Speech coding

전재진

Speech coder 특성

- Signal bandwidth : narrow/wideband, audio
- Bit rate
- quality of reconstructed speech : SNR, MOS
- noise robustness
- computational complexity
- Delay
- channel-error sensitivity
- standards

MOS/ SNR

- Averaging opinion scores for a set of between 20 and 60 untrained subjects
- MOS of 4.0 or higher defines *good* or *toll* quality, where the reconstructed speech signal is generally indistinguishable from the original signal (16bit quantization)
- MOS between 3.5 and 4.0 defines *communication* quality

Table 7.1 Mean Opinion Score (MOS) is a numeric value computed as an average for a number of subjects, where each number maps to the above subjective quality.

Excellent	Good	Fair	Poor	Bad
5	4	3	2	1

- SNR: ratio of the signal's energy and the noise's energy in terms of dB

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\{x^2[n]\}}{E\{e^2[n]\}}$$

Delay/Channel

- Algorithmic delay/ Computational delay/ Multiplexing delay/ Transmission delay/ Decoder delay

- Source coding/channel coding/joint coding
  - Source coding : 용량 최소화/ channel coding : 에러 최소화, 용량 늘어남
  - 에러가 음질에 미치는 영향을 최소화
- Channel errors : random error/burst error

Scalar waveform coders

- Approximate the waveform, and, if a large enough bit rate is available, will get arbitrarily close to it.
- Linear Pulse code modulation(PCM)
  - With B bits, 2^B quantization level

uniform quantization with quantization step size Δ $\quad x_i - x_{i-1} = \Delta$

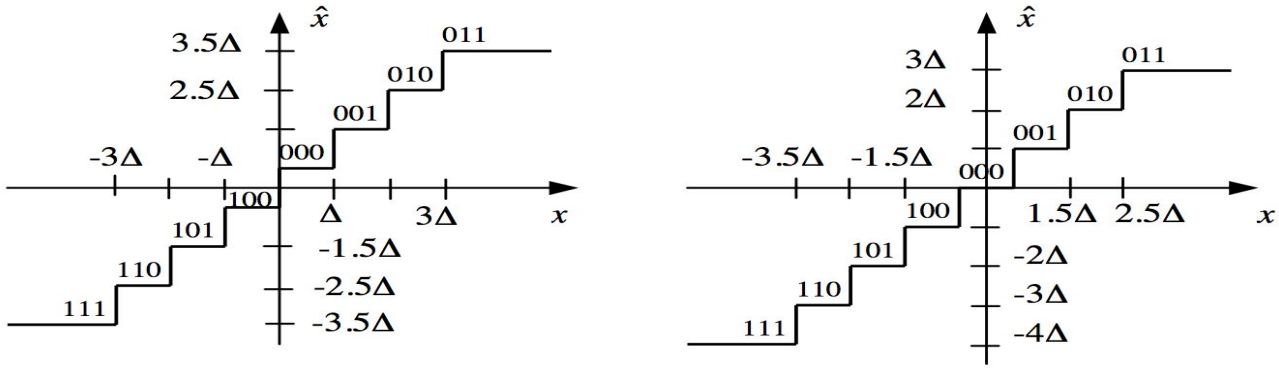

Figure 7.1 Three-bit uniform quantization characteristics: (a) mid-riser, (b) mid-tread.

PCM

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

$$-\frac{\Delta}{2} \le e[n] \le \frac{\Delta}{2}$$

1. $e[n]$ is white: $E\{e[n]e[n+m]\} = \sigma_e^2 \delta[m]$
2. $e[n]$ and $x[n]$ are uncorrelated: $E\{x[n]e[n+m]\} = 0$
3. $e[n]$ is uniformly distributed in the interval $(-\Delta/2, \Delta/2)$

PCM

$$\sigma_e^2 = \frac{\Delta^2}{12} = \frac{X_{\max}^2}{3 \times 2^{2B}}$$

$$SNR(dB) = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) = (20\log_{10} 2)B + 10\log_{10} 3 - 20\log_{10}\left(\frac{X_{\max}}{\sigma_x}\right)$$

each bit contributes to 6 dB of *SNR*, since $20\log_{10} 2 \cong 6$

A-law / mu-law PCM

- Human perception is affected by SNR, because adding noise to a signal is not as noticeable if the signal energy is large enough
  - Log expander 사용

$$y[n] = \ln|x[n]|$$

$$\hat{y}[n] = y[n] + \varepsilon[n]$$: Uniform quantizer

$$\hat{x}[n] = \exp\{\hat{y}[n]\}\text{sign}\{x[n]\} = x[n]\exp\{\varepsilon[n]\}$$

$$\hat{x}[n] \cong x[n](1+\varepsilon[n]) = x[n] + x[n]\varepsilon[n]$$

$SNR = 1/\sigma_\varepsilon^2$ is constant for all levels : Ideal case

A-law / mu-law PCM

$$y[n] = X_{\max} \frac{1+\log\left[\frac{A|x[n]|}{X_{\max}}\right]}{1+\log A}\operatorname{sign}\{x[n]\} \qquad y[n] = X_{\max} \frac{\log\left[1+\mu\frac{|x[n]|}{X_{\max}}\right]}{\log[1+\mu]}\operatorname{sign}\{x[n]\}$$

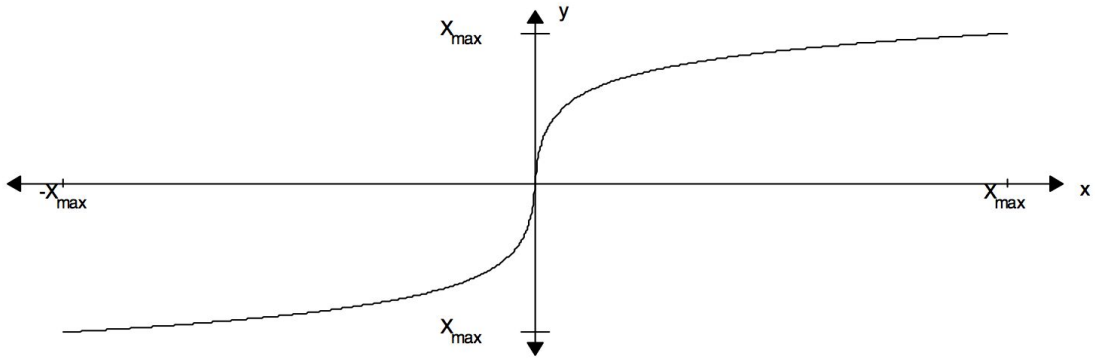

Figure 7.2 Nonlinearity used in the μ-law compression.

G.711

- 8bit 8khz
- mu=255, A=87.56
- SNR 35dB
- Uniform quantizer 12bit
- MOS 4.0

Adaptive PCM(APCM)

- Quantization step size to be large enough to accommodate the maximum peak-to-peak range of the signal and avoid clipping
- Small step size to minimize the quantization noise.
  - To adapt the step size to the level of the input signal.

$$\Delta[n] = \Delta_0 \sigma[n] \qquad G[n] = G_0 / \sigma[n]$$

$$\sigma^2[n] = \alpha\sigma^2[n-1] + (1-\alpha)x^2[n-1] \qquad \sigma^2[n] = \frac{1}{M}\sum_{m=n-M}^{n-1} x^2[m]$$

$$\Delta_{\min} \le \Delta[n] \le \Delta_{\max}$$

$$G_{\min} \le G[n] \le G_{\max}$$

$\Delta_{\max} / \Delta_{\min}$ and $G_{\max} / G_{\min}$ determining the dynamic range

- SNR over a range of 40 dB, these ratios can be 100.

APCM

Feedforward adaptation schemes require us to transmit, in addition to the quantized signal, either the step size $\Delta[n]$ or the gain $G[n]$. Because these values evolve slowly with time, they can be sampled and quantized at a low rate.

Feedback adaptation to avoid having to send information about the step size or gain. In this case, the step size and gain are estimated from the quantizer output, so that they can be recreated at the decoder without any extra information

APCM exhibits an improvement between 4–8 dB over μ-law PCM for the same bit rate

Differential PCM(DPCM)

There is considerable correlation between adjacent samples, because on the average the signal doesn't change rapidly from sample to sample

$$d[n] = x[n] - \tilde{x}[n]$$: predicted value tilda $x[n]$

$$\hat{d}[n] = Q\{d[n]\} = d[n] + e[n]$$

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

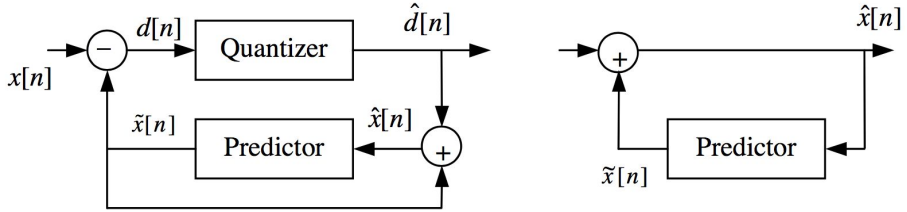

Figure 7.3 Block diagram of a DPCM encoder and decoder with feedback prediction.

DPCM

Delta modulation(DM) is 1-bit DPCM

$$\tilde{x}[n] = x[n-1] \qquad d[n] = \begin{cases} \Delta & x[n] > x[n-1] \\ -\Delta & x[n] \le x[n-1] \end{cases}$$

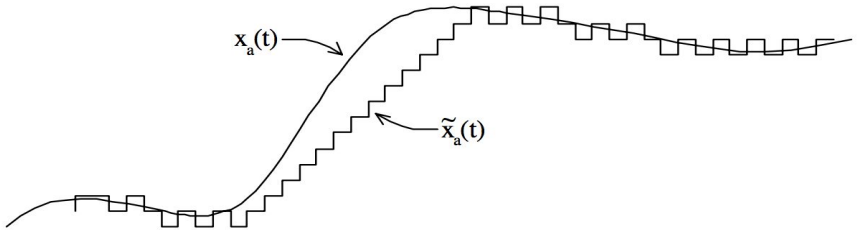

Figure 7.4 An example of slope overload distortion and granular noise in a DM encoder.

Delta size 에 따라 slope overload distortion/granular noise

Coder is indeed very simple, sampling rates of over 200 kbps are needed for SNRs comparable to PCM, so DM is rarely used as a speech coder

DPCM : sigma-delta modulation

If the signal is oversampled by a factor *N,* and the step size is reduced by the same amount (i.e., Δ/N), the slope overload will be the same, but the granular noise will decrease by a factor *N*

Delta modulation is useful in the design of analog-digital converters, in a variant called sigma-delta modulation

Advantages of this technique as an analog-digital converter are that inexpensive analog filters can be used and only a simple 1-bit A/D is needed.

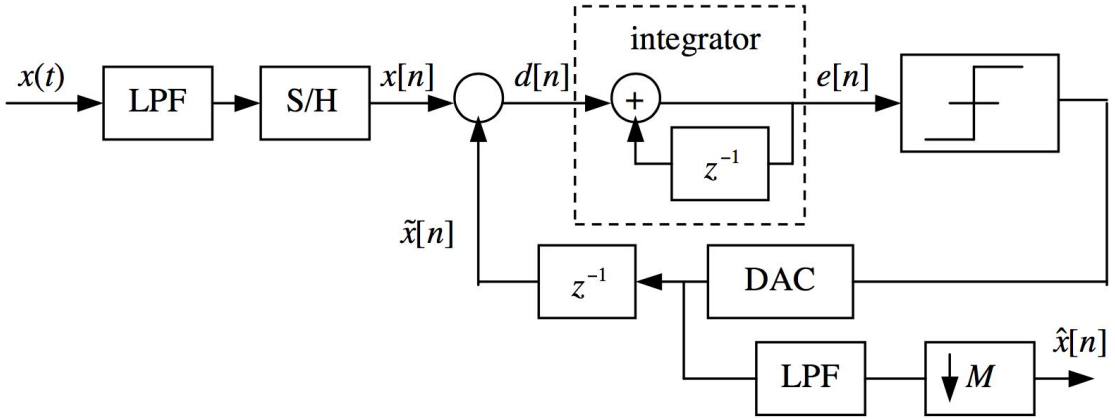

Figure 7.5 A sigma-delta modulator used in an oversampling analog-digital converter.

Adaptive Delta Modulation(ADM)/ Improved DPCM

$$\Delta[n] = \begin{cases} \alpha\Delta[n-1] + k_1 & \text{if } e[n], e[n-1] \text{ and } e[n-2] \text{ have same sign} \\ \alpha\Delta[n-1] + k_2 & \text{otherwise} \end{cases}$$

$$0 < \alpha < 1 \text{ and } 0 < k_2 << k_1$$

Improved PCM

Prediction: linear combination of past quantized values

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k \hat{x}[n-k]$$

성능비교

Table 7.2 Common scalar waveform standards used.

Standard	Bit Rate (kbits/sec)	MOS	Algorithm	Sampling Rate (kHz)
Stereo CD Audio	1411	5.0	16-bit linear PCM	44.1
WAV, AIFF, SND	Variable	-	16/8-bit linear PCM	8, 11.025, 16, 22.05, 44.1, 48
G.711	64	4.3	μ-law/A-law PCM	8
G.727	40, 32, 24, 16	4.2 (32k)	ADPCM	8
G.722	64, 56, 48		Subband ADPCM	16

Scalar Frequency Domain Coder

The samples of a speech signal have a great deal of correlation among them, whereas frequency domain components are approximately uncorrelated

The perceptual effects of masking can be more easily implemented in the frequency domain

Frequency-domain coding has been mostly used for CD-quality signals and not for 8-kHz speech signals

MP3 - masked threshold

$$x[n] = \frac{s[n]}{N2^{b-1}}$$: Normalize

$$w[n] = 0.5 - 0.5\cos\left(2\pi n / N\right)$$:hanning window

$$P[k] = P_0 + 10\log_{10}\left(\sum_{n=0}^{N-1} w[n]x[n]e^{-j2\pi nk/N}\right)$$: spectrum

$$P[k] > P[k \pm l] + 7dB \qquad 1 < l \le \Delta_k$$

$$\Delta_k = \begin{cases} 2 & 2 < k < 63 & (170\text{Hz} - 5.5\text{kHz}) \\ 3 & 63 \le k < 127 & (5.5\text{kHz}, 11\text{kHz}) \\ 6 & 127 \le k \le 256 & (11\text{kHz}, 22\text{kHz}) \end{cases}$$

:tonal component detect

$$P_{TM}[k] = 10\log_{10}\left(\sum_{j=-1}^{j} 10^{0.1P[k+j]}\right)$$: Tonal Mask

$$P_{NM}[\bar{k}] = 10\log_{10}\left(\sum_{j} 10^{0.1P[j]}\right)$$: Noise Mask

The noise maskers are computed from as the sum of power spectrum of the remaining frequency bins $\bar{k}$ in a critical band not within a neighborhood Δ_k of the tonal maskers:

$$T[k] = \max\left(T_h[k], \max_i\left(T_i[k]\right)\right)$$: Overall masked threshold (see chapter 2 ??)

Transform coder

Adaptive Spectral Entropy Coding of High Quality Music Signals (ASPEC) algorithm: basis for the MPEG1 Layer 1
The DFT coefficients are grouped into 128 subbands/128 scalar quantizers are used to transmit all the DFT coefficients
Each subband *j* has a quantizer having *k j* levels and step size of *Tj* as

$$k_j = 1 + 2 \times \mathrm{rnd}\left(P_j / T_j \right)$$

where *Tj* is the quantized JND threshold, *Pj* is the quantized magnitude of the largest real or imaginary component of the *j* subband
Entropy coding is used to encode the coefficients of that subband
Both *Tj* and *Pj* are quantized on a dB scale using 8-bit uniform quantizers with a 170-dB dynamic range, thus with a step size of 0.66 dB, then they are transmitted as side information.

Two frequency-domain representation:
1. Through subband filtering via a filterbank. When a filterbank is used, the bandwidth of each band is chosen to increase with frequency following a perceptual scale, such as the Bark scale.
2. Through frequency-domain transforms. Instead of using a DFT, higher efficiency can be obtained by the use of an MDCT (see Chapter 5)

기타 코덱

Dolby digital/MPEG/DTS/Perceptual audio coder(PAC)/DAB

LPC Vocoder

Linear predictors removes redundancy in the signal, so that coding of the residual signal can be done with simpler quantizers

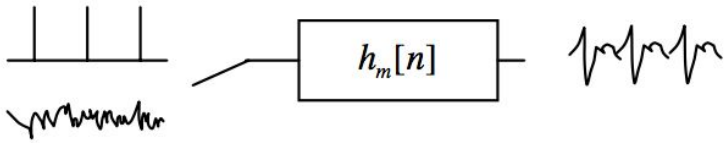

Figure 7.6 Block diagram of an LPC vocoder.

전송 정보 : gain, LPC, voiced/unvoiced, pitch info. (voiced 경우)

장점 : highly intelligible , 2.4 kbps

단점 : voiced 신호 음질이 나쁨, pitch estimator error 에 민감, 배경잡음에 취약

Code Excited Linear Prediction(CELP)

Quantize the LPC residual using VQ

전송정보 : LPC coefficients, codeword index

The prediction using LPC coefficients is called *short-term prediction*/The prediction of the residual based on pitch is called *long-term prediction*.

We first estimate the *p*th-order LPC coefficients from the samples *x*[*n*] for frame *t* using the autocorrelation method

$$e[n] = \sum_{i=1}^{p} a_i x[n-i] \qquad H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \sum_{i=0}^{\infty} h_i z^{-i}$$

CELP

$$h[n] = \begin{cases} 1 & n = 0 \\ \sum_{i=1}^{n} a_i h[n-i] & 0 < n < p \\ \sum_{i=1}^{p} a_i h[n-i] & p \le n < M \end{cases}$$

: M coefficient of h[n]

so that if we quantize a frame of M samples of the residual $\mathbf{e} = (e[0], e[1], \cdots e[M-1])^T$ to $\mathbf{e}_i = (e_i[0], e_i[1], \cdots e_i[M-1])^T$, we can compute the reconstructed signal $\hat{x}_i[n]$ as

$$\hat{x}_i[n] = \sum_{m=0}^{n} h[m] e_i[n-m] + \sum_{m=n+1}^{\infty} h[m] e[n-m] \qquad r_0[n] = \sum_{m=n+1}^{\infty} h[m] e[n-m]$$

$$\hat{\mathbf{x}}_i = \mathbf{H}\mathbf{e}_i + \mathbf{r}_0$$

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \cdots & 0 & 0 \\ h_1 & h_0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ h_{M-1} & h_{M-2} & \cdots & h_0 & 0 \\ h_M & h_{M-1} & \cdots & h_1 & h_0 \end{bmatrix}$$

CELP

$\mathbf{e}_i = \lambda \mathbf{c}_i$ where λ is the gain and $\mathbf{c}_i$ is the codebook entry i.

$\hat{\mathbf{x}}_i = \lambda \mathbf{H} \mathbf{c}_i + \mathbf{r}_0$: reconstructed signal

$\boldsymbol{\varepsilon} = \mathbf{x} - \hat{\mathbf{x}}_i$: estimation error

$$E(i,\lambda) = \left|\mathbf{x} - \hat{\mathbf{x}}_i\right|^2 = \left|\mathbf{x} - \lambda \mathbf{H} \mathbf{c}_i - \mathbf{r}_0\right|^2 = \left|\mathbf{x} - \mathbf{r}_0\right|^2 + \lambda^2 \mathbf{c}_i^T \mathbf{H}^T \mathbf{H} \mathbf{c}_i - 2\lambda \mathbf{c}_i^T \mathbf{H}^T (\mathbf{x} - \mathbf{r}_0)$$: minimize err function

$$\lambda_i = \frac{\mathbf{c}_i^T \mathbf{H}^T (\mathbf{x} - \mathbf{r}_0)}{\mathbf{c}_i^T \mathbf{H}^T \mathbf{H} \mathbf{c}_i} \qquad j = \arg\min_i \left\{ -\frac{\left(\mathbf{c}_i^T \mathbf{H}^T (\mathbf{x} - \mathbf{r}_0)\right)^2}{\mathbf{c}_i^T \mathbf{H}^T \mathbf{H} \mathbf{c}_i} \right\}$$

LPC coeff. Est → code book index → gain lambda calc. : open-loop estimation

Closed loop est : jointly estimate parameters/ computationally expensive and low squared error

CELP

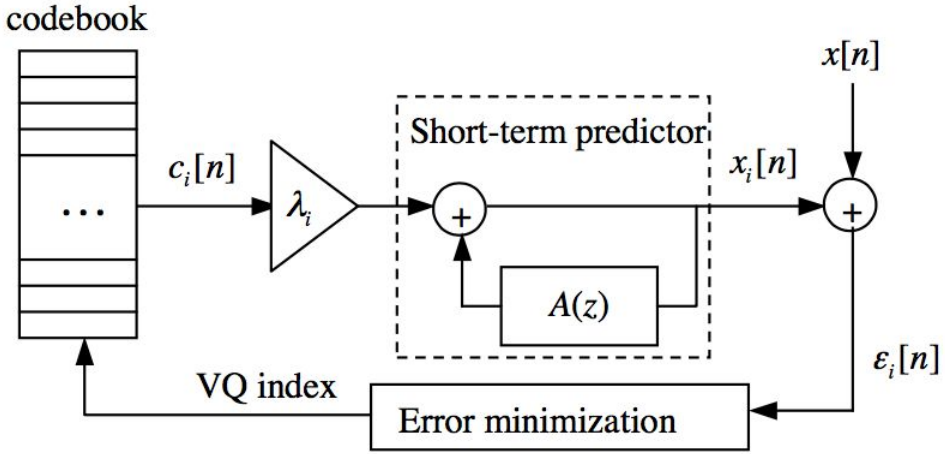

Figure 7.7 Analysis-by-synthesis principle used in a basic CELP.

Pitch Estimation : Adaptive Codebook

The fact that speech is highly periodic during voiced segments can also be used to reduce redundancy in the signal

Predicting the residual signal *e*[*n*] at the current vector with samples from the past residual signal shifted a pitch period *t*

$$e[n] = \lambda_t^a e[n-t] + \lambda_i^f c_i^f[n] = \lambda_t^a c_t^a[n] + \lambda_i^f c_i^f[n]$$

$$\mathbf{e}_{ti} = \lambda_t^a \mathbf{c}_t^a + \lambda_i^f \mathbf{c}_i^f$$: combine adaptive/fixed codebook

$$\mathbf{c}_t^a = (e[-t], e[1-t], \cdots, e[M-1-t])^T$$

t is the delay which specifies the start of the adaptive codebook entry *t*

Perceptual Weighting and Post Filtering

minimization of the error is not necessarily the best criterion

perceptual weighting filter tries to shape the noise so that it gets masked by the speech signal

$$W(z) = \frac{A(z/\beta)}{A(z/\gamma)} \qquad A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$$

Choosing γ and β so that and $0 < \gamma < \beta \leq 1$, implies that the roots of $A(z/\beta)$ and $A(z/\gamma)$ will move closer to the origin of the unit circle than the roots of $A(z)$, thus resulting in a frequency response with wider resonances. This perceptual filter therefore deemphasizes the contribution of the quantization error near the formants. A common choice of parameters is

CELP with perceptual weighting

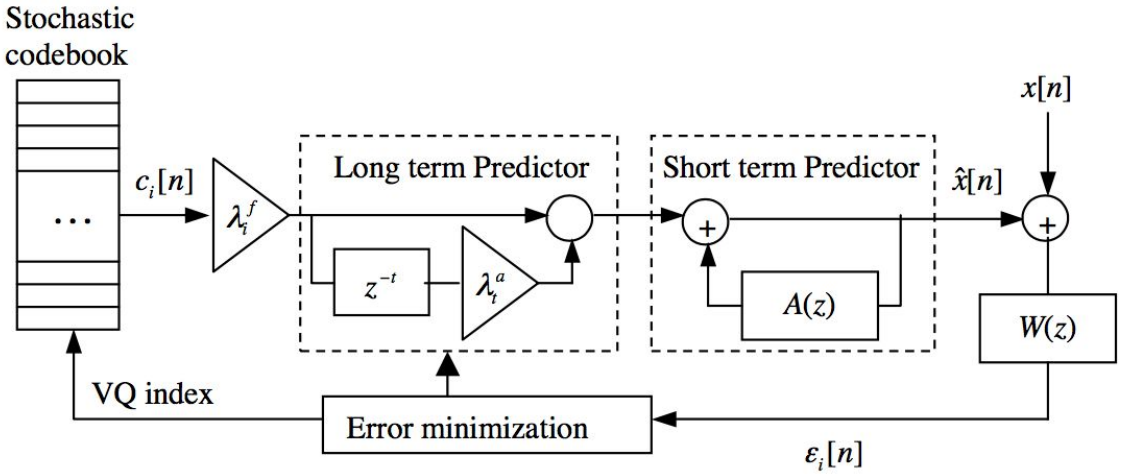

Figure 7.8 Diagram of a CELP coder. Both long-term and short-term predictors are used, together with a perceptual weighting.

CELP Standards

Table 7.4 Several CELP standards used in the H.323 specification used for teleconferencing and voice streaming through the internet.

Standard	Bit Rate (kbps)	MOS	Algorithm	H.323	Comments
G.728	16	4.0	No pitch prediction	Optional	Low -delay
G.729	8	3.9	ACELP	Optional	
G.723.1	5.3, 6.3	3.9	ACELP for 5.3k	Optional	

Table 7.5 CELP standards used in cellular telephony.

Standard	Bit Rate (kbps)	MOS	Algorithm	Cellular	Comments
Full-rate GSM	13	3.6	VSELP RTE-LTP	GSM	
EFR GSM	12.2	4.5	ACELP	GSM	
IS-641	7.4	4.1	ACELP	PCS1900	
IS-54	7.95	3.9	VSELP	TDMA	
IS-96a	max 8.5	3.9	QCELP	CDMA	Variable-rate

Low bit rate speech coders

mixed-excitation LPC vocoder, harmonic coding, and waveform interpolation 등이 있음

Speech synthesis 에 사용

Waveform approximating coders (scalar waveform coders, frequency domain coders, CELP) 는 bit rate 증가시 SNR 증가

Low bit rate coders : Perceptually equivalent 한 신호 생성이 목적이라 SNR 이 무한히 증가하지 않음

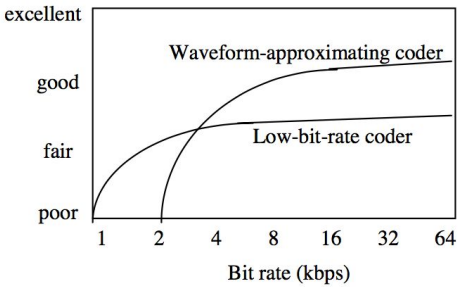

Figure 7.9 Typical subjective performance of waveform-approximating and low-bit-rate coders as a function of the bit rate. Note that waveform-approximating coders are a better choice for bit rates higher than about 3 kbps, whereas parametric coders are a better choice for lower bit rates. The exact cutoff point depends on the specific algorithms compared.

Mixed Excitation LPC Vocoder

The main weakness of the LPC vocoder is the binary decision between voiced and unvoiced speech

By having a separate voicing decision for each of a number of frequency bands, the performance can be enhanced significantly

MELP LPC Vocoder : MOS 3.3 @ 2.4kbps

Harmonic Coding

Sinusoidal coding decomposes the speech signal or the LP residual signal into a sum of sinusoids
For unvoiced speech, using a default pitch of 100 Hz results in acceptable quality

$$\tilde{s}[n] = \sum_{l=0}^{T_0-1} A_l \cos(nl\omega_0 + \phi_l)$$

$$e[n] = s[n] - \tilde{s}[n]$$

$$s_k[n] = s[n]w_k[n] = s[n]w[kN - n]$$: frame 별 처리

sinusoid parameters for frame k (ω_0^k, A_l^k and ϕ_l^k)

$$\hat{s}[n] = w[n]\tilde{s}^{k-1}[n] + w[n-N]\tilde{s}^{k}[n]$$: overlap-add

$$w[n] + w[n-N] = 1$$: perfect reconstruction

Parameter Estimation

$$\tilde{s}[n] = \sum_{l}^{T_0-1} A_l \exp\{j(nl\omega_0 + \phi_l)\}$$: complex form

$$\tilde{S}_W(\omega) = \sum_{l=0}^{T_0-1} A_l e^{j\phi_l} W(\omega - l\omega_0)$$: DFT

$$E = |S(\omega) - \tilde{S}_W(\omega)|^2$$

$$A_l = \frac{|S(l\omega_0)|}{W(0)}$$: magnitude $\qquad \phi_l = \arg S(l\omega_0)$: Phase

Voiced/unvoiced decisions can be computed from the ratio between the energy of the signal and that of the reconstruction error

Frames with SNR higher than 13 dB are generally voiced and lower than 4 dB unvoiced

$$SNR = \frac{\sum_{n=-N}^{N} |s[n]|^2}{\sum_{n=-N}^{N} |s[n] - \tilde{s}[n]|^2}$$

Phase Modeling

$$e[n] = T_0 \sum_{k=-\infty}^{\infty} \delta[n - n_0 - kT_0] = \sum_{l=0}^{T_0-1} e^{j(n-n_0)\omega_0 l}$$: impulse train

$$s[n] = \sum_{l=0}^{T_0-1} A(l\omega_0) \exp\{j[(n-n_0)\omega_0 l + \Phi(l\omega_0)]\}$$

$$\phi_l = -n_0\omega_0 l + \Phi(l\omega_0)$$

Since the sinusoidal model has too many parameters to lead to low-rate coding, a common technique is to not encode the phases → minimum phase system 으로 가정

The magnitude spectrum is known at the pitch harmonics, and the remaining values can be filled in by interpolation: e.g., linear or cubic splines
This interpolated magnitude spectrum can be approximated through the real cepstrum

Phase Modeling

$$|\tilde{A}(\omega)| = c_0 + 2\sum_{k=1}^{K} c_k \cos(k\omega)$$

$$\tilde{\Phi}(\omega) = -2\sum_{k=1}^{K} c_k \sin(k\omega)$$

: minimum phase

$$\phi_0(t) = \phi_0((k-1)N) + \int_{(k-1)N}^{t} \omega_0(t)dt$$

: phase of the first harmonic between frames (k – 1) and k

$$\omega_0(t) = \omega_0^{k-1} + \frac{\omega_0^k - \omega_0^{k-1}}{N}t$$

: frequency vary linearly between frames (k – 1) and k

$$\phi_0^k = \phi_0(kN) = \phi_0((k-1)N) + (\omega_0^{k-1} + \omega_0^k)(N/2)$$

: 위 두식을 합하고, t=kN 일때

$$\phi_l^k = \Phi^k(l\omega_0) + l\phi_0^k$$

: Phase of the sinusoid at w_0 as a function of the fundamental frequencies at frames (k –1), k and the phase at frame (k – 1)

Phase Modeling

Phases model : good approximation for voiced sounds
For unvoiced sounds, random phases are needed
Voiced fricatives and many voiced sounds have an aspiration component, so that a mixed excitation is needed to represent them
In these cases, the source is split into different frequency bands and each band is classified as either voiced or unvoiced
Sinusoids in voiced bands use the phases described above, whereas sinusoids in unvoiced bands have random phases

참고: Parameter quantization : Variable-Dimension Vector Quantization(VDVQ)

Waveform Interpolation

Pitch pulse changes slowly over time for voiced speech

$$x_m[n] = w_m[n]x[n]$$

$$x[n] = \sum_{m=-\infty}^{\infty} x_m[n - t_m]$$

$$x_s[n] = \tilde{u}[n, sT] = \frac{\sum_m h[sT - t_m]u[n, t_m]}{\sum_m h[sT - t_m]}$$

: slowly evolving waveform(SEW)
Lowpass filtered signal

$$\tilde{w}_m[n] = \tilde{u}[n, t_m] = \frac{\sum_s h[t_m - sT]w_s[n]}{\sum_s h[t_m - sT]}$$

:reconstruct each pitch waveform from the SEW by interpolation between adjacent pitch waveforms

Waveform Interpolation

$$u[n,l] = \tilde{u}[n,l] + \hat{u}[n,l]$$

$$x_m[n] = \tilde{x}_m[n] + \hat{x}_m[n]$$

신호를 REW (Rapidly evolving Waveform)과 SEW 로 나눔
이러한 방식을 Residual 신호에 주로 적용함

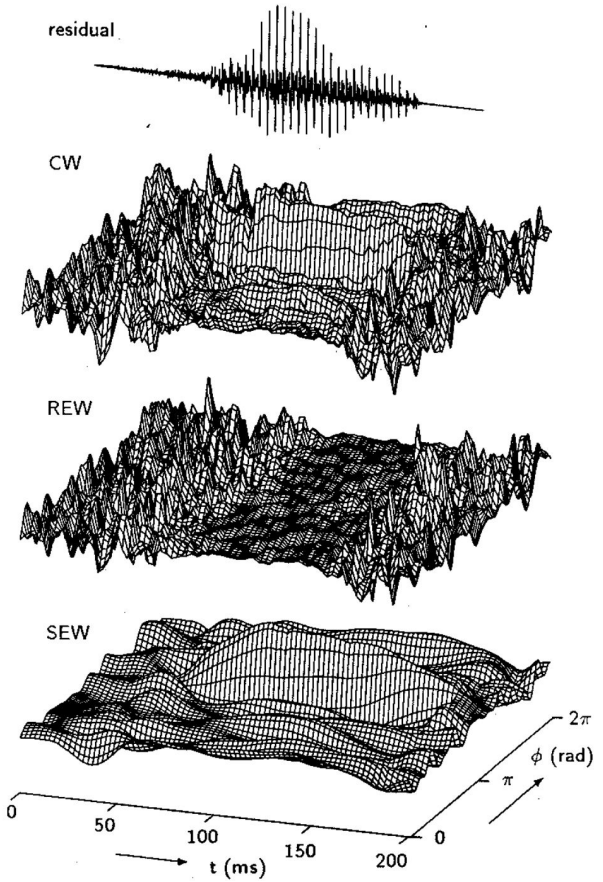

Figure 7.10 LP residual signal and its associated surface $u(t,\phi)$. In the ϕ axis we have a normalized pitch pulse at every given time t. Decomposition of the surface into a slowly evolving waveform and a rapidly evolving waveform (After Kleijn [30], reprinted by permission of IEEE).

Waveform Interpolation

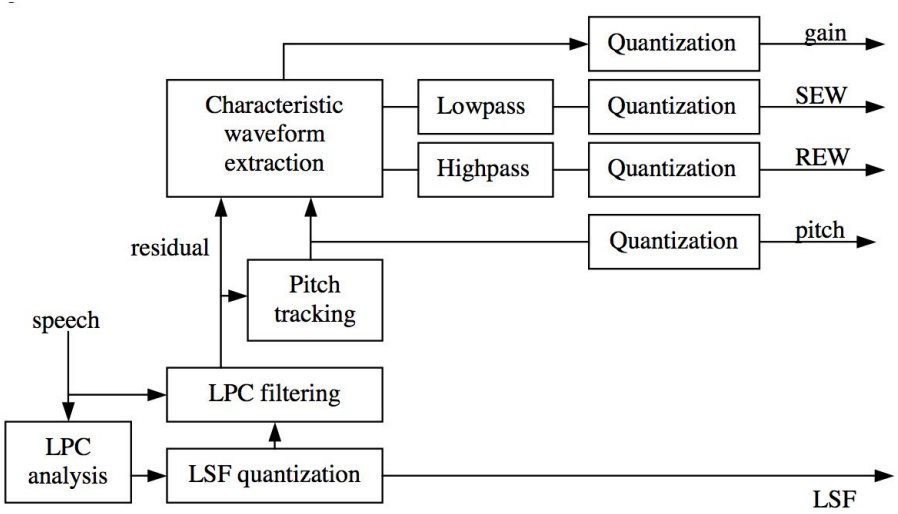

Figure 7.11 Block diagram of the WI encoder.

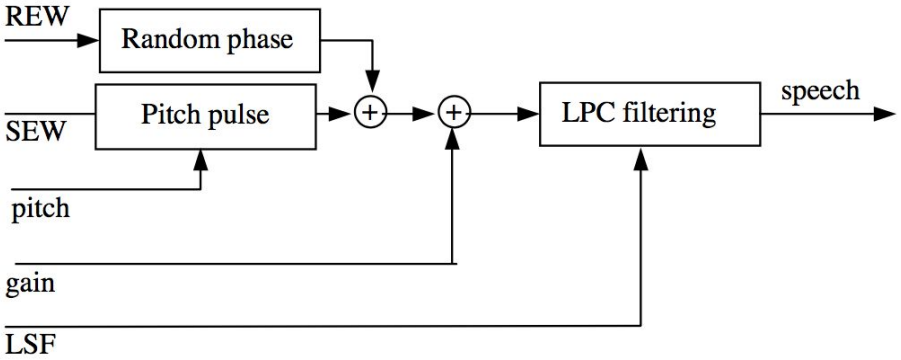

Figure 7.12 Block diagram of the WI decoder.

끝