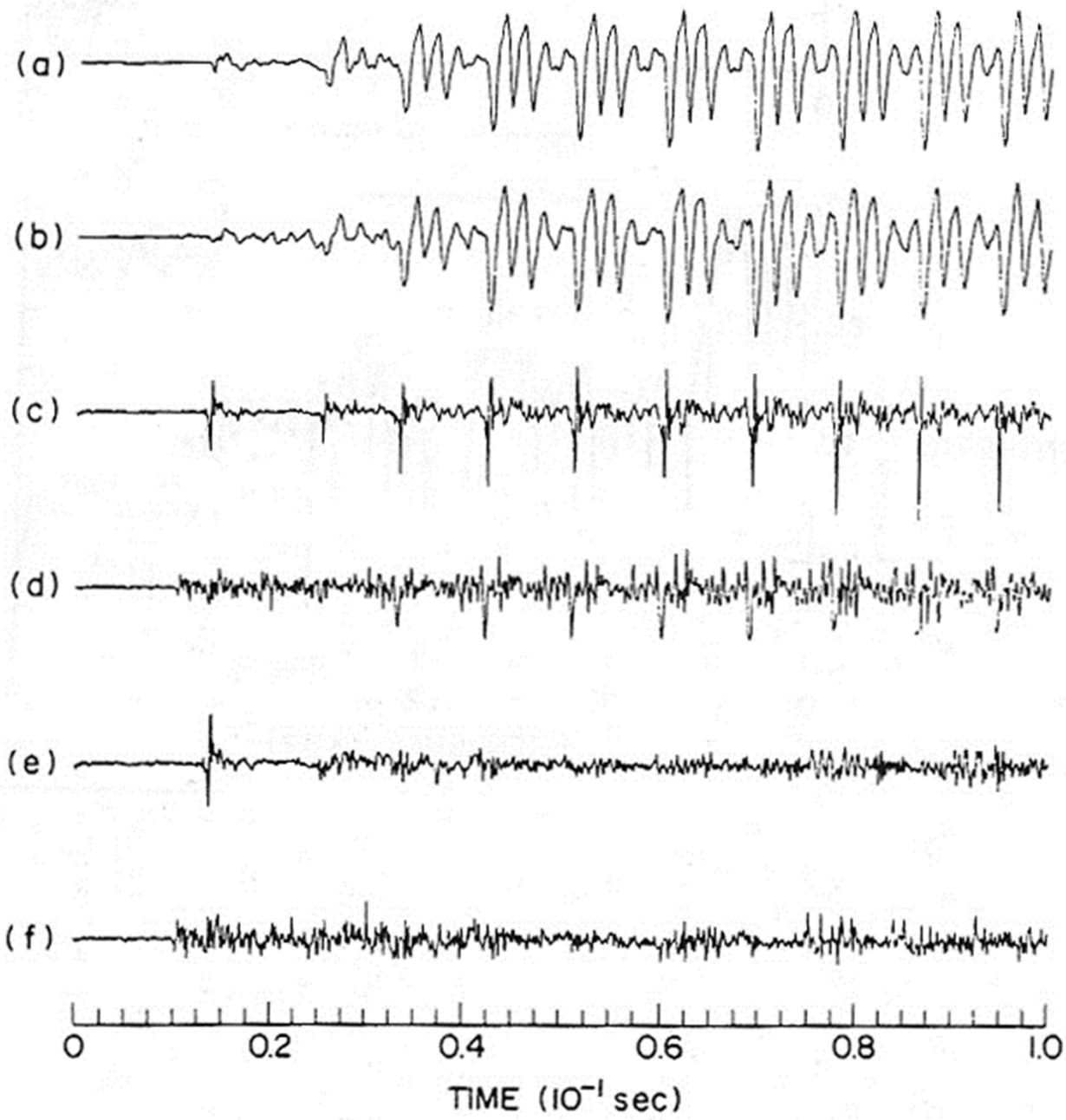
CELP Stochastic Codebook



- "Vocal tract" and excitation frame lengths are usually different; e.g. in the *Proposed Federal Standard 1016* CELP coder, the vocal tract frame length is 30 msec. (240 samples) and the excitation subframe length is 7.5 msec. (60 samples).



CELP Waveforms



(a) original speech

(b) synthetic speech output

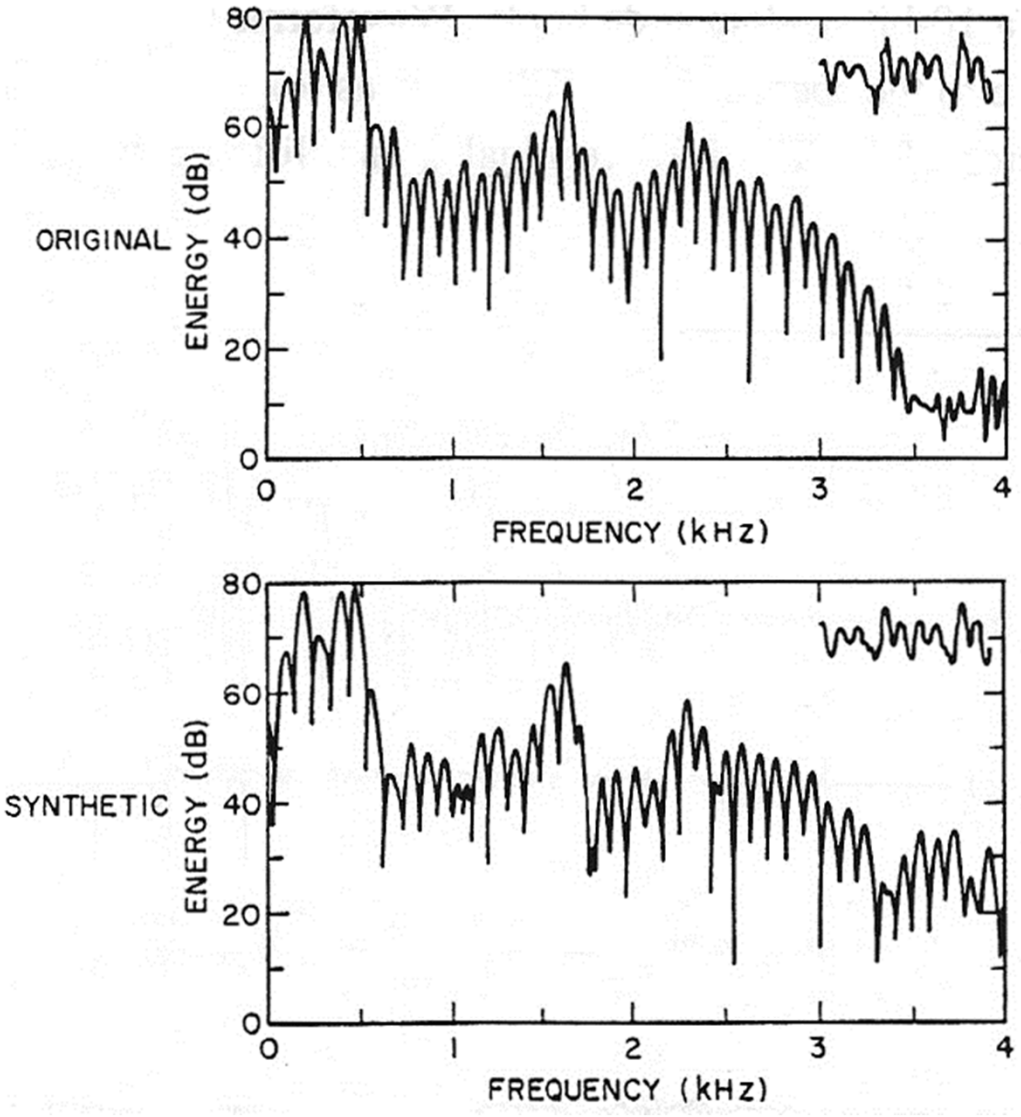
(c) LPC prediction residual

(d) reconstructed LPC residual

(e) prediction residual after pitch prediction

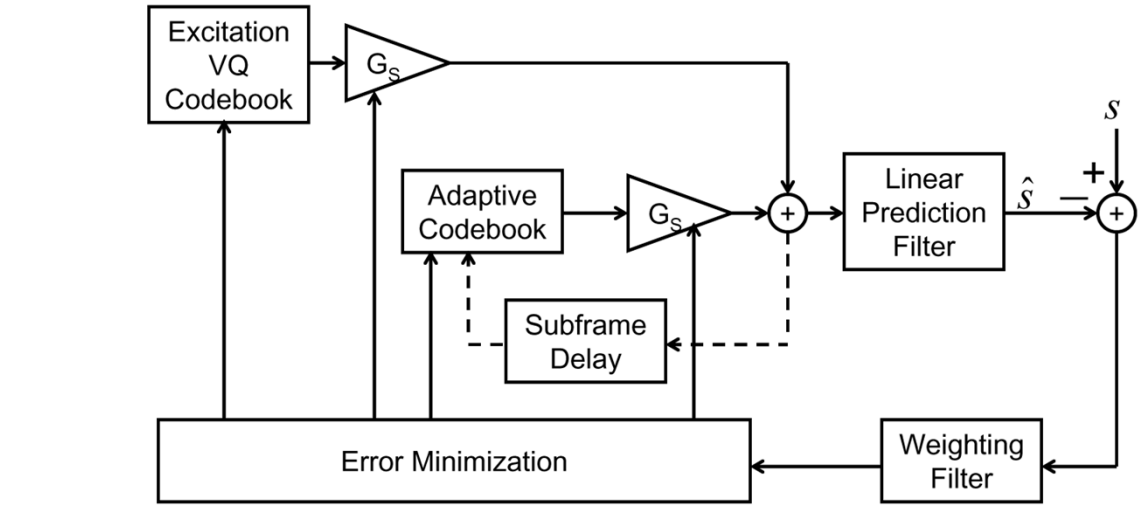
(f) coded residual from 10-bit random codebook

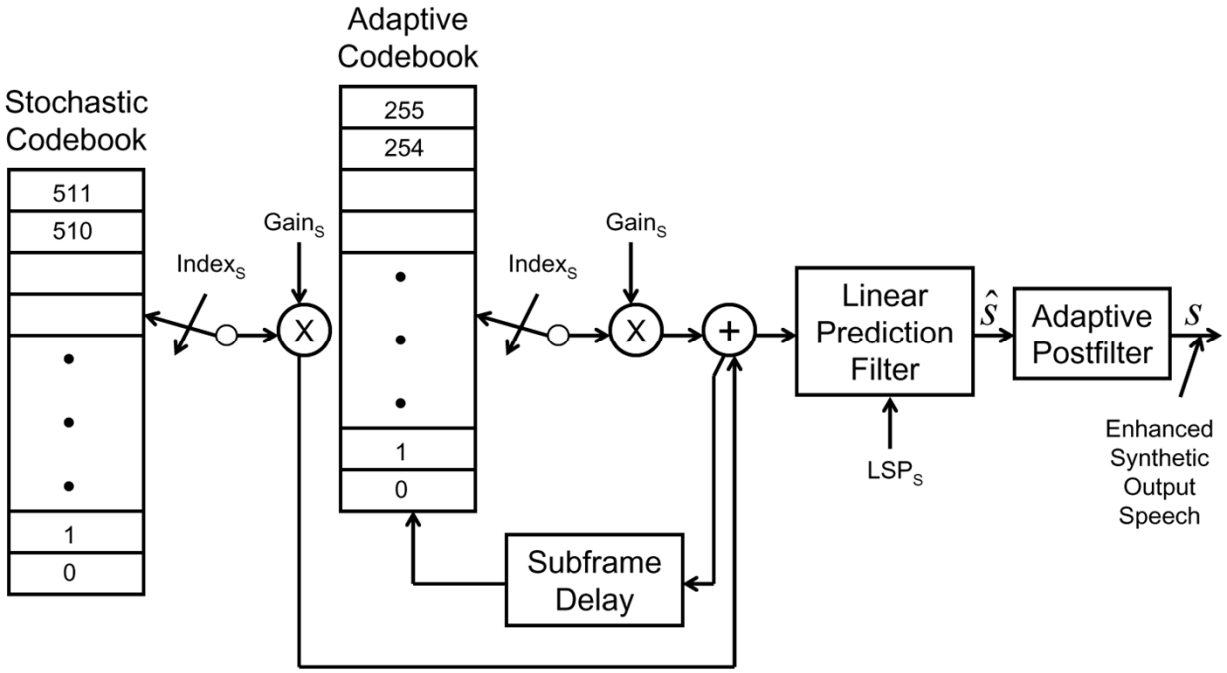
CELP Speech Spectra



CELP Coder at 4800 bps

FS-1016 Encoder/Decoder





FS-1016 Features

- Encoder uses a shochastic codebook with 512 codewords and an adaptive codebook with 256 codewords to estimate the long-term correlation (the pitch period)
- Each codeword in the stochastic codebook is sparsely populated with ternary valued samples (-1, 0, +1) with codewords overlapped and shifted by 2 samples, thereby enabling a fast convolution solution for selection of the optimum codeword for each frame of speech
- LPC analyzer uses a frame size of 30 msec and an LPC predictor of order $p=10$ using the autocorrelation method with a Hamming window
- The 30 msec frame is broken into 4 sub-frames and the adaptive and stochastic codewords are updated every sub-frame, whereas the LPC analysis is only performed once every full frame

FS-1016 Features

- Three sets of features are produced by the encoding system, namely:
  1. the LPC spectral parameters (coded as a set of 10 LSP parameters) for each 30 msec frame
  2. the codeword and gain of the adaptive codebook vector for each 7.5 msec sub-frame
  3. the codeword and gain of the stochastic codebook vector for each 7.5 msec sub-frame

FS-1016 Bit Allocation

Parameter	Subframe				Frame
	1	2	3	4	
LSP1					3
LSP2					4
LSP3					4
LSP4					4
LSP5					4
LSP6					3
LSP7					3
LSP8					3
LSP9					3
LSP10					3
Pitch Delay	8	6	8	6	28
Pitch Gain	5	5	5	5	20
Codeword Index	9	9	9	9	36
Codeword Gain	5	5	5	5	20
Future Expansion					1
Hamming Parity					4
Synchronization					1
Total					144

Table 11.9: Bit allocation for FS-1016 4800 bps CELP coder.

Low Delay CELP

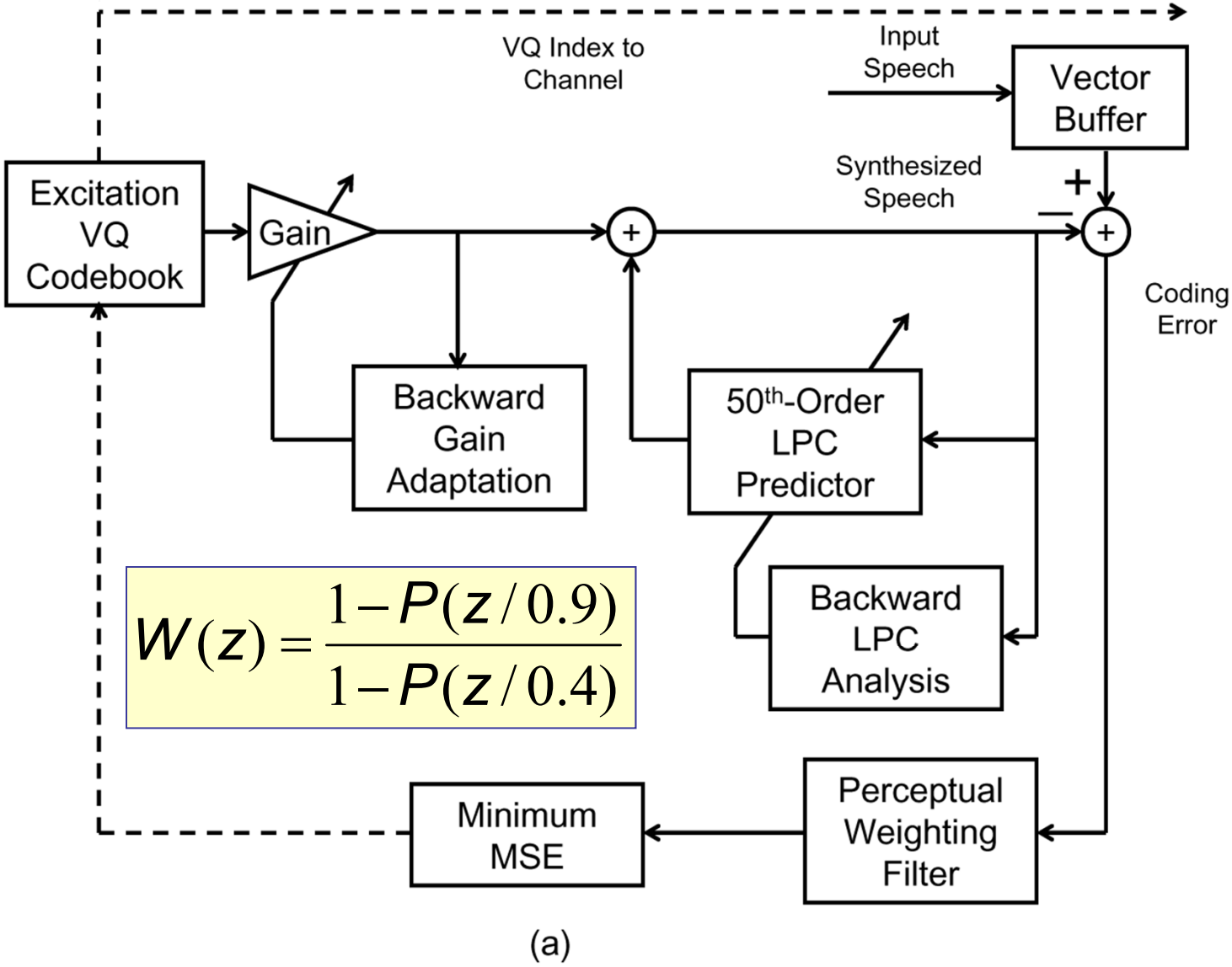
Low Delay CELP Coder

- Total delay of any coder is the time taken by the input speech sample to be processed, transmitted, and decoded, plus any transmission delay, including:
  - buffering delay at the encoder (length of analysis frame window)-~20-40 msec
  - processing delay at the encoder (compute and encode all coder parameters)-~20-40 msec
  - buffering delay at the decoder (collect all parameters for a frame of speech)-~20-40 msec
  - processing delay at the decoder (time to compute a frame of output using the speech synthesis model)-~10-20 msec
- Total delay (exclusive of transmission delay, interleaving of signals, forward error correction, etc.) is order of 70-130 msec

Low Delay CELP Coder

- For many applications, delays are just too large due to forward adaptation for estimating the vocal tract and pitch parameters
  - backward adaptive methods generally produced poor quality speech
  - Chen showed how a backward adaptive CELP coder could be made to perform as well as a conventional forward adaptive coder at bit rates of 8 and 16 kbps

Low Delay (LD) CELP Coder



(a)

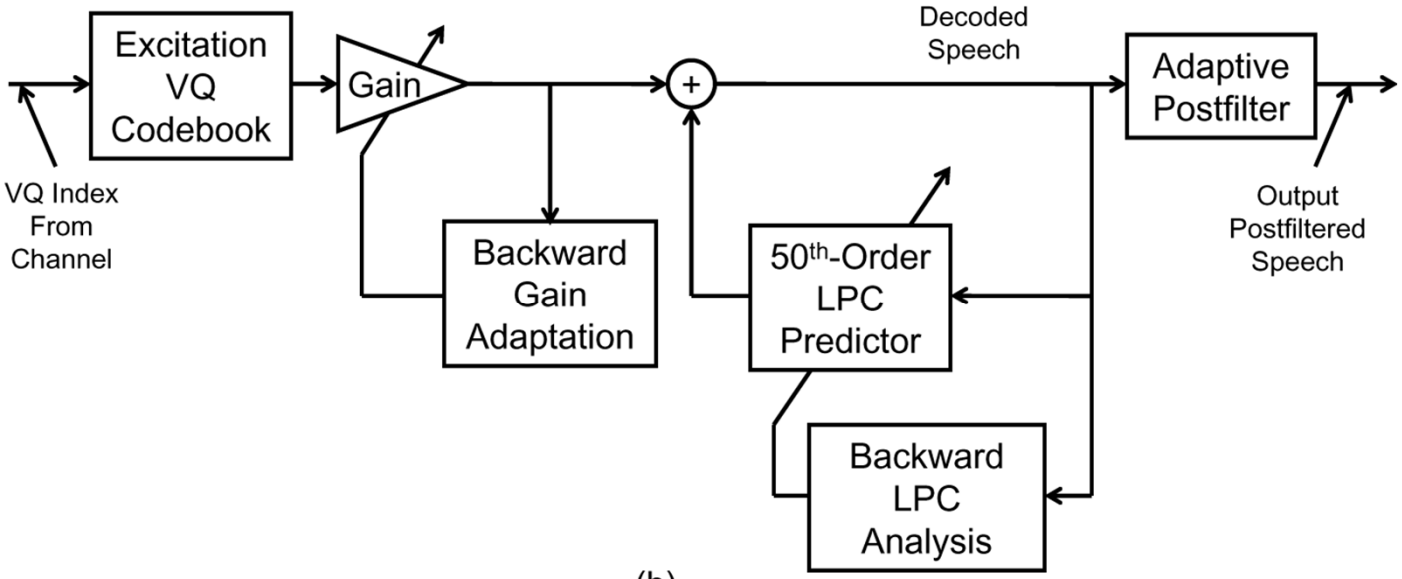
Key Features of LD-CELP

- only the excitation sequence is transmitted to the receiver; the long and short-term predictors are combined into one 50^{th} order predictor whose coefficients are updated by performing LPC analysis on the previously quantized speech signal
- the excitation gain is updated by using the gain information embedded in the previously quantized excitation
- the LD-CELP excitation signal, at 16 kbps, uses 2 bits/sample at an 8 kHz rate; using a codeword length of 5 samples, each excitation vector is coded using a 10-bit codebook (3-bit gain codebook and a 7-bit shape codebook)
- a closed loop optimization procedure is used to populate the shape codebook using the same weighted error criterion as is used to select the best codeword in the CELP coder

16 kbps LD CELP Characteristics

- 8 kHz sampling rate
  - 2 bits/sample for coding residual
- 5 samples per frame are encoded by VQ using a 10-bit “gain-shape” codebook
  - 3 bits (2 bits and sign) for gain (backward adaptive on synthetic speech)
  - 7 bits for wave shape
- recursive autocorrelation method used to compute autocorrelation values from past synthetic speech.
- 50^{th}-order predictor captures pitch of female voice

LD-CELP Decoder



(b)

- all predictor and gain values are derived from coded speech as at the encoder
- post filter improves perceived quality:

$$H_p(z) = K\frac{1-P_{10}(\alpha_1^{-1}z)}{1-P_{10}(\alpha_2^{-1}z)}(1+bz^{-M})(1+\alpha_3^{-1}k_1z^{-1})$$

Lots of CELP Variations

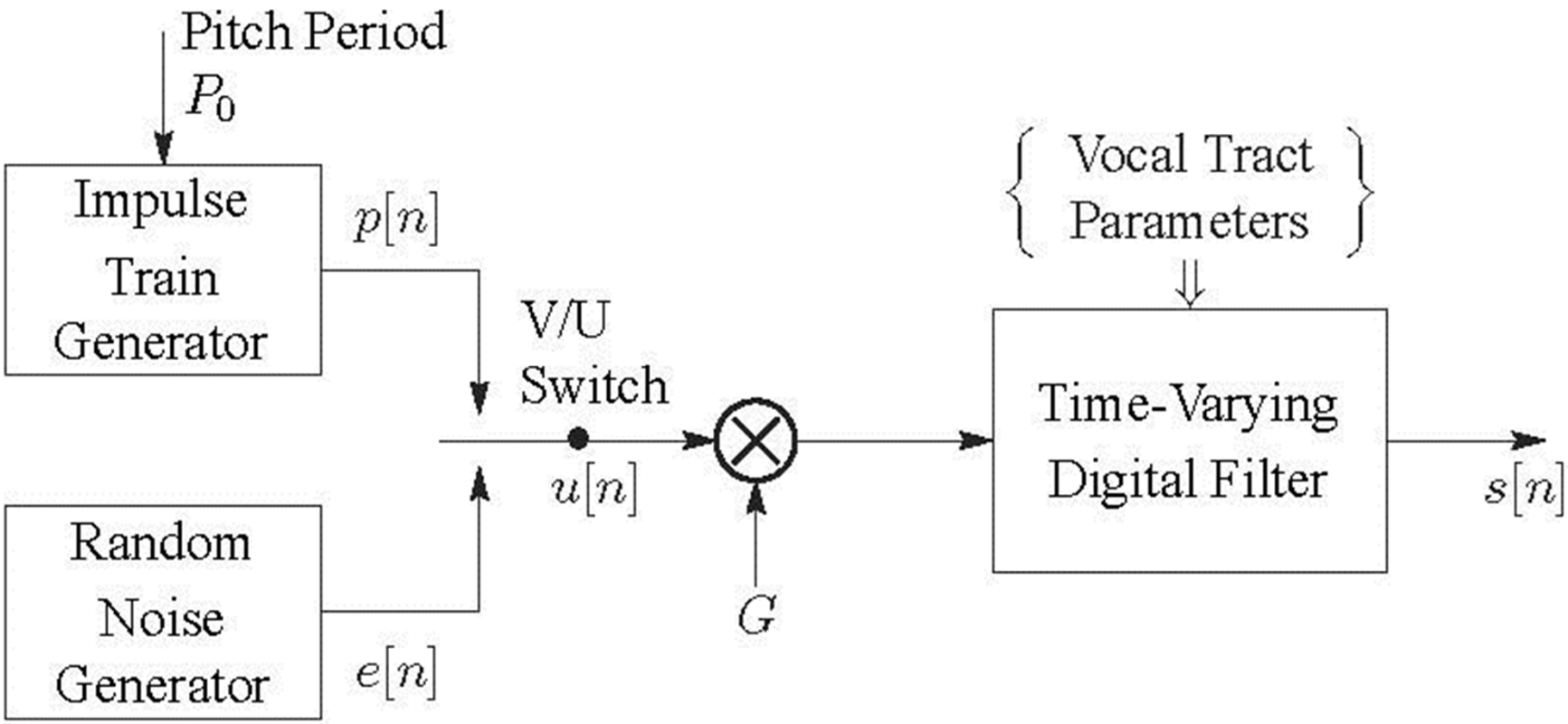
- **ACELP**：Algebraic Code Excited Linear Prediction
- **CS-ACELP**：Conjugate-Structure ACELP
- **VSELP**：Vector-Sum Excited Linear Predictive coding
- **EVSELP**：Enhanced VSELP
- **PSI-CELP**：Pitch Synchronous Innovation-Code Excited Linear Prediction
- **RPE-LTP**：Regular Pulse Exciting-Long Term Prediction-linear predictive coder
- **MP-MLQ :** Multipulse-Maximum Likelihood Quantization

Summary of ABS Speech Coding

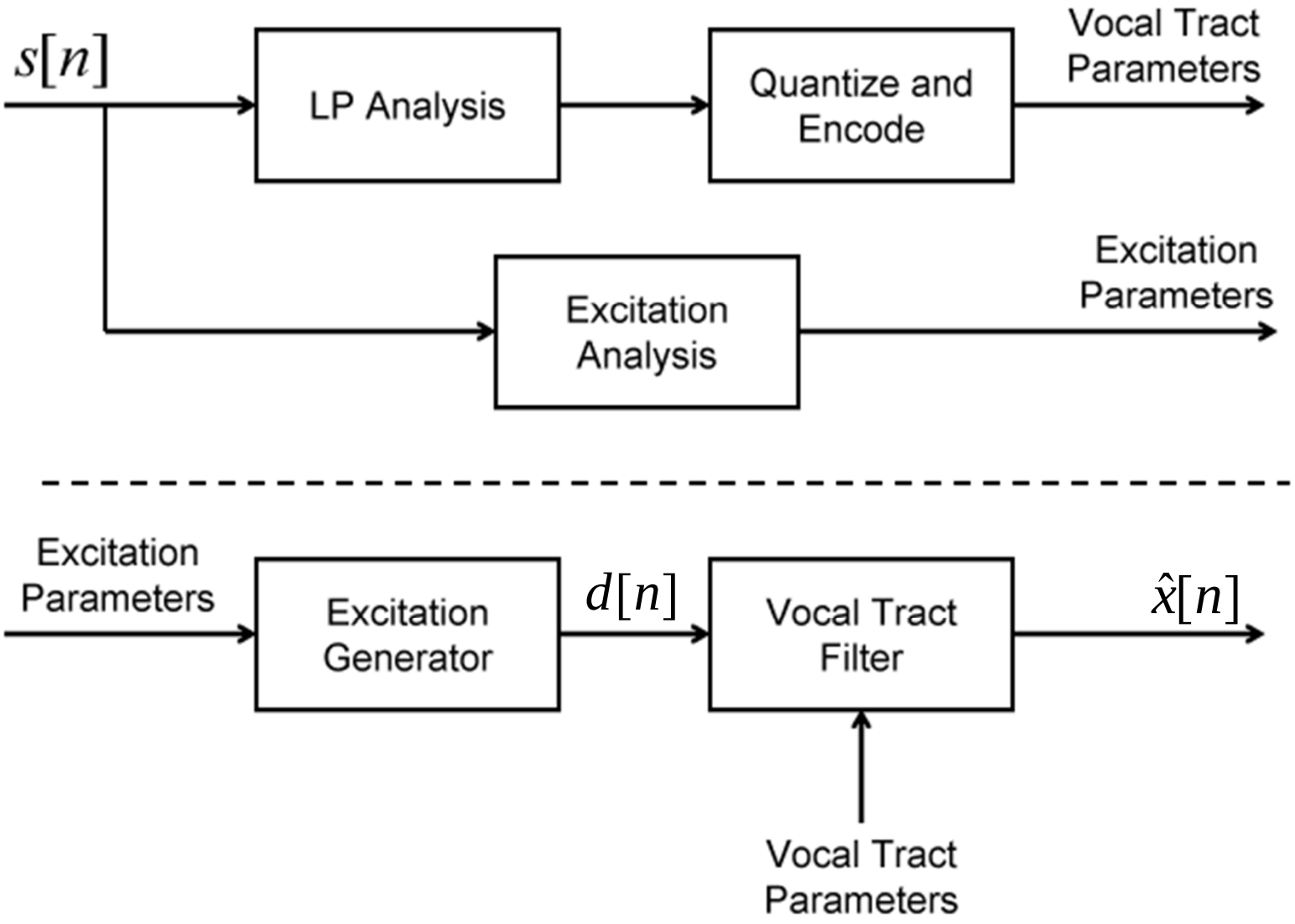
- analysis-by-synthesis methods can be used to derive an excitation signal that produces very good synthetic speech while being efficient to code
  - multipulse LPC
  - code-excited LPC
  - many speech coding standards are based on the CELP idea

Open-Loop Speech Coders

Two-State Excitation Model



Using LP in Speech Coding



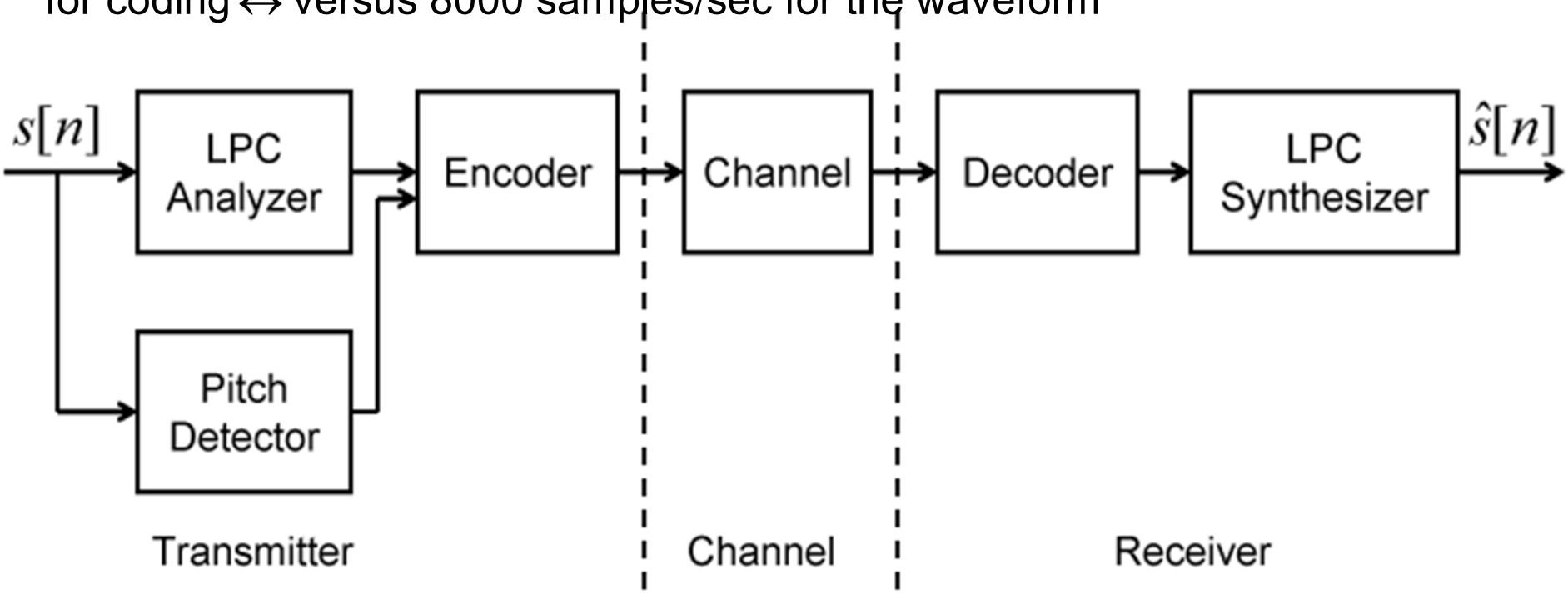
Model-Based Coding

- assume we model the vocal tract transfer function as

$$H(z) = \frac{X(z)}{S(z)} = \frac{G}{A(z)} = \frac{G}{1 - P(z)}$$

$$P(z) = \sum_{k=1}^{p} a_k z^{-k}$$

- LPC coder => 100 frames/sec, 13 parameters/frame (p = 10 LPC coefficients, pitch period, voicing decision, gain) => 1300 parameters/second for coding $\leftrightarrow$ versus 8000 samples/sec for the waveform



LPC Parameter Quantization

- don't use predictor coefficients (large dynamic range, can become unstable when quantized) => use LPC poles, PARCOR coefficients, etc.
- code LP parameters optimally using estimated pdf's for each parameter

1. V/UV-1 bit	100 bps
2. Pitch Period-6 bits (uniform)	600 bps
3. Gain-5 bits (non-uniform)	500 bps
4. LPC poles-10 bits (non-uniform)-5 bits for BW and 5 bits for CF of each of 6 poles	6000 bps
Total required bit rate	7200 bps

S5-original

S5-synthetic

- **no loss in quality from uncoded synthesis (but there is a loss from original speech quality)**
- **quality limited by simple impulse/noise excitation model**

LPC Coding Refinements

1. log coding of pitch period and gain
2. use of PARCOR coefficients ($|k_i|<1$) => log area ratios $g_i=\log(A_{i+1}/A_i)$—almost uniform pdf with small spectral sensitivity => 5-6 bits for coding

- can achieve 4800 bps with almost same quality as 7200 bps system above
- can achieve 2400 bps with 20 msec frames => 50 frames/sec

LPC-10 Vocoder

LPC-10 Vocoder

- U.S. Government standard
  - covariance LP analysis (10th-order)
  - AMDF pitch detector (see Chapter 4)
- Bit rate

Frame rate = 44.44 frames/sec

param.	$k_1 - k_4$	$k_5 - k_8$	k_9	k_{10}	pitch	ampl.	sync.	Total
# bits	5 ea.	4 ea.	3	2	7	5	1	54

Bit rate = 2400 bits/sec

LPC-Based Speech Coders

- the key problems with speech coders based on all-pole linear prediction models
  - inadequacy of the basic source/filter speech production model
  - idealization of source as either pulse train or random noise
  - lack of accounting for parameter correlation using a one-dimensional scalar quantization method => aided greatly by using VQ methods

VQ-Based LPC Coder

- train VQ codebooks on PARCOR coefficients

***bottom line*: to dramatically improve quality need improved excitation model**

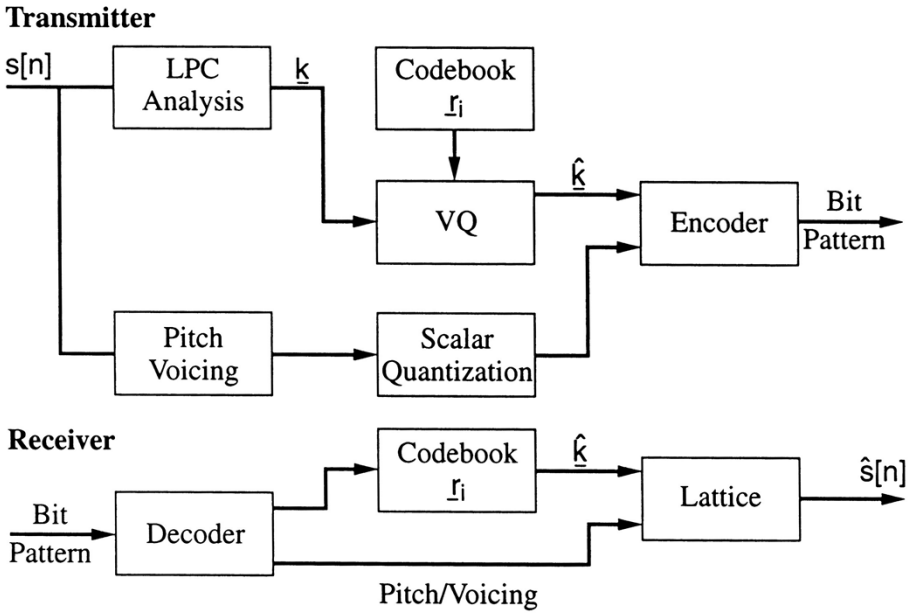


Figure 12.22 A VQ LPC vocoder. In this example, the synthesizer uses an all-pole lattice structure.

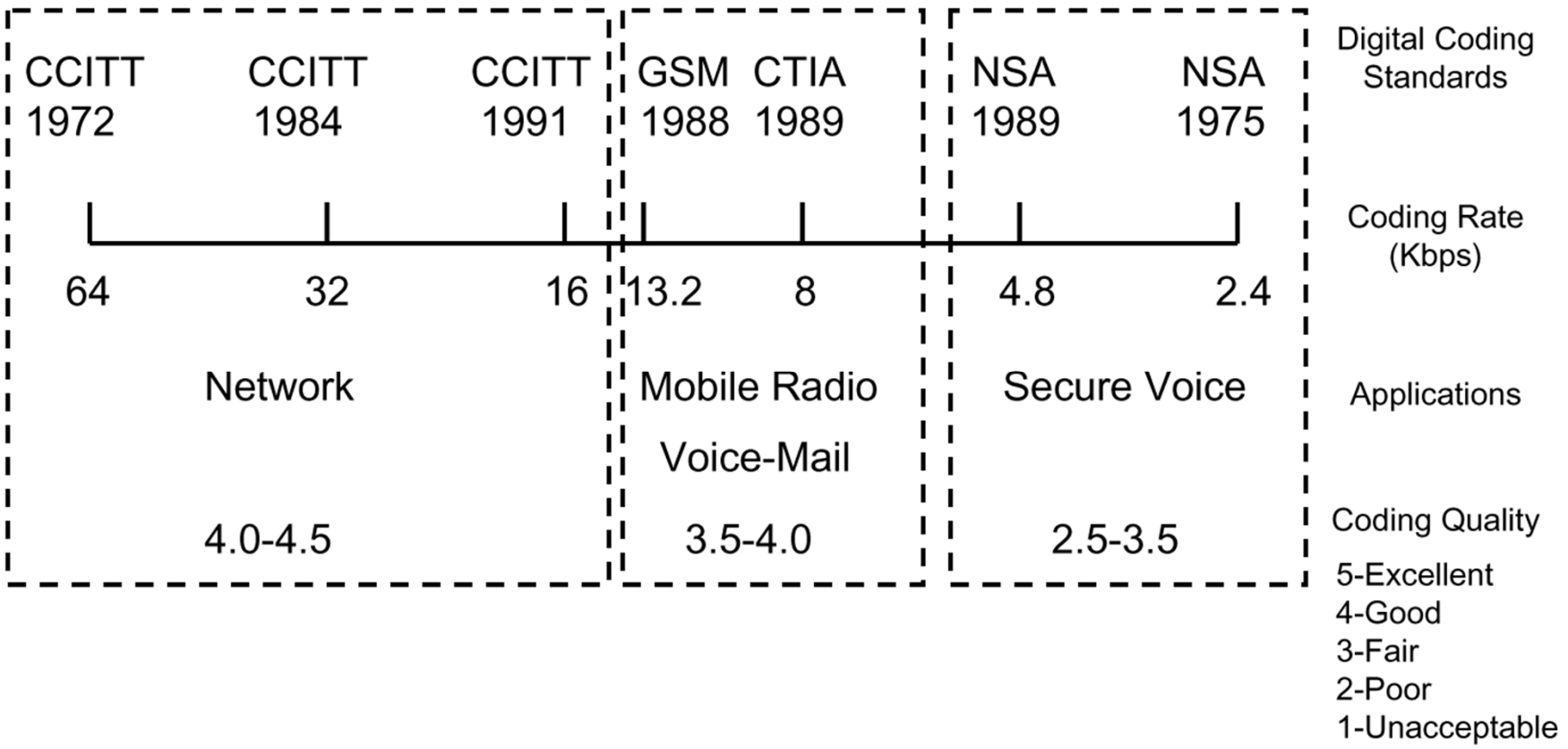
- Case 1: same quality as 2400 bps LPC vocoder
  - 10-bit codebook of PARCOR vectors
  - 44.4 frames/sec
  - 8-bits for pitch, voicing, gain
  - 2-bit for frame synchronization
- total bit rate of 800 bps

- Case 2: same bit rate, higher quality
  - 22 bit codebook => 4.2 million codewords to be searched
  - never achieved good quality due to computation, storage, graininess of quantization at cell boundaries

Applications of Speech Coders

- network-64 Kbps PCM (8 kHz sampling rate, 8-bit log quantization)
- international-32 Kbps ADPCM
- teleconferencing-16 Kbps LD-CELP
- wireless-13, 8, 6.7, 4 Kbps CELP-based coders
- secure telephony-4.8, 2.4 Kbps LPC-based coders (MELP)
- VoIP-8 Kbps CELP-based coder
- storage for voice mail, answering machines, announcements-16 Kbps LC-CELP

Applications of Speech Coders



Speech Coder Attributes

- bit rate-2400 to 128,000 bps
- quality-subjective (MOS), objective (*SNR*, intelligibility)
- complexity-memory, processor
- delay-echo, reverberation; block coding delay, processing delay, multiplexing delay, transmission delay-~100 msec
- telephone bandwidth-200-3200 Hz, 8kHz sampling rate
- wideband speech-50-7000 Hz, 16 kHz sampling rate

Network Speech Coding Standards

Coder	*Type*	*Rate*	*Usage*
G.711	companded PCM	64 Kbps	toll
G.726/727	ADPCM	16-40 Kbps	toll
G.722	SBC/ADPCM	48, 56,64 Kbps	wideband
G.728	LD-CELP	16 Kbps	toll
G.729A	CS-ACELP	8 Kbps	toll
G.723.1	MPC-MLQ & ACELP	6.3/5.3 Kbps	toll

Cellular Speech Coding Standards

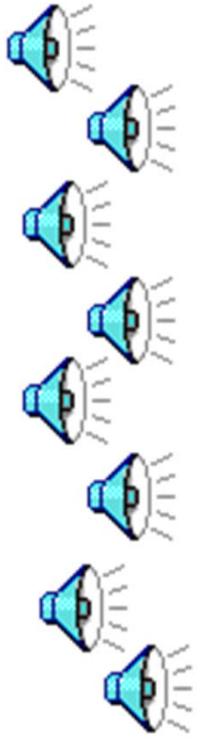
Coder	*Type*	*Rate*	*Usage*
GSM	RPE-LTP	13 Kbps	<toll
GSM ½ rate	VSELP	5.6 Kbps	GSM
IS-54	VSELP	7.95 Kbps	GSM
IS-96	CELP	0.8-8.5 Kbps	<GSM
PDC	VSELP	6.7 Kbps	<GSM
PDC ½ rate	PSI-CELP	3.45 Kbps	PDC

Secure Telephony Speech Coding Standards

Coder	*Type*	*Rate*	*Usage*
FS-1015	LPC	2.4 Kbps	high DRT
FS-1016	CELP	4.8 Kbps	<IS-54
?	model-based	2.4 Kbps	>FS-1016

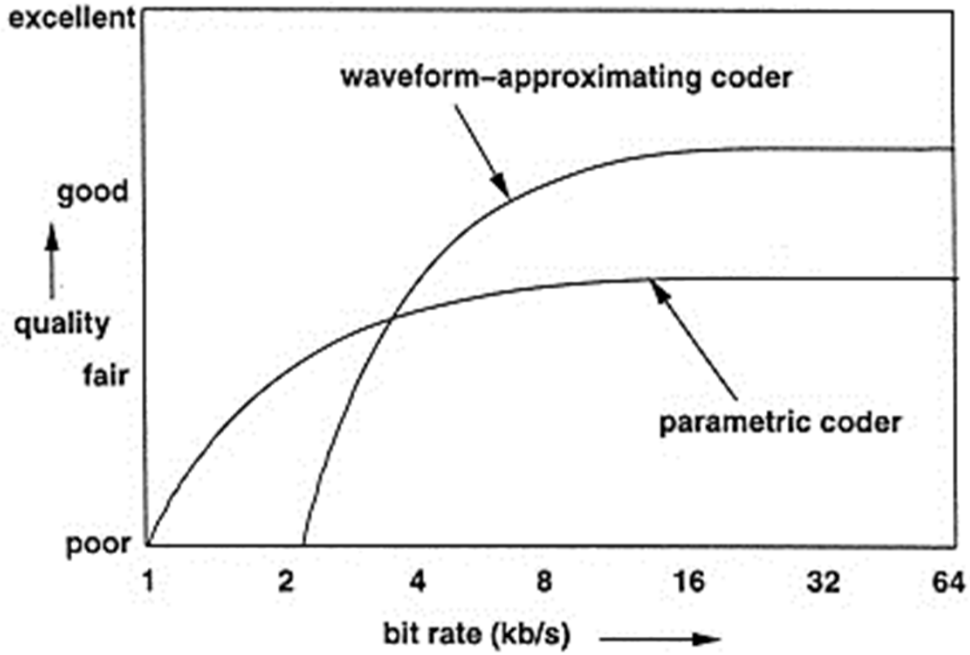
Demo: Coders at Different Rates

G.711	64 kb/s
G.726 ADPCM	32 kb/s
G.728 LD-CELP	16 kb/s
G.729 CS-ACELP	8 kb/s
G.723.1 MP-MLQ	6.3 kb/s
G.723.1 ACELP	5.3 kb/s
RCR PSI-CELP	3.45 kb/s
NSA 1998 MELP	2.4 kb/s



Speech Coding Quality Evaluation

- 2 types of coders
  - waveform approximating-PCM, DPCM, ADPCM-coders which produce a reconstructed signal which converges toward the original signal with decreasing quantization error
  - parametric coders (model-based)-SBC, MP-LPC, LPC. MB-LPC, CELP-coders which produce a reconstructed signal which does not converge to the original signal with decreasing quantization error



• waveform coder converges to quality of original speech

• parametric coder converges to model-constrained maximum quality (due to the model inaccuracy of representing speech)

Factors on Speech Coding Quality

- **talker and language dependency** - especially for parametric coders that estimate pitch that is highly variable across men, women and children; language dependency related to sounds of the language (e.g., clicks) that are not well reproduced by model-based coders
- **signal levels** - most waveform coders designed for speech levels normalized to a maximum level; when actual samples are lower than this level, the coder is not operating at full efficiency causing loss of quality
- **background noise** - including babble, car and street noise, music and interfering talkers; levels of background noise varies, making optimal coding based on clean speech problematic
- **multiple encodings** - tandem encodings in a multi-link communication system, teleconferencing with multiple encoders
- **channel errors** - especially an issue for cellular communications; errors either random or bursty (fades)-redundancy methods oftern used
- **non-speech sounds** - e.g., music on hold, dtmf tones; sounds that are poorly coded by the system

Measures of Speech Coder Quality

$$SNR = 10\log_{10}\frac{\sum_{n=0}^{N-1}\left[s[n]\right]^2}{\sum_{n=0}^{N-1}\left[s[n]-\hat{s}[n]\right]^2}, \quad \text{over whole signal}$$

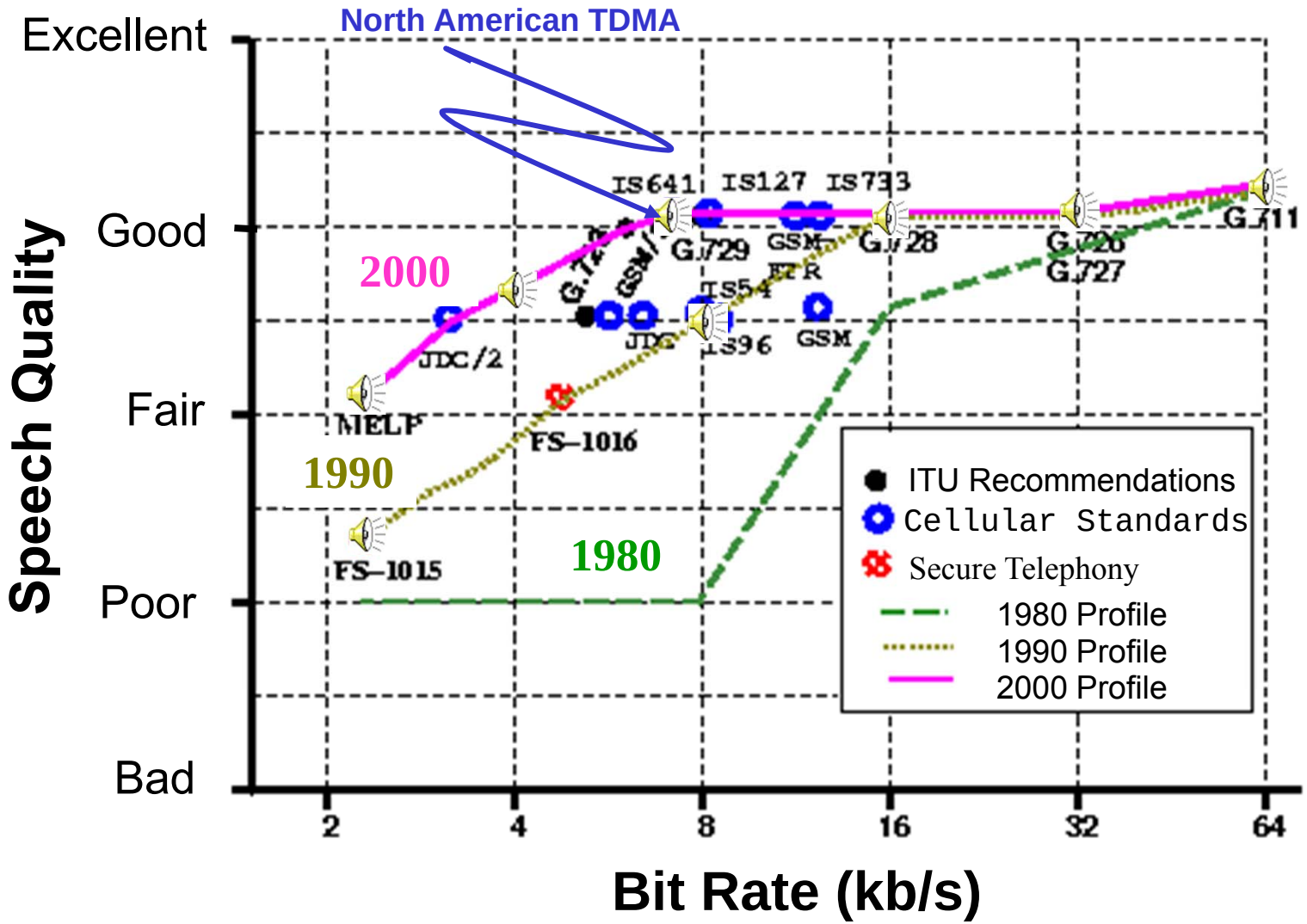
$$SNR_{seg} = \frac{1}{K}\sum_{k=1}^{K} SNR_k \quad \text{over frames of 10-20 msec}$$

- good primarily for waveform coders

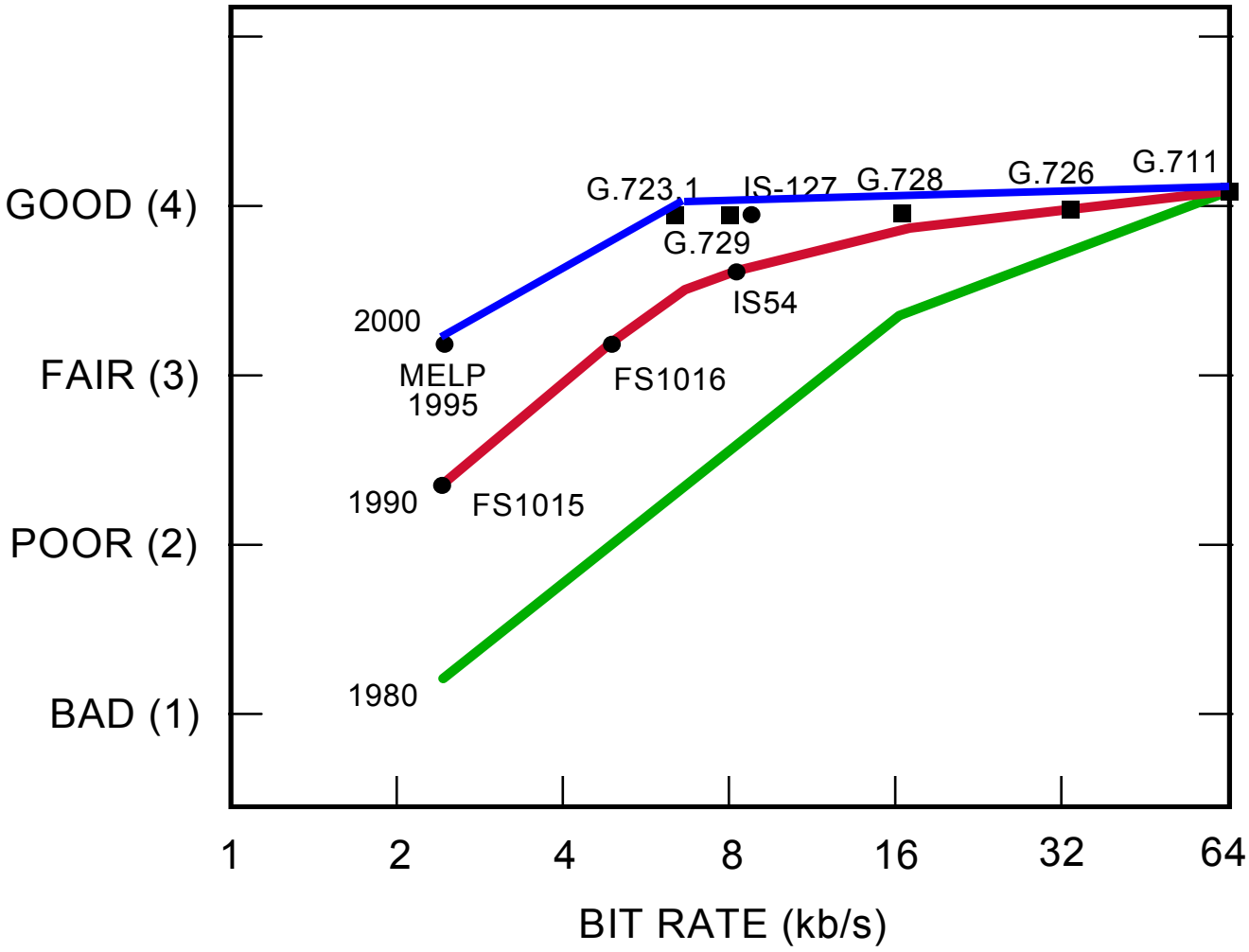
Measures of Speech Coder Quality

- Intelligibility-Diagnostic Rhyme Test (DRT)
  - compare words that differ in leading consonant
  - identify spoken word as one of a pair of choices
  - high scores (~90%) obtained for all coders above 4 Kbps
- Subjective Quality-Mean Opinion Score (MOS)
  - 5 excellent quality
  - 4 good quality
  - 3 fair quality
  - 2 poor quality
  - 1 bad quality
- MOS scores for high quality wideband speech (~4.5) and for high quality telephone bandwidth speech (~4.1)

Evolution of Speech Coder Performance



Speech Coder Subjective Quality



Speech Coder Demos

Telephone Bandwidth Speech Coders

- 64 kbps Mu-Law PCM
- 32 kbps CCITT G.721 ADPCM
- 16 kbps LD-CELP
- 8 kbps CELP
- 4.8 kbps CELP for STU-3
- 2.4 kbps LPC-10E for STU-3

Wideband Speech Coder Demos

Wideband Speech Coding

- Male talker
  - 3.2 kHz-uncoded
  - 7 kHz-uncoded
  - 7 kHz-coded at 64 kbps (G.722)
  - 7 kHz-coded at 32 kbps (LD-CELP)
  - 7 kHz-coded at 16 kbps (BE-CELP)
- Female talker
  - 3.2 kHz-uncoded
  - 7 kHz-uncoded
  - 7 kHz-coded at 64 kbps (G.722)
  - 7 kHz-coded at 32 kbps (LD-CELP)
  - 7 kHz-coded at 16 kbps (BE-CELP)