

Adaptive Quantization

$$SNR(dB) = 10 \log_{10} \left[\frac{\sigma_x^2}{\sigma_e^2} \right] = 6B + 4.77 - 20 \log_{10} \left[\frac{X_{\max}}{\sigma_x} \right]$$

Adaptive Quantization

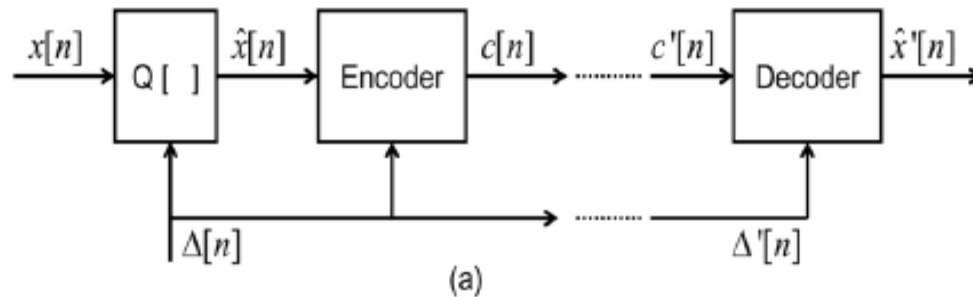
$$\Delta = \frac{2X_{\max}}{2^B}$$

- Uniform quantization => SNR depends on σ_x being constant (this is clearly not the case)
- instantaneous companding => SNR only weakly dependent on X_{\max}/σ_x for large μ -law compression (100- 500)
- **Quantization dilemma**: want to choose quantization step size large enough to accomodate maximum peak-to-peak range of $x[n]$; at the same time need to make the quantization step size small so as to minimize the quantization error
 - the non-stationary nature of speech (variability across sounds, speakers, backgrounds) compounds this problem greatly

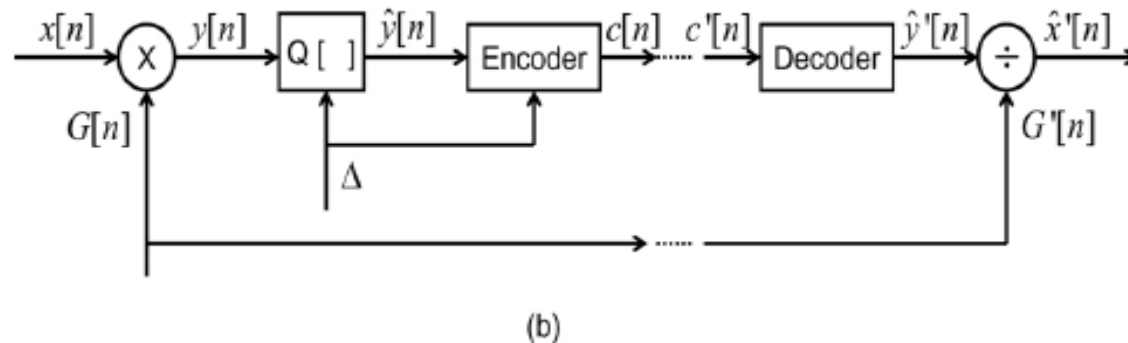
$$SNR(dB) = 6B + 4.77 - 20 \log_{10} [\ln(1 + \mu)] - 10 \log_{10} \left[1 + \left(\frac{X_{\max}}{\mu \sigma_x} \right)^2 + \sqrt{2} \left(\frac{X_{\max}}{\mu \sigma_x} \right) \right]$$

Solutions to Quantization Dilemma

- **Solution 1** – let Δ vary to match the variance of the input signal $\Rightarrow \Delta[n]$
 - $\Delta[n]$ proportional to $\sigma_x \Rightarrow$ quantization levels and ranges would be linearly scaled to match $\sigma_x^2 \Rightarrow$ need to reliably estimate σ_x^2



- **Solution 2** - use a variable gain, $G[n]$, followed by a fixed quantizer step size, $\Delta \Rightarrow$ keep signal variance of $y[n]=G[n]x[n]$ constant
 - $G[n]$ proportional to $1/\sigma_x$ to give $\sigma_y^2 \approx \text{constant}$



Differential Quantization

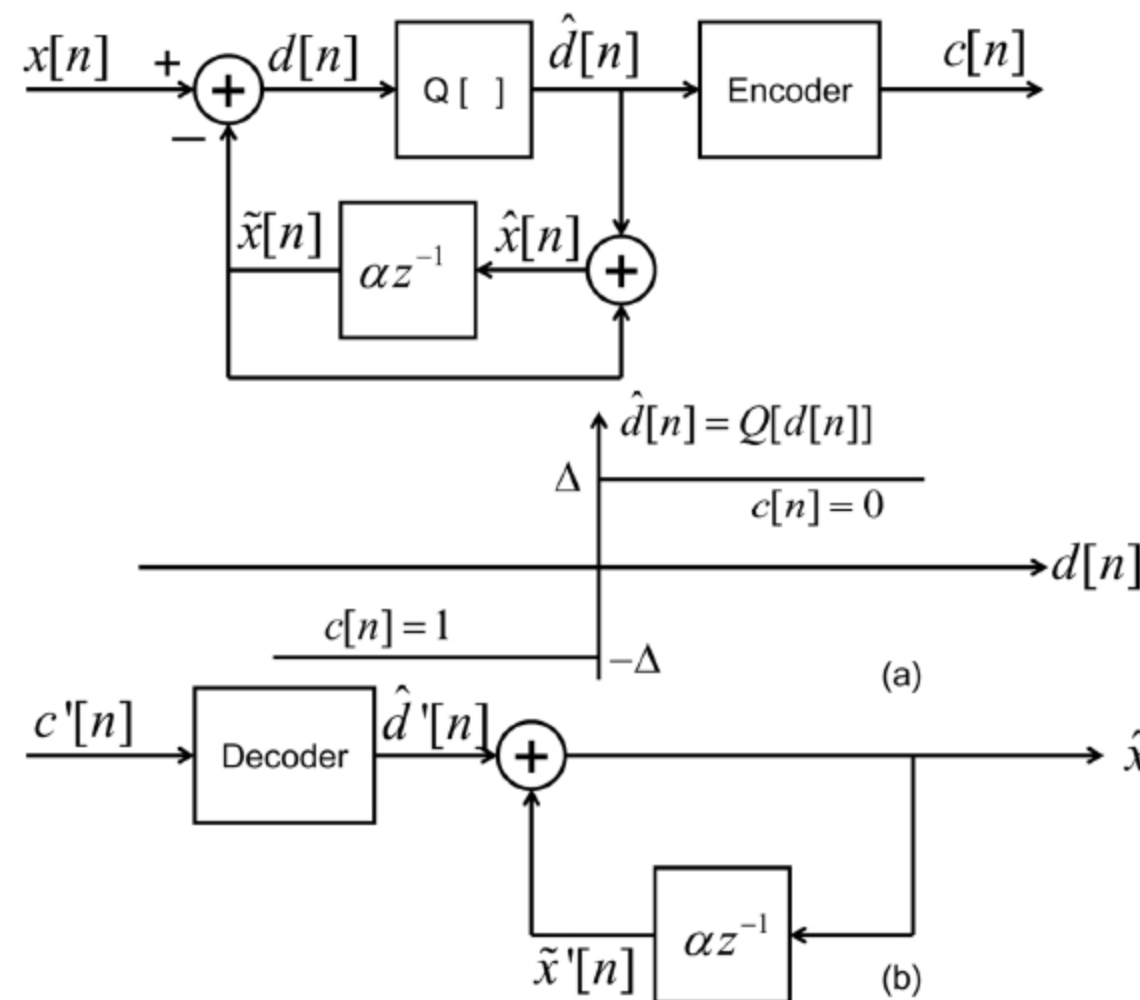
Differential Quantization

- we have carried instantaneous quantization of $x[n]$ as far as possible
- consider correlations between speech samples separated in time => differential quantization(差分量化)
- high correlation values => signal does not change rapidly in time => difference between adjacent samples should have lower variance than the signal itself
- differential quantization can increase SNR at a given bit rate, or lower bit rate for a given SNR

Delta Modulation

- simplest form of differential quantization is in **delta modulation** (DM) 增量调制
- sampling rate chosen to be many times the Nyquist rate for the input signal => adjacent samples are highly correlated
- this leads to a high ability to predict $x[n]$ from past samples, with the variance of the prediction error being very low,
=> can use simple 1-bit (2-level) quantizer
=> the bit rate for DM systems is just the (high) sampling rate of the signal

Linear Delta Modulation(LDM)



- 2-level quantizer with fixed step size, Δ , with quantizer form

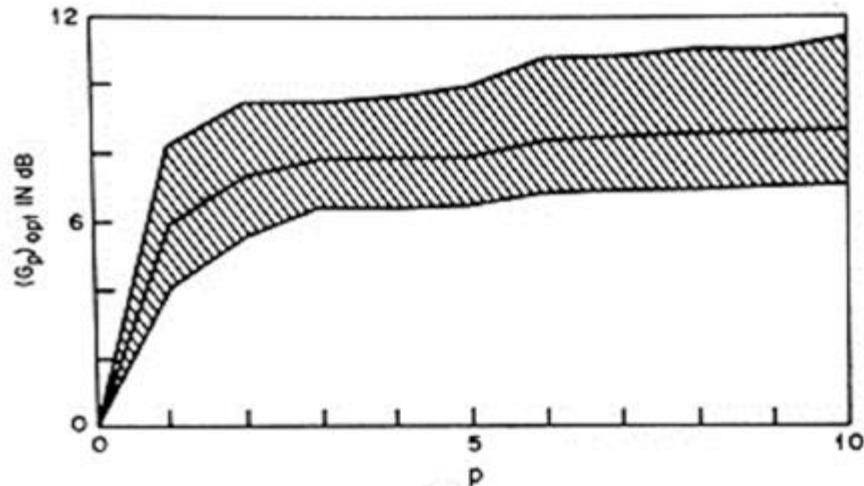
$$\begin{aligned}\hat{d}[n] &= \Delta \quad \text{if } d[n] > 0 \quad (c[n] = 0) \\ &= -\Delta \quad \text{if } d[n] < 0 \quad (c[n] = 1)\end{aligned}$$

- using simple first order predictor
- basic equations of DM are

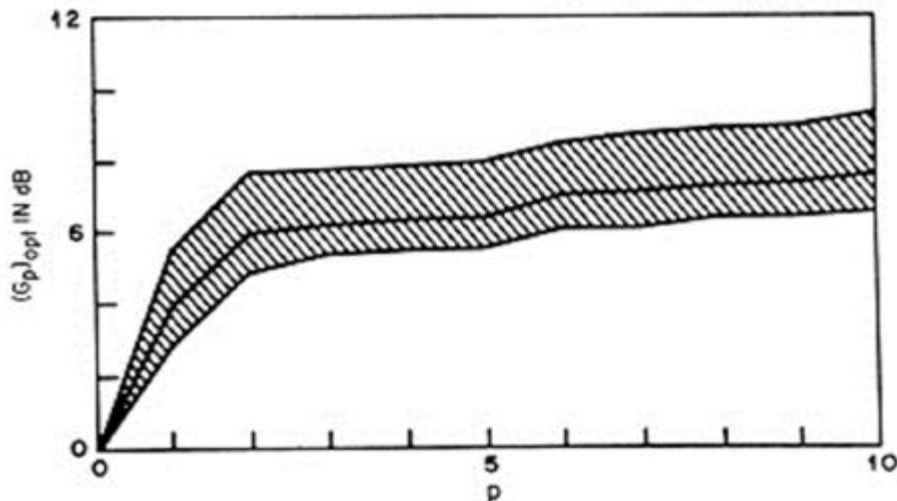
$$\hat{x}[n] = \alpha \hat{x}[n-1] + \hat{d}[n]$$

- since $\hat{x}[n]$ can only increase by fixed increments of Δ , fixed DM is called **linear DM** or **LDM**

Differential PCM (DPCM)



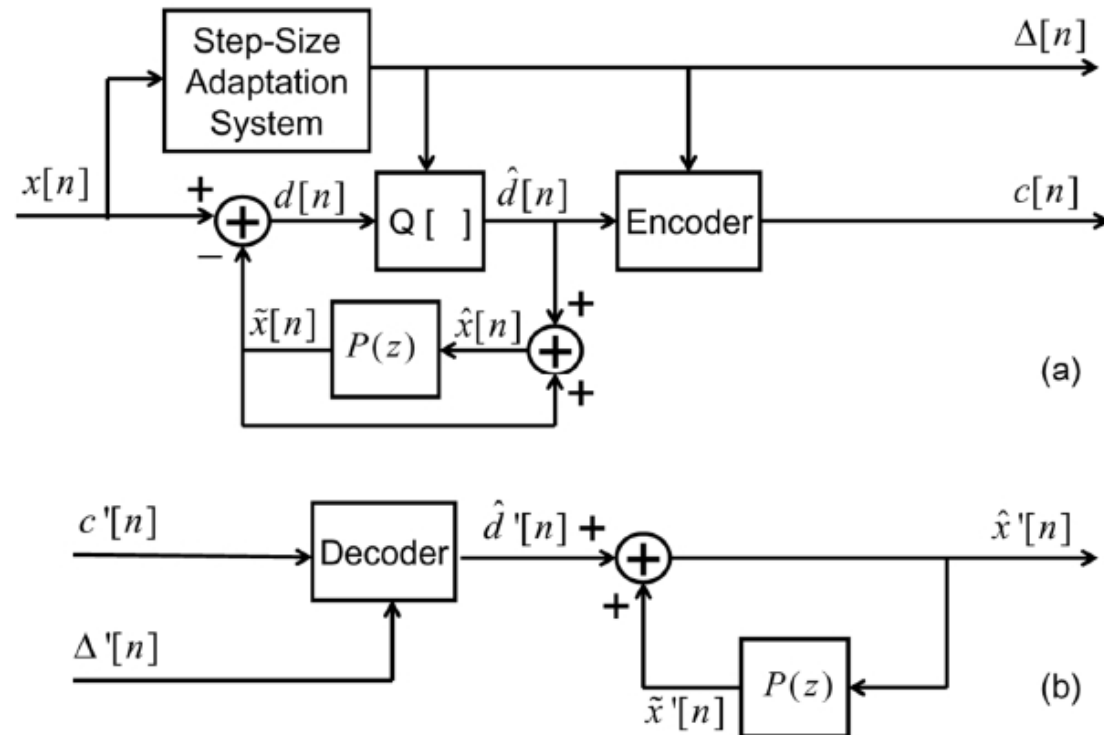
(a) lowpass filtered speech



(b) bandpass filtered speech

- Differential quantization
 - fixed predictors can give from 4-11 dB SNR improvement over direct quantization (PCM)
 - most of the gain occurs with first order predictor
 - prediction up to 4th or 5th order helps => **DPCM**

DPCM with Adaptive Quantization



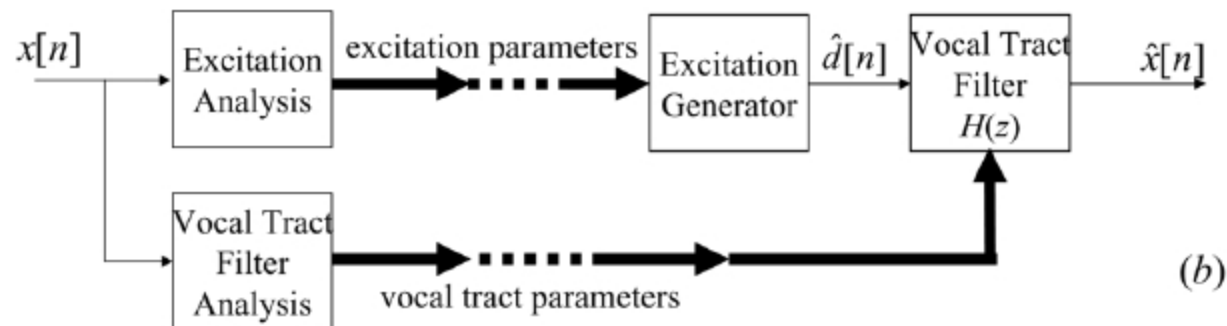
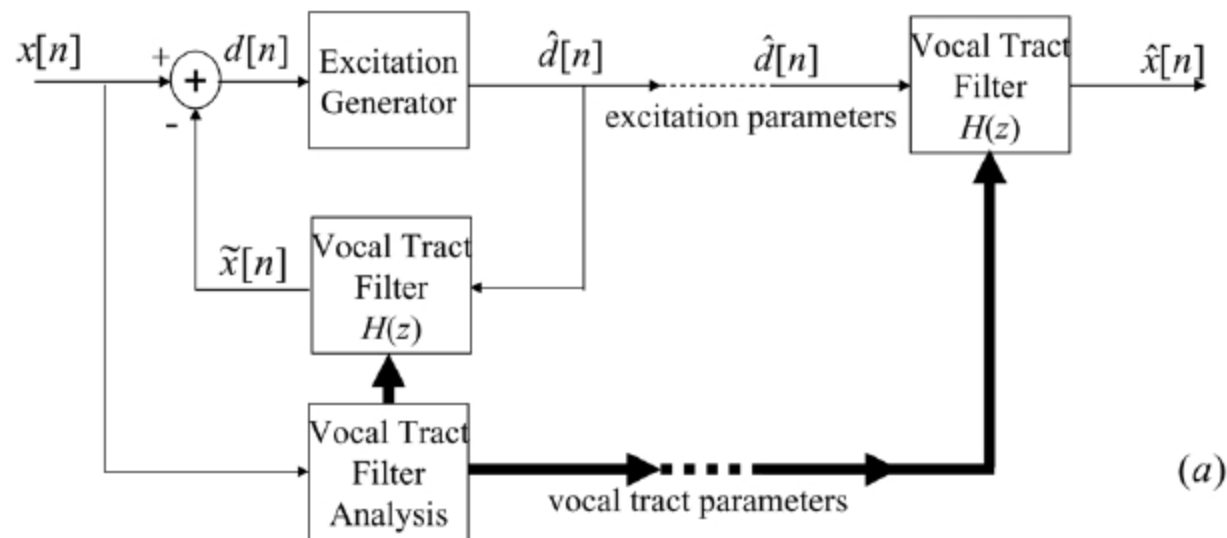
- quantizer step size proportional to variance at quantizer input
- ADPCM is about 10-11 dB SNR better than μ -law non-adaptive PCM
 - get 5 dB improvement in SNR using adaptation procedures
 - get 6 dB improvement in SNR using differential configuration with fixed prediction

Model-Based Speech Coding

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

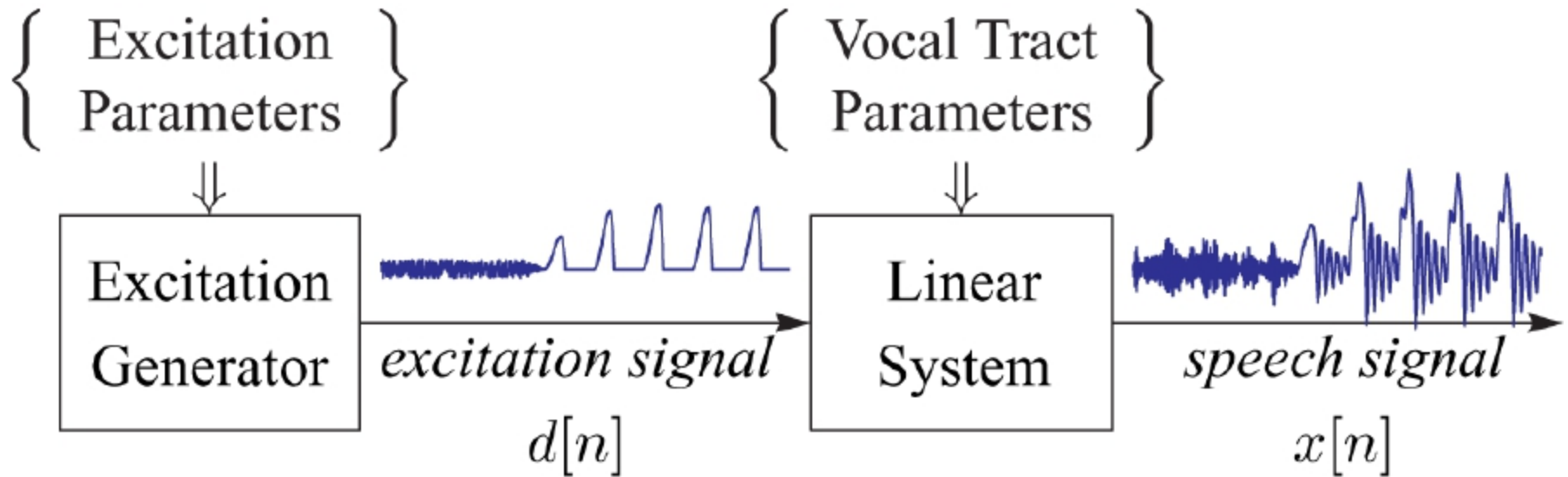
Closed-Loop and Open-Loop Speech Coders



- **Closed-loop** – used in a feedback loop where the synthetic speech output is compared to the input signal, and the resulting difference used to determine the excitation for the vocal tract model ([Analysis-by-Synthesis](#)).
- **Open-loop** – the parameters of the model are estimated directly from the speech signal with no feedback as to the quality of the resulting synthetic speech.

Quantization of Speech Model Parameters

Quantization of Speech Model Parameters



- Excitation and vocal tract (linear system) are characterized by sets of parameters which can be estimated from a speech signal by LP or cepstral processing
- We can use the set of estimated parameters to synthesize an approximation to the speech signal whose quality depends of a range of factors

Quantization of Speech Model Parameters

- Quality and data rate of synthesis depends on:
 - the ability of the model to represent speech
 - the ability to reliably and accurately estimate the parameters of the model
 - the ability to quantize the parameters in order to obtain a low data rate digital representation that will yield a high quality reproduction of the speech signal

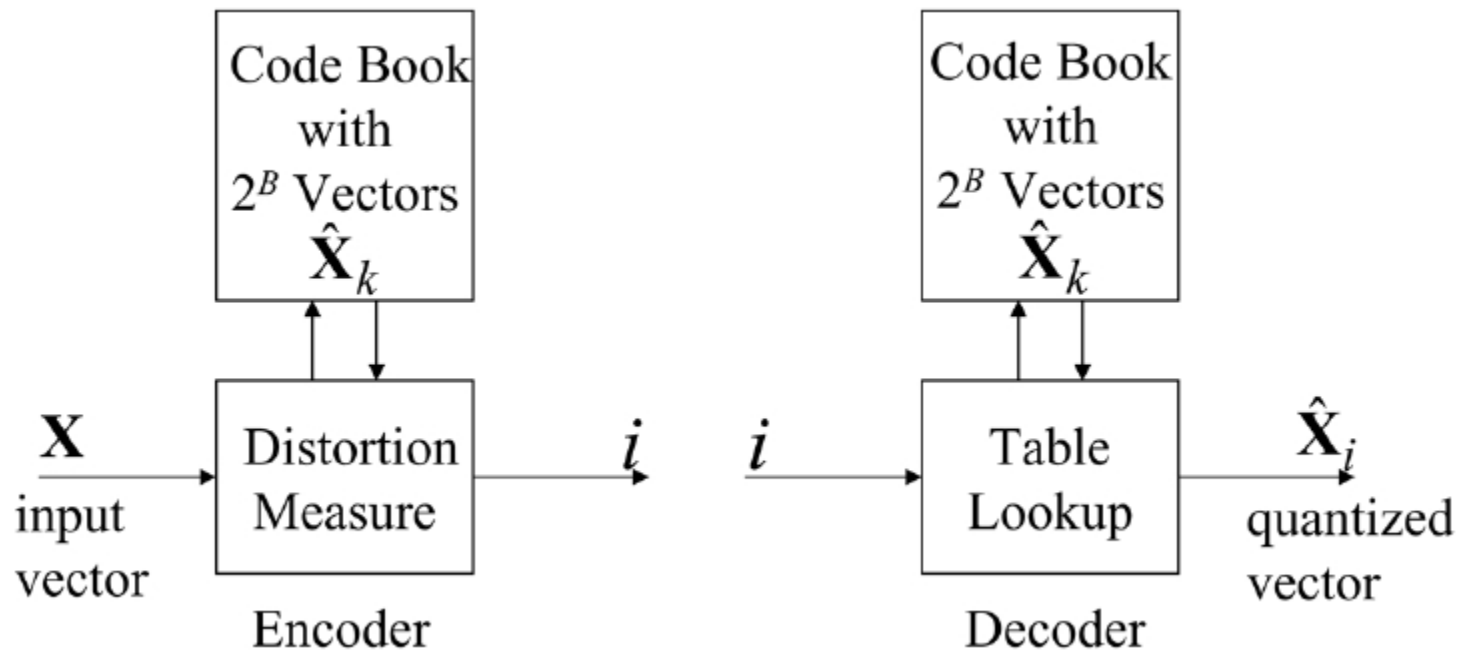
Scalar Quantization 标量量化

- Scalar quantization – treat each model parameter separately and quantize using a fixed number of bits
 - need to measure (estimate) statistics of each parameter, i.e., mean, variance, minimum/maximum value, pdf, etc.
 - each parameter has a different quantizer with a different number of bits allocated
- Example of scalar quantization
 - pitch period typically ranges from 20-150 samples (at 8 kHz sampling rate) => need about 128 values (7-bits) uniformly over the range of pitch periods, including value of zero for unvoiced/background
 - amplitude parameter might be quantized with a μ -law quantizer using 4-5 bits per sample
 - using a frame rate of 100 frames/sec, you would need about 700 bps for pitch period and 400-500 bps for amplitude

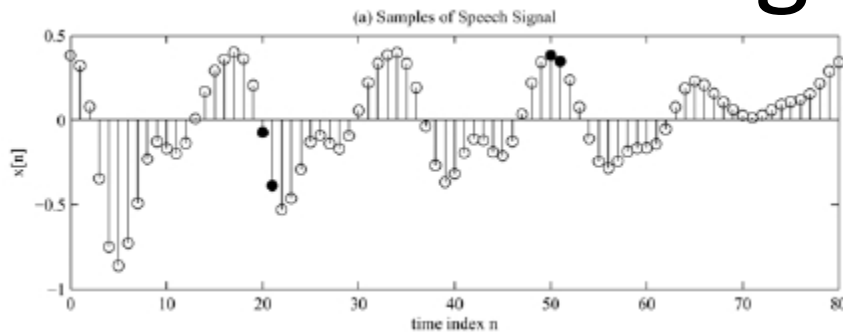
Vector Quantization 矢量量化

- code **block of scalars** as a vector, rather than individually
- design an **optimal quantization** method based on mean-squared distortion metric
- essential for model-based coders
- **VQ works because the scalar components of each vector are correlated**
- if scalar components are independent => VQ offers no advantage over scalar quantization

Vector Quantization



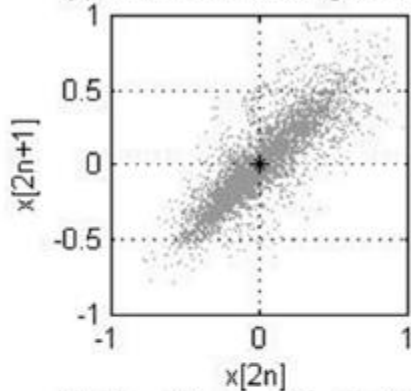
Waveform Coding Vector Quantizer



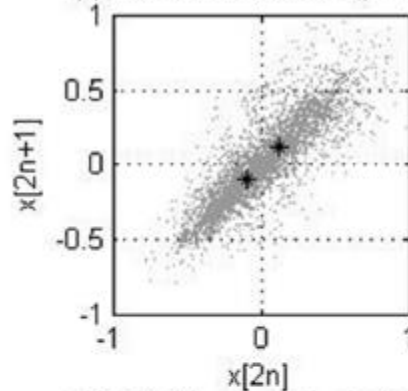
VQ code pairs of waveform samples,

$$X[n] = (x[2n], x[2n+1]);$$

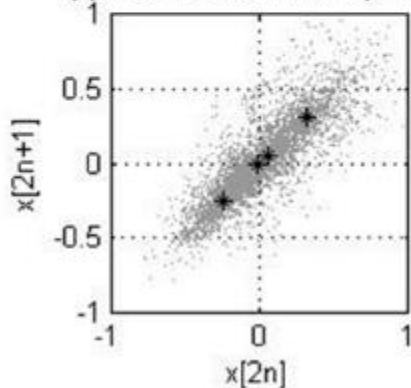
(b) Centroid of Training Vectors



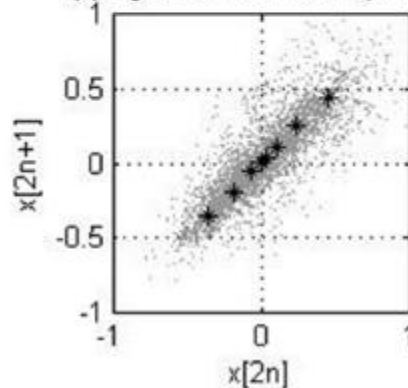
(c) Two Vectors After Splitting



(d) Four Vectors After Splitting



(e) Eight Vectors After Splitting



(b) Single element codebook with cluster centroid (0-bit codebook)

(c) Two element codebook with two cluster centers (1-bit codebook)

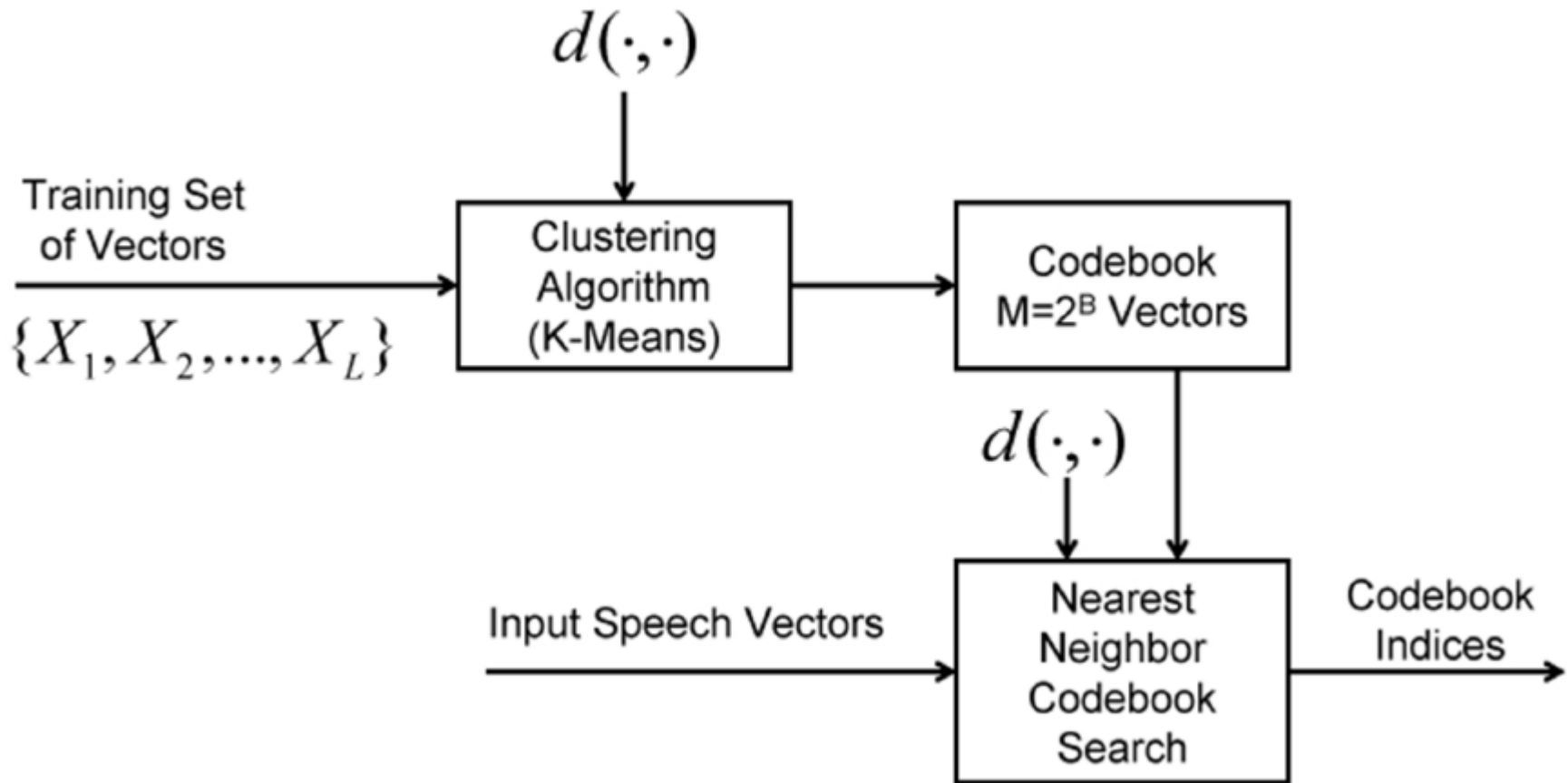
(d) Four element codebook with four cluster centers (2-bit codebook)

(e) Eight element codebook with eight cluster centers (3-bit codebook)

Elements of a VQ Implementation

1. A large training set of analysis vectors; $X=\{X_1, X_2, \dots, X_L\}$, L should be much larger than the size of the codebook, M , i.e., 10-100 times the size of M .
2. A measure of distance, $d_{ij}=d(X_i, X_j)$, between a pair of analysis vectors, both for clustering the training set as well as for classifying test set vectors into unique codebook entries.
3. A centroid computation procedure
4. A classification procedure for arbitrary analysis vectors that chooses the codebook vector closest in distance to the input vector, providing the codebook index of the resulting nearest codebook vector.

VQ Implementation



1. The VQ Training Set

- The VQ training set of $L \geq 10M$ vectors should span the anticipated range of:
 - talkers, ranging in age, accent, gender, speaking rate, speaking levels, etc.
 - speaking conditions, range from quiet rooms, to automobiles, to noisy work places
 - transducers and transmission systems, including a range of microphones, telephone handsets, cellphones, speakerphones, etc.
 - speech, including carefully recorded material, conversational speech, telephone queries, etc.

2. The VQ Distance Measure

- The VQ distance measure depends critically on the nature of the analysis vector, X .
 - If X is a log spectral vector, then a possible distance measure would be an log spectral distance, of the form:

$$d(X_i, X_j) = \left[\sum_{k=1}^R |x_i^k - x_j^k|^p \right]^{1/p}$$

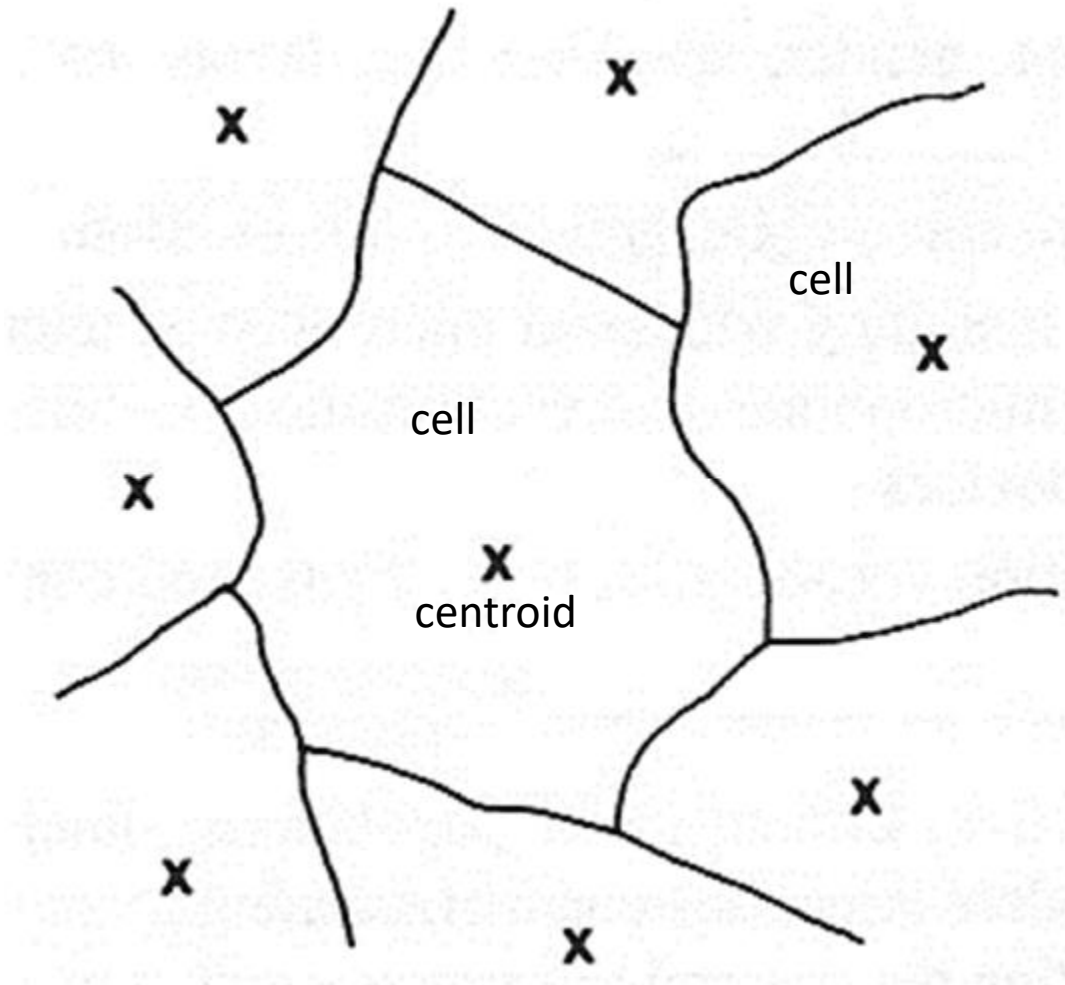
- If X is a cepstral vector, then the distance measure might well be a cepstral distance of the form:

$$d(X_i, X_j) = \left[\sum_{k=1}^R (x_i^k - x_j^k)^2 \right]^{1/2}$$

3. Clustering Training Vectors

- Goal is to cluster the set of L training vectors into a set of M codebook vectors using generalized Lloyd algorithm (also known as the K-means clustering algorithm) with the following steps:
 1. **Initialization** – arbitrarily choose M vectors (initially out of the training set of L vectors) as the initial set of codewords in the codebook
 2. **Nearest Neighbor Search** – for each training vector, find the codeword in the current codebook that is closest (in distance) and assign that vector to the corresponding cell
 3. **Centroid Update** – update the codeword in each cell to the centroid of all the training vectors assigned to that cell in the current iteration
 4. **Iteration** – repeat steps 2 and 3 until the average distance between centroids at successive iterations falls below a preset threshold

3. Clustering Training Vectors



Partitioning of a two-dimensional vector space into VQ cells with each cell represented by a centroid vector (denoted by x)

3. Clustering Training Vectors

- Assume we have a set of V vectors, $X^C = \{X_1^C, X_2^C, \dots, X_V^C\}$ where all V vectors are assigned to cluster C
- The centroid of the set X^C is defined as the vector \bar{Y} that minimizes the average distortion

$$\bar{Y} = \min_Y \frac{1}{V} \sum_{i=1}^V d(X_i^C, Y)$$

- The solution for the centroid is highly dependent on the choice of distance measure. When both X_i^C and Y are measured in a K -dimensional space with the L2-norm, the centroid is the mean of the vector set

$$\bar{Y} = \frac{1}{V} \sum_{i=1}^V X_i^C$$

- When using an L1 distance measure, the centroid is the median vector of the set of vectors assigned to the given class

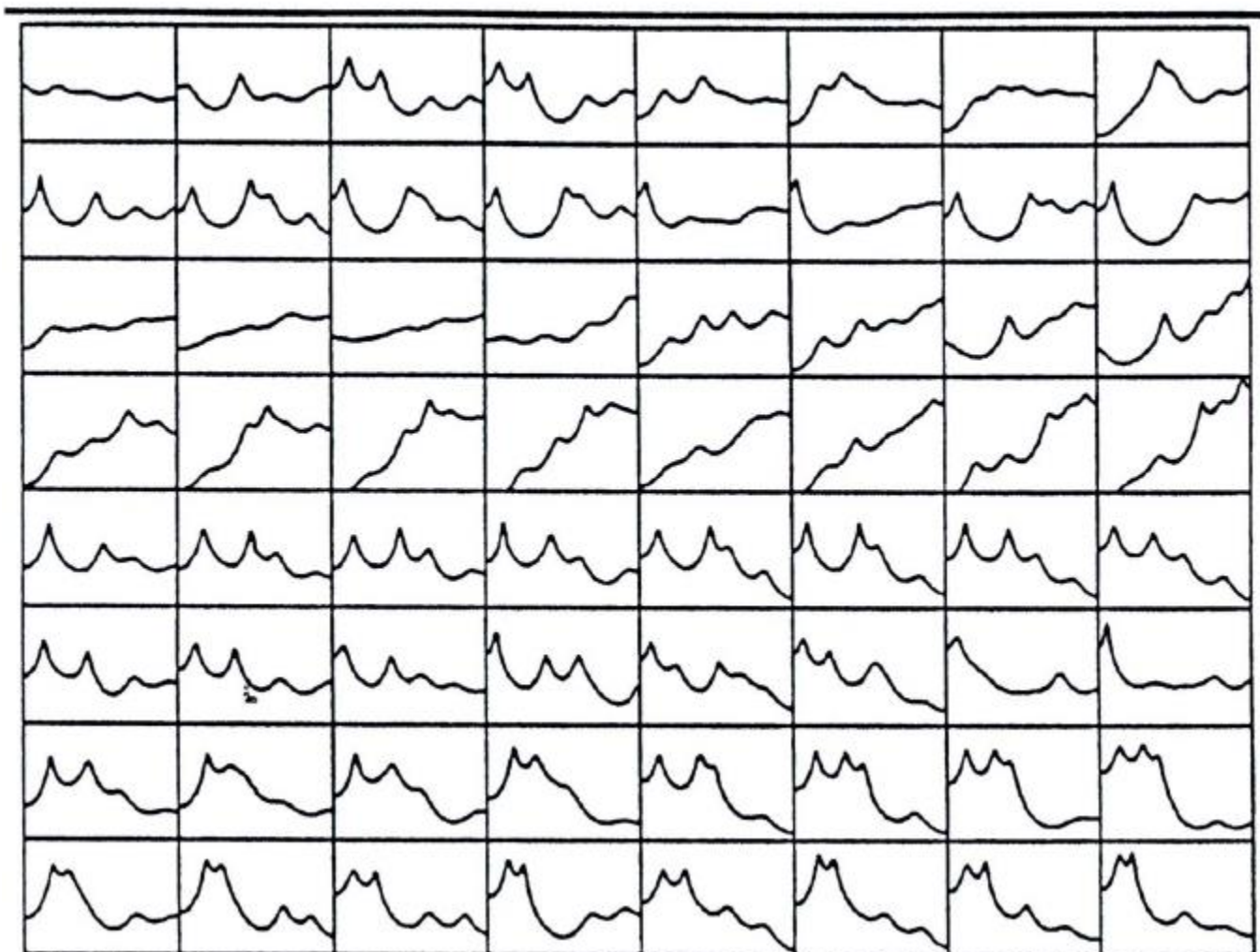
4. Vector Classification Procedure

- The classification procedure for arbitrary test set vectors is a full search through the codebook to find the "best" (minimum distance) match.
- If we denote the codebook vectors of an M -vector codebook as CB_i , for $1 \leq i \leq M$, and we denote the vector to be classified (and vector quantized) as X , then the index, i^* , of the best codebook entry is:

$$i^* = \arg \min_{1 \leq i \leq M} d(X, CB_i)$$

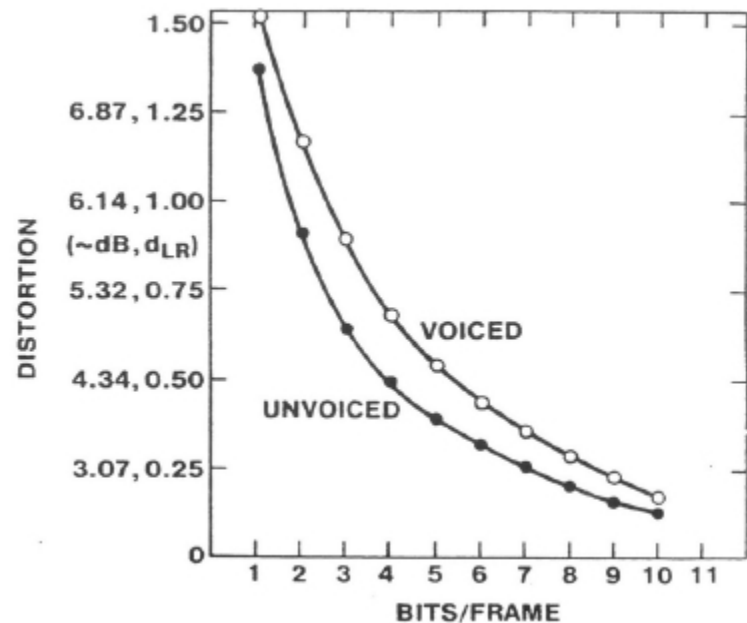
VQ Codebook Example

A VQ Codebook

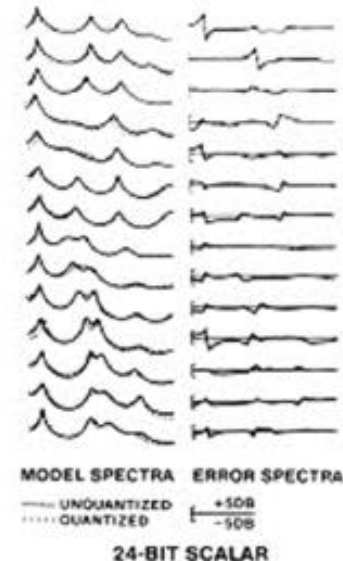
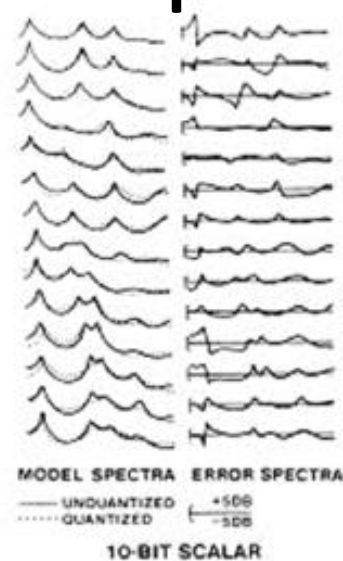


Spectral shapes
corresponding to
codebook vectors
in an $M = 64$
codebook

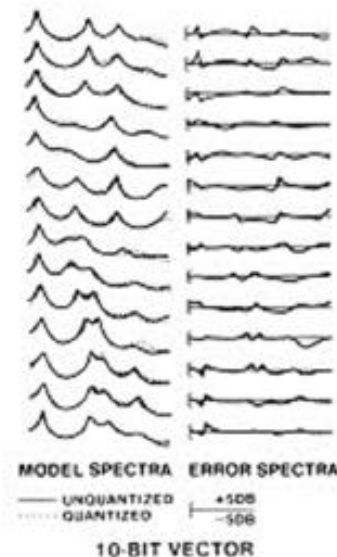
VQ Coding for Speech



'distortion' in coding computed using a spectral distortion measure related to the difference in log spectra between the original and the codebook vectors

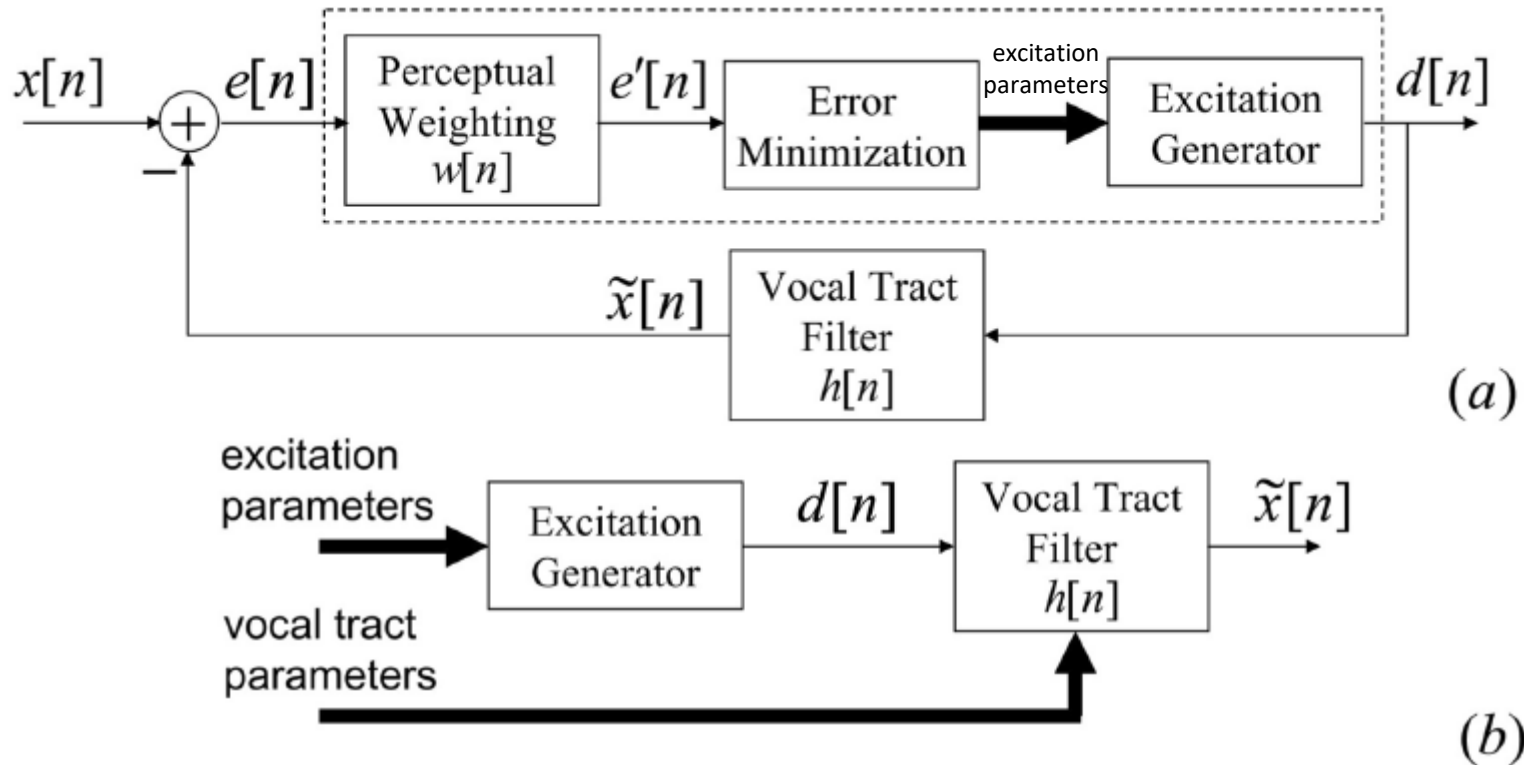


10-bit VQ comparable to 24-bit scalar quantization for these examples



Analysis-by-Synthesis Speech Coders

A-b-S Speech Coding



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

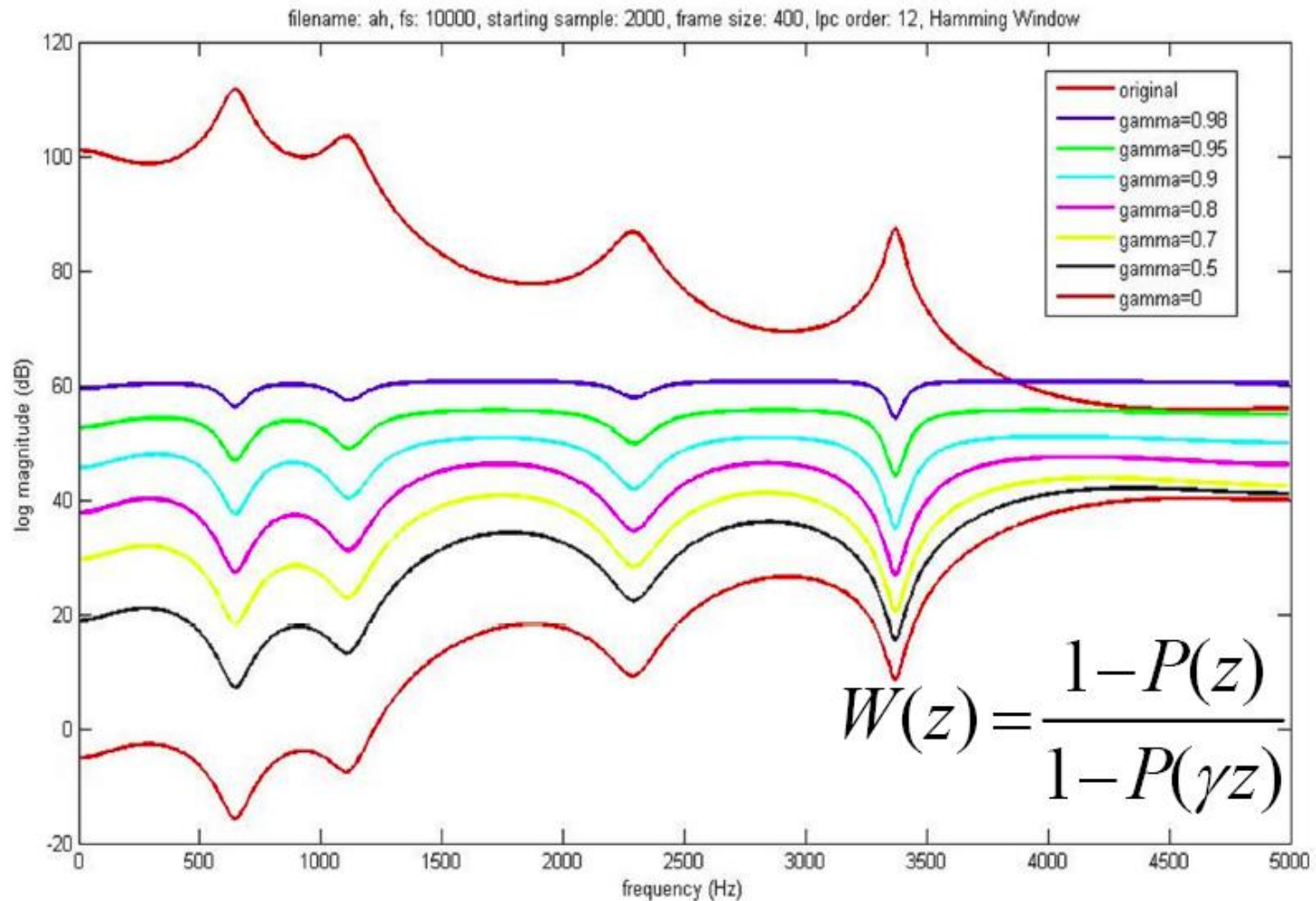
$$H(z) = \frac{1}{1 - P(z)} = \frac{1}{1 - \sum_{i=1}^p \alpha_i z^{-1}}$$

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

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

3. the error minimization box and the excitation generator create a sequence $d[n]$ of error signals that iteratively (once per loop) minimize 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.

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 in 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:

$$\mathfrak{F}_\gamma = \{f_1[n], f_2[n], \dots, f_Q[n]\}, \quad 0 \leq n \leq L-1$$

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 \mathfrak{F}_γ that maximally reduces the perceptually weighted mean square error, E :

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

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

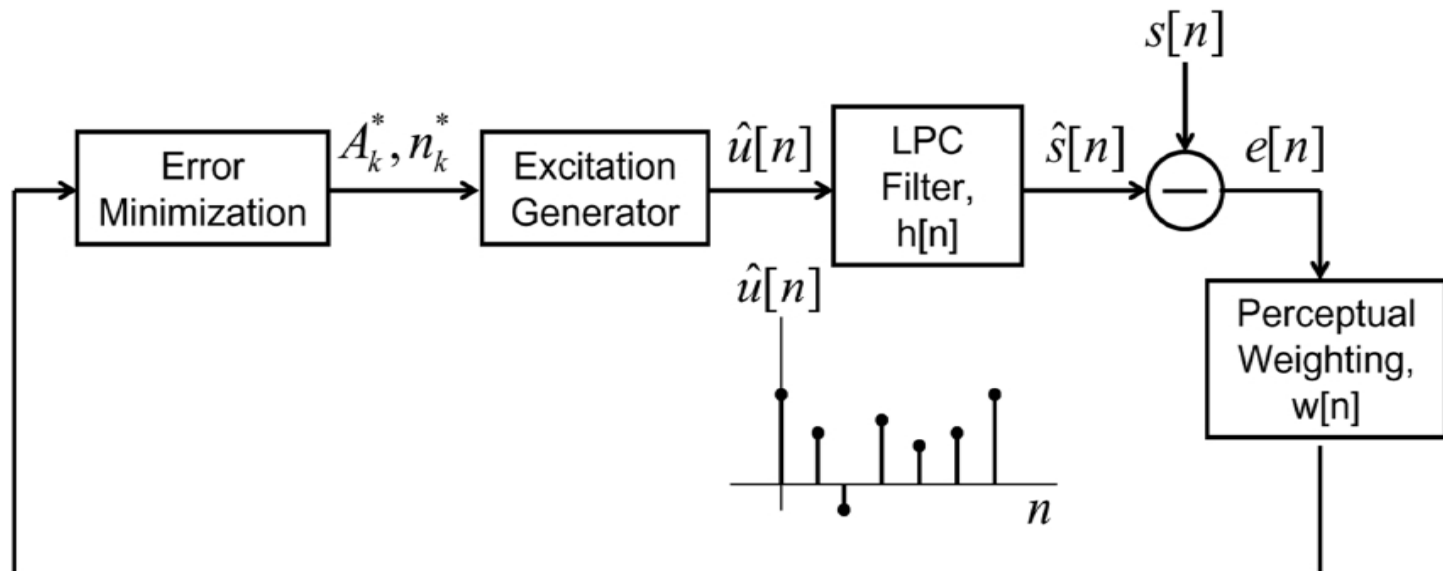
- We denote the optimal basis function at the k -th iteration as $f_{\gamma_k}[n]$, giving the excitation signal $d_k[n] = \beta_k f_{\gamma_k}[n]$ where β_k is the optimal weighting coefficient for basis function $f_{\gamma_k}[n]$
- The A-b-S iteration continues until the perceptually weighted error falls below some desired threshold, or until a maximum number of iterations, , is reached

MPLPC

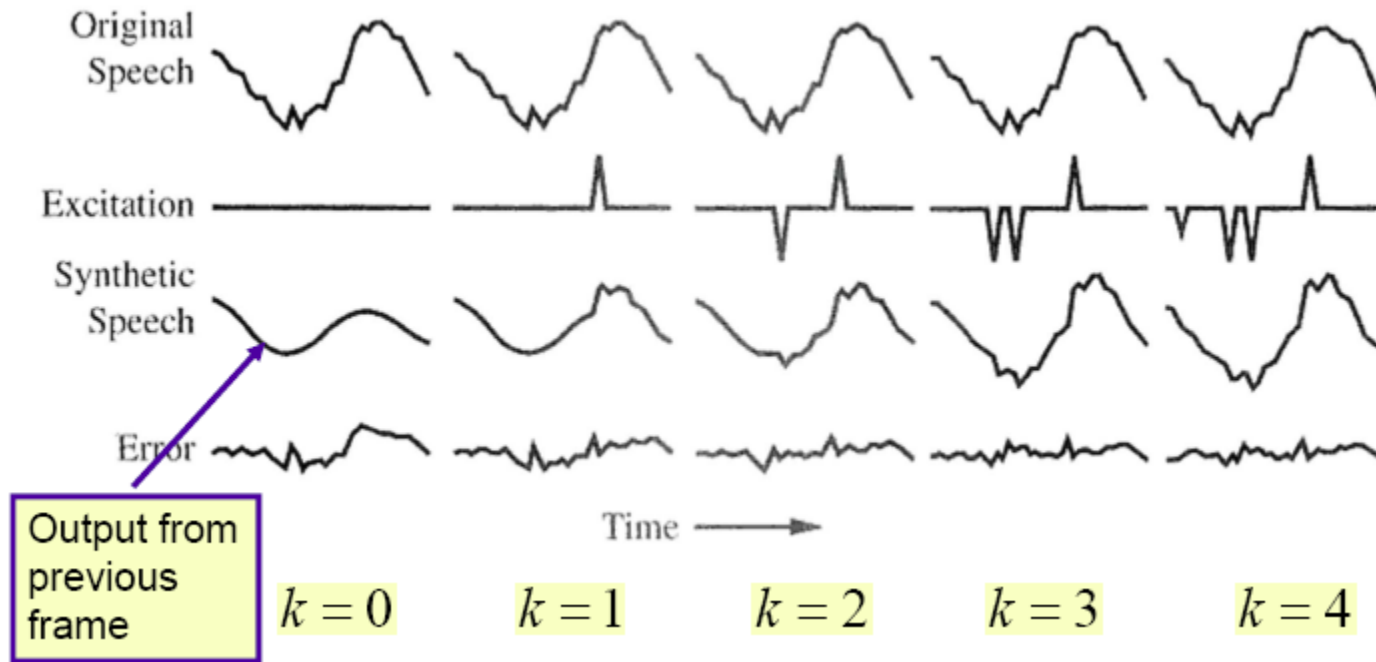
- Multipulse linear predictive coding (MPLPC)

$$f_\gamma[n] = \delta[n - \gamma] \quad 0 \leq \gamma \leq Q - 1 = L - 1$$

- 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.



MPLPC



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.

MPLPC

- 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
- To further reduce the bitrate
 - code pulse locations differentially ($\Delta_i = N_i - N_{i-1}$) to reduce range of variable
 - amplitudes normalized to reduce dynamic range

CELP

- Code-excited linear predictive coding (CELP)

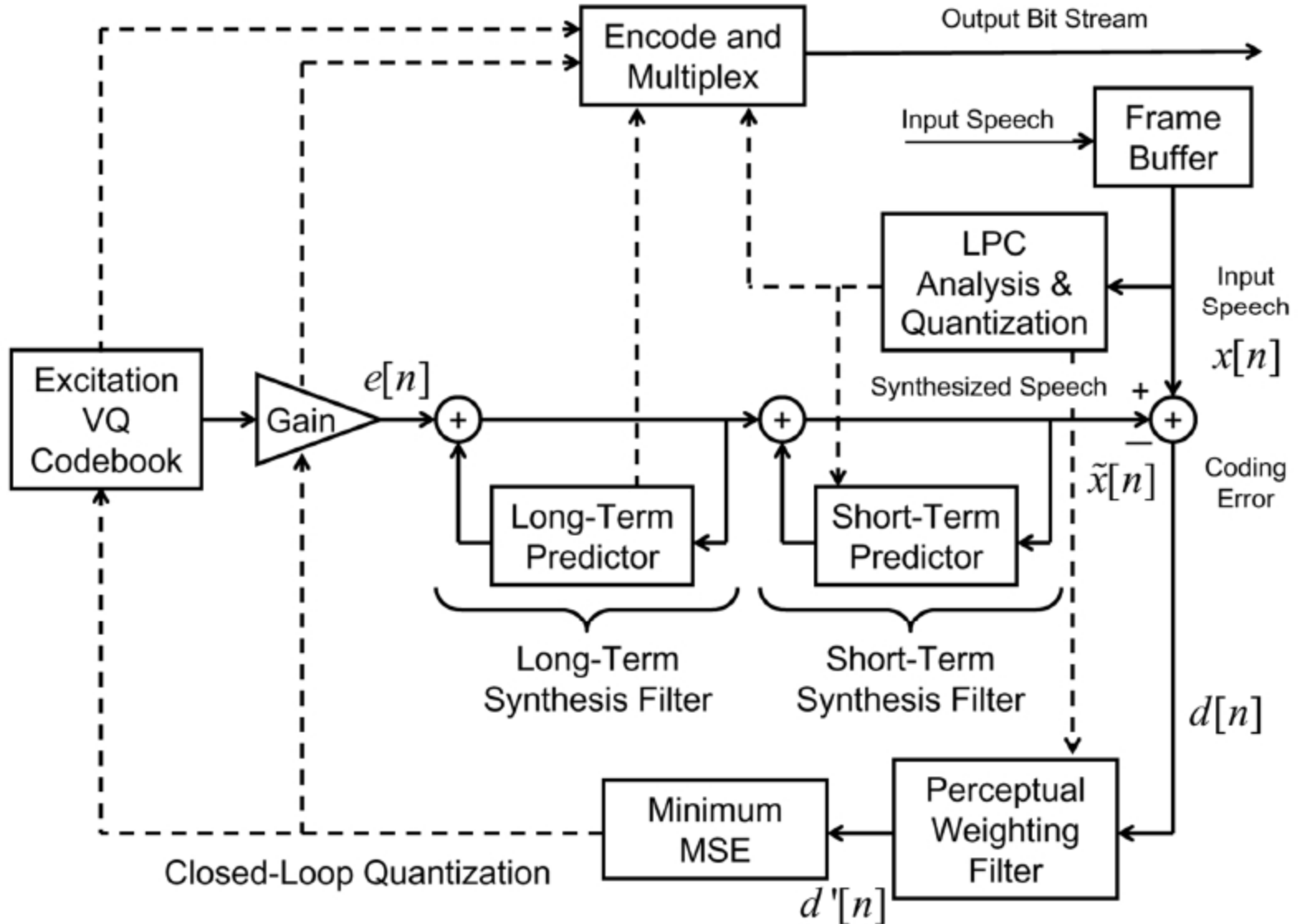
$f_\gamma[n]$ = vector of white Gaussian noise, $1 \leq \gamma \leq Q = 2^M$

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

CELP

- 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
- replace 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

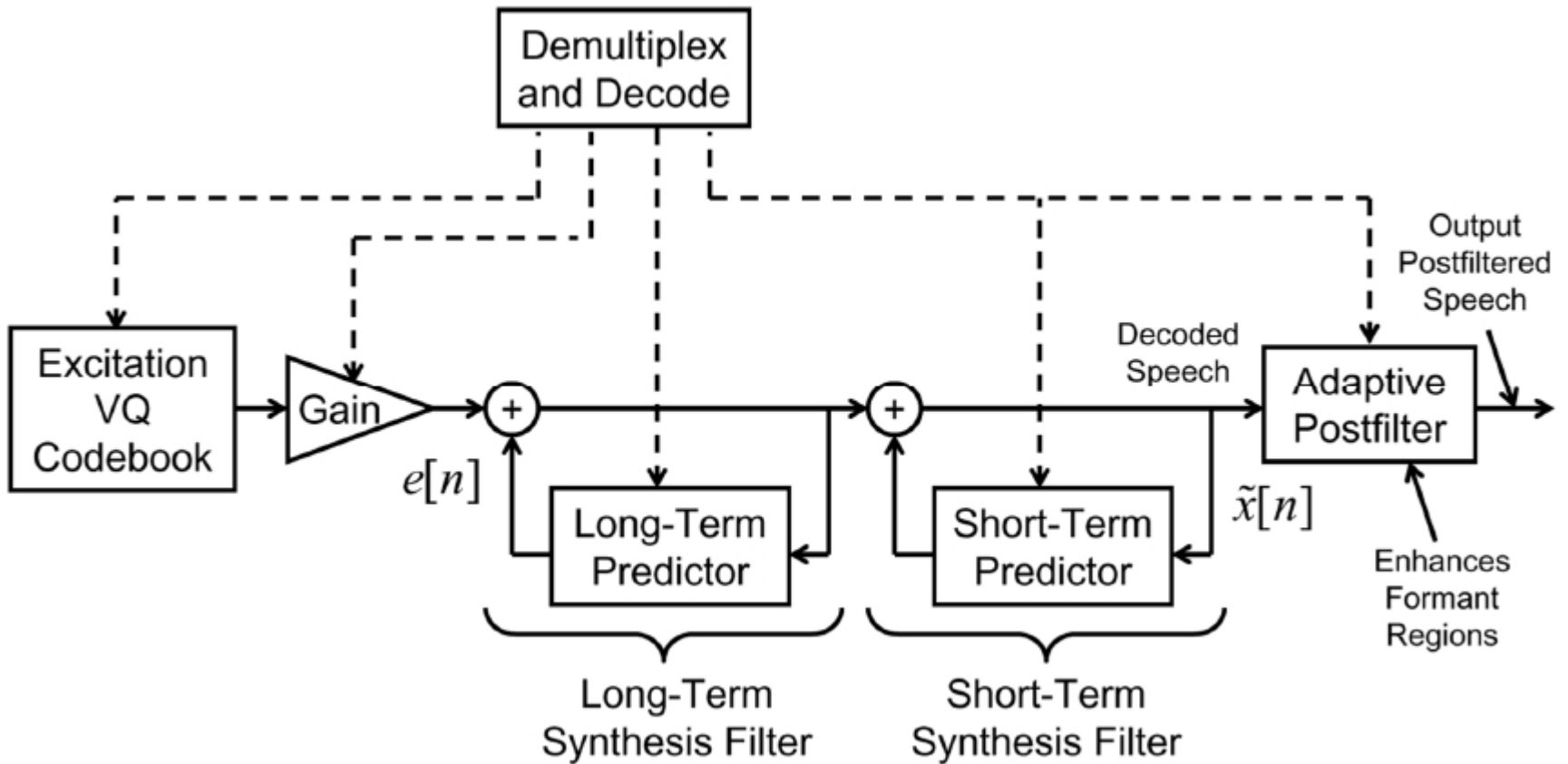
CELP Encoder



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 LP predictors, 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

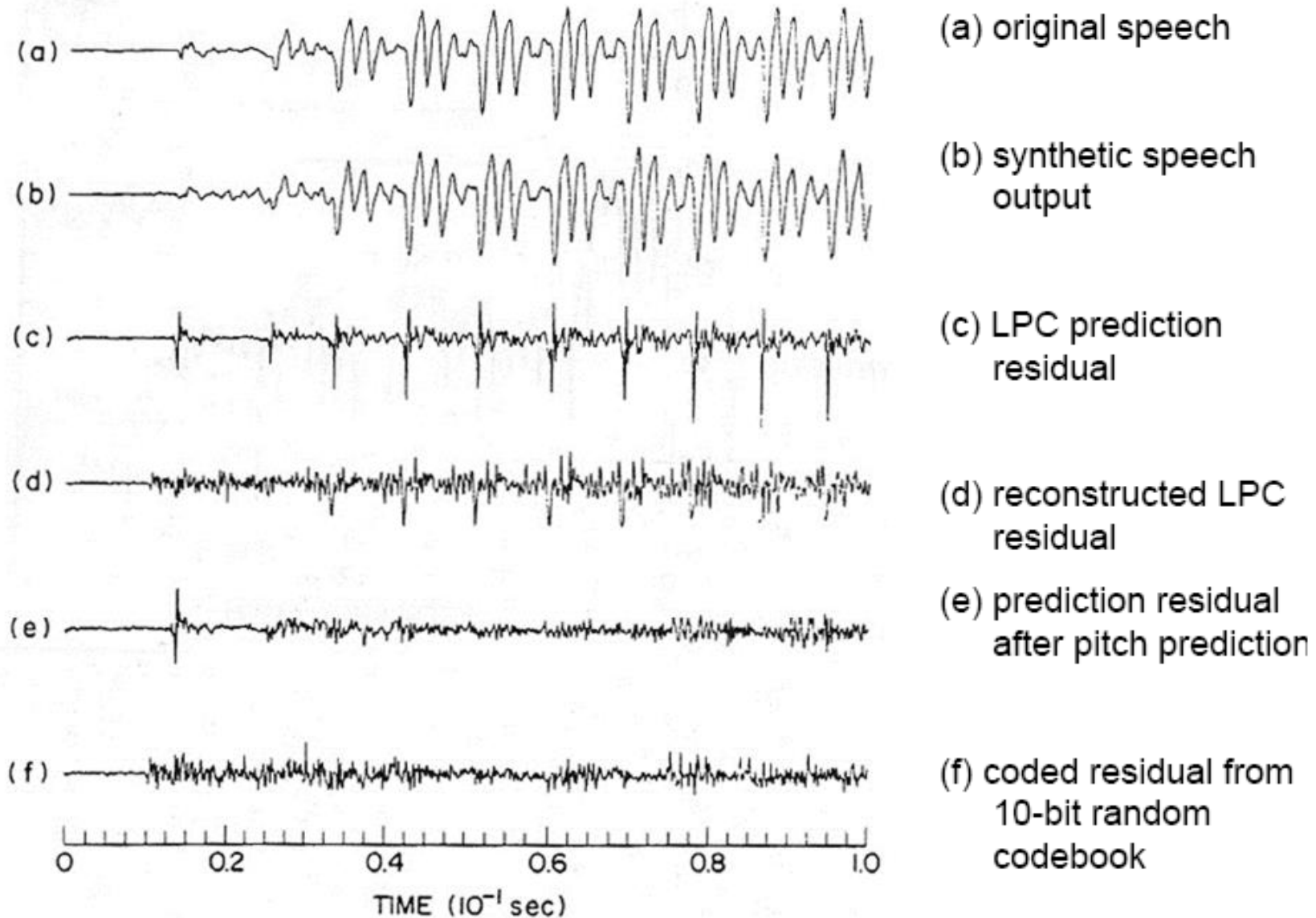
CELP Decoder



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

CELP Waveforms



Lots of CELP Variations

- **ACELP**: Algebraic Code Excited Linear Prediction (G.723.1)
- **CS-ACELP**: Conjugate-Structure ACELP (G.729)
- **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 (GSM)
- **MP-MLQ** : Multipulse-Maximum Likelihood Quantization

Speech Coding Applications and Standards

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 (low delay)
- 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 (low complexity)








Speech Coder Attributes

- **bit rate**-2400 to 128,000 bps
- **quality**-subjective (MOS), objective (*SNR*)
- **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	64kbps	Toll
G.726/727	ADPCM	16-40kbps	Toll
G.722	SBC/ADPCM	48/56/64kbps	Wideband
G.728	LD-CELP	16kbps	Toll
G.729	CS-ACELP	8kbps	Toll
G.723.1	MPC-MLQ&ACELP	6.3/5.3kbps	Toll

Demo: Coders at Different Rates

- original speech signal 
- G.711 standard, 64 Kbps mu-law PCM 
- G.726 standard, 32 Kbps ADPCM 
- G.728 standard, 16 Kbps LD-CELP 
- GSM standard, 13 Kbps RPE-LTP 
- FS1015 standard, 2.4 Kbps LPC 
- 2.4 Kbps MELP 

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 often used
- **non-speech sounds** - e.g., music on hold, DTMF tones; sounds that are poorly coded by the system

Measures of Speech Coder Quality

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

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

- 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)

Speech Coder Subjective Quality

