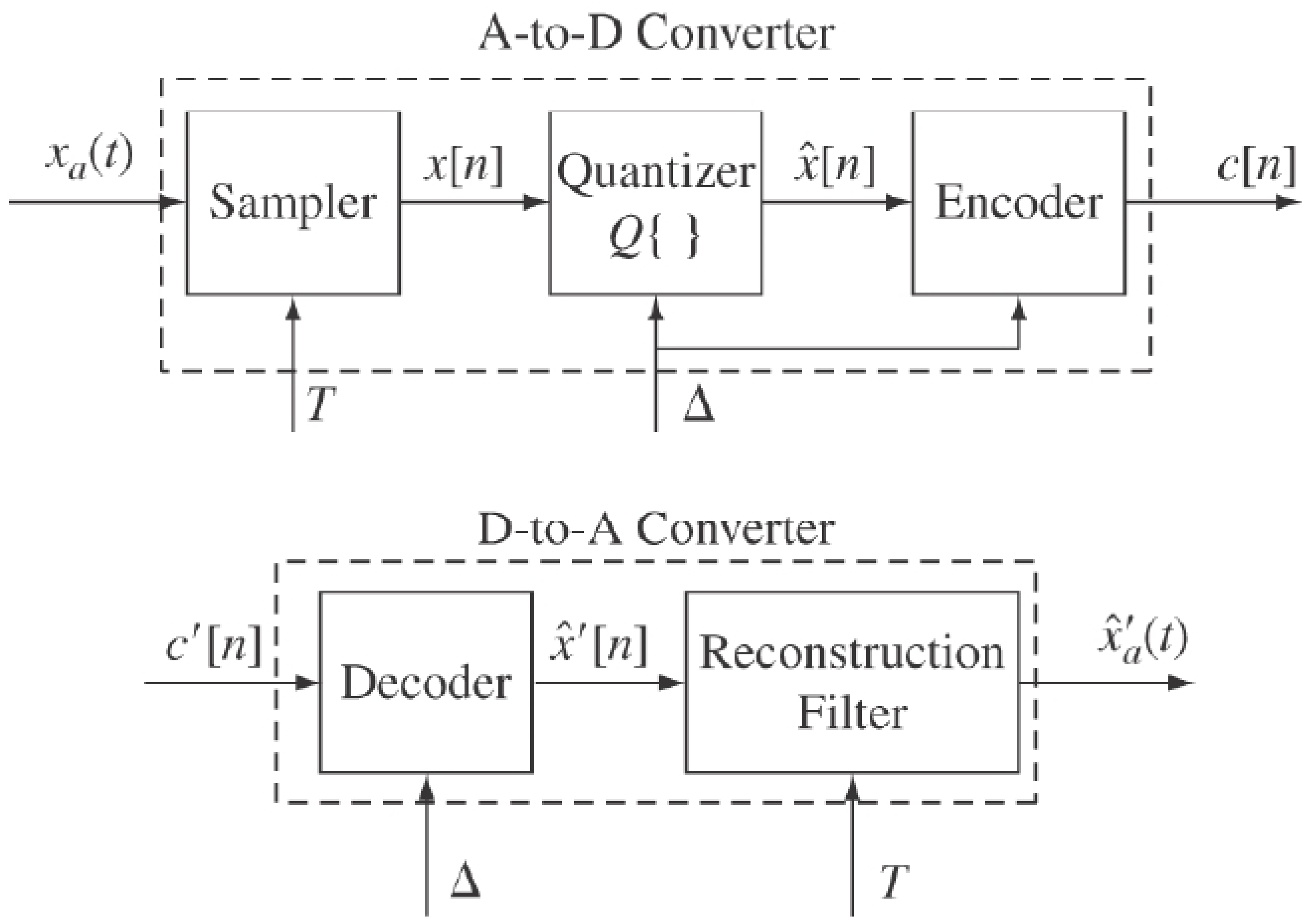
Chapter 11

Digital Coding of Speech Signal
语音信号数字编码

Introduction

Analog-to-Digital Conversion (Sampling and Quantization)



Class of "waveform coders" can be represented in this manner

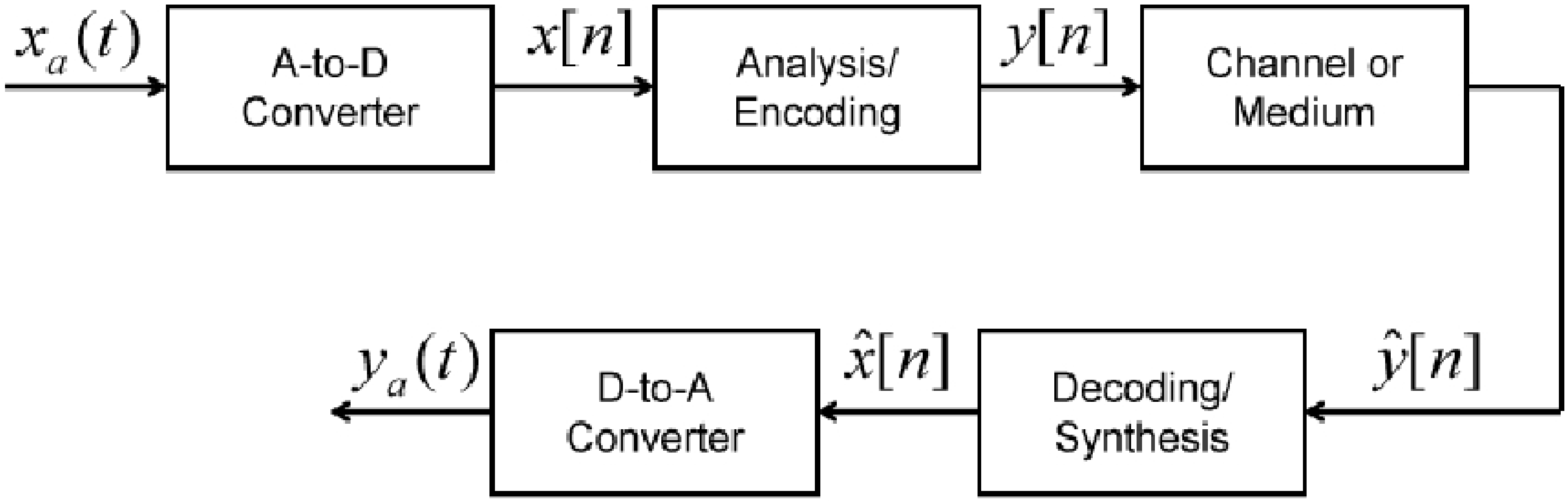
Information Rate

- Waveform coder information rate, I_w , of the digital representation of the signal, $x_a(t)$, defined as

$$I_w = B \cdot F_S = B / T$$

where *B* is the number of bits used to represent each sample and F_S=1/*T* is the number of sample/second

Speech Analysis/Synthesis Systems



- Second class of digital speech coding systems:
  - analysis/synthesis systems
  - model-based systems
  - hybrid coders
  - vocoder (voice coder) systems
- Detailed waveform properties generally not preserved
  - coder estimates parameters of a model for speech production
  - coder tries to preserve intelligibility and quality of reproduction from the digital representation

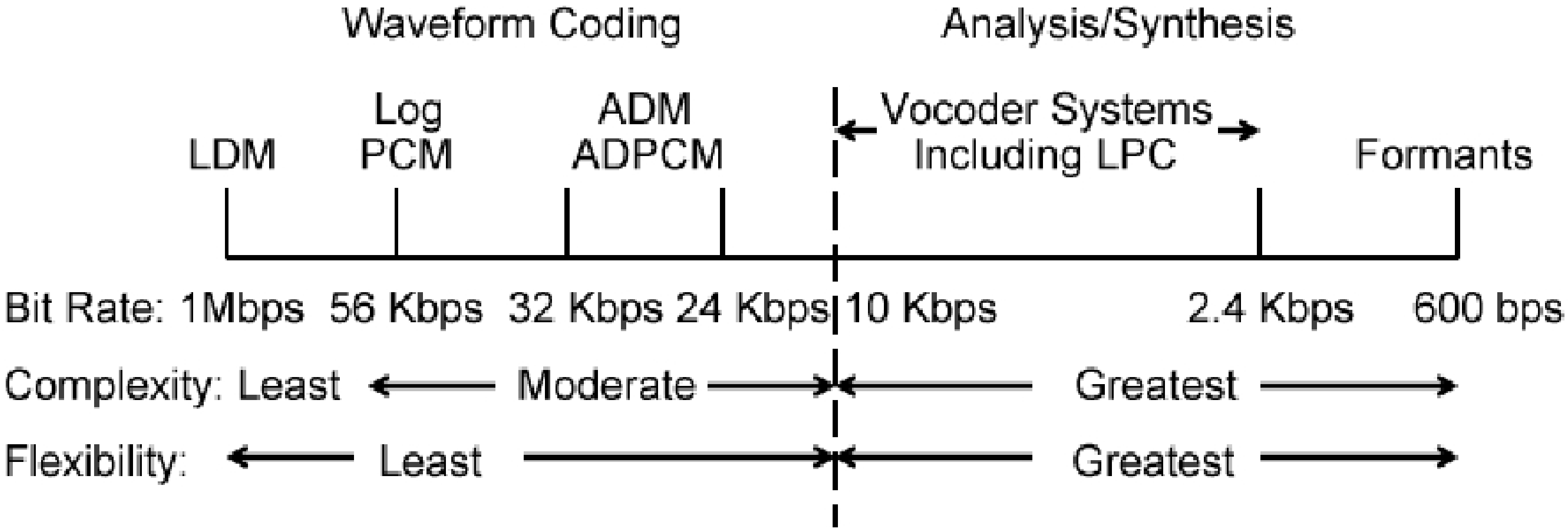
Speech Analysis/Synthesis Systems

- Speech parameters (the chosen representation) are encoded for transmission or storage
  - analysis and encoding gives a data parameter vector
  - data parameter vector computed at a sampling rate much lower than the signal sampling rate
  - denote the "frame rate" of the analysis as F_{fr}
  - total information rate for model-based coders is:

$$I_m = B_c \cdot F_{fr}$$

  - where B_c is the total number of bits required to represent the parameter vector

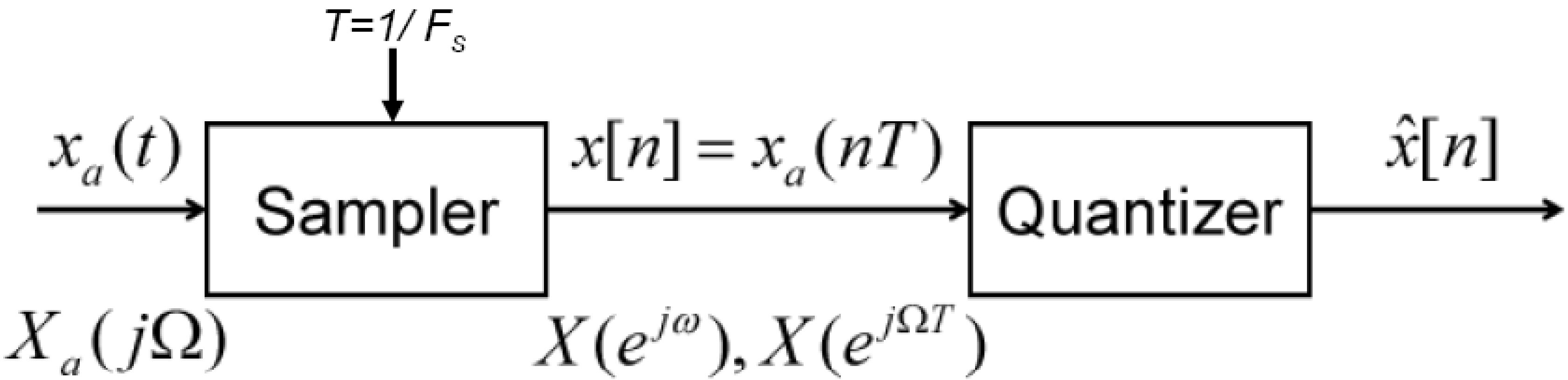
Speech Coder Comparisons



- waveform coders characterized by:
  - high bit rates (24 Kbps – 1 Mbps)
  - low complexity
  - low flexibility
- analysis/synthesis systems characterized by:
  - low bit rates (10 Kbps – 600 bps)
  - high complexity
  - great flexibility (e.g., time expansion/compression)

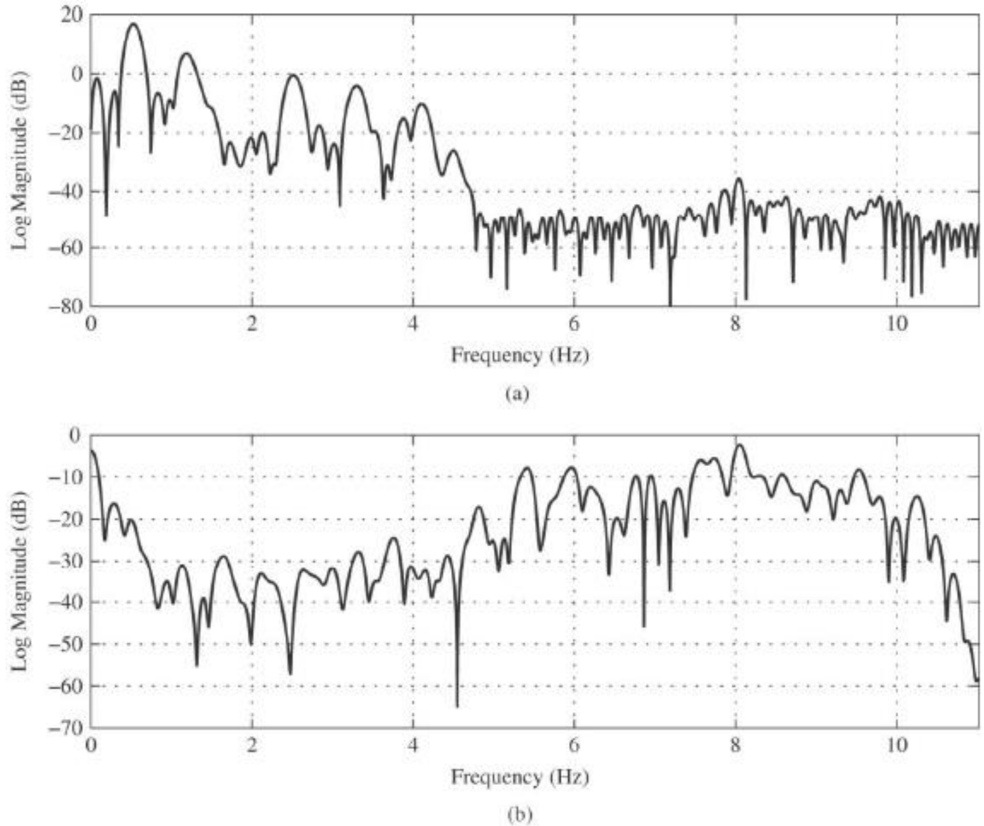
Sampling Speech Signals

Sampling Speech Signals



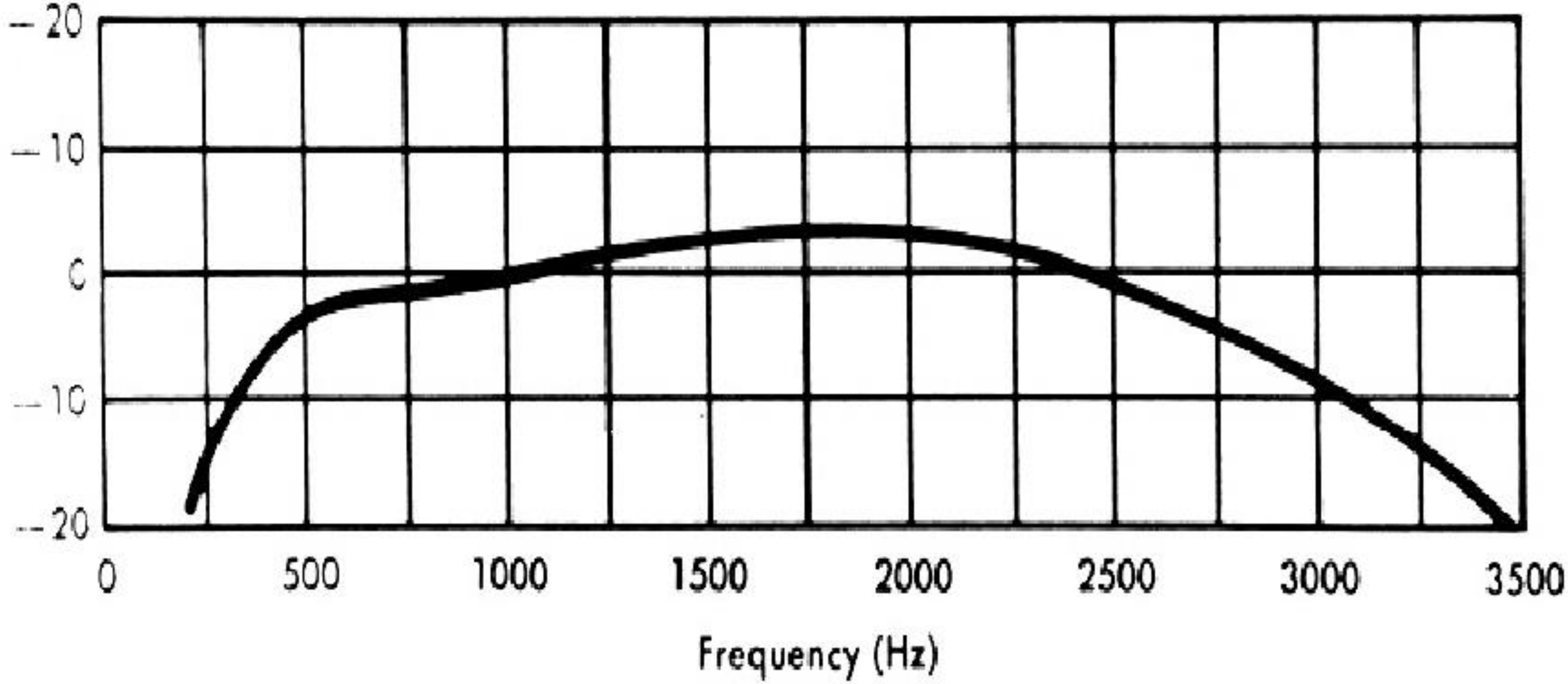
- to perfectly recover $x_a(t)$ from the set of digital samples (as yet unquantized) we require that $Fs = 1/T >$ twice the highest frequency in the input signal
- this implies that $x_a(t)$ must first be lowpass filtered since speech is not inherently lowpass
  - for telephone bandwidth the frequency range of interest is 200-3200 Hz (filtering range) => Fs = 6400 Hz, 8000 Hz
  - for wideband speech the frequency range of interest is 100-7000Hz (filtering range) => Fs = 16000 Hz

Sampling Speech Sounds



- notice high frequency components of vowels and fricatives (up to 10 kHz) =>need *Fs* > 20 kHz
  - need only about 4 kHz to estimate formant frequencies
  - need only about 3.2 kHz for telephone speech coding

Telephone Channel Response



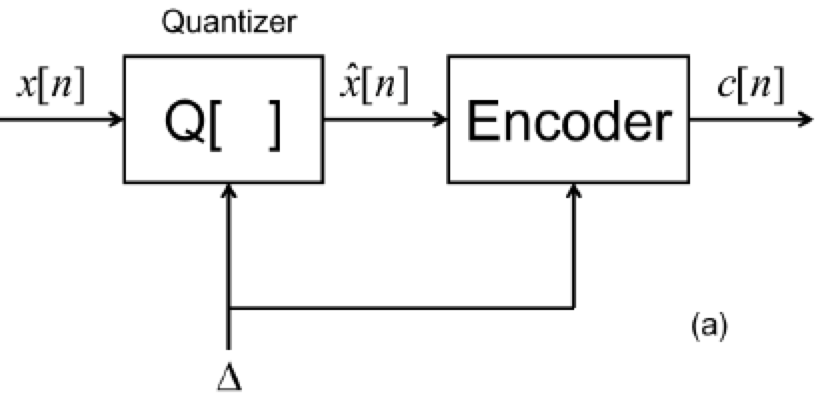
it is clear that 4 kHz bandwidth is sufficient for most applications using telephone speech because of inherent channel band limitations from the transmission path

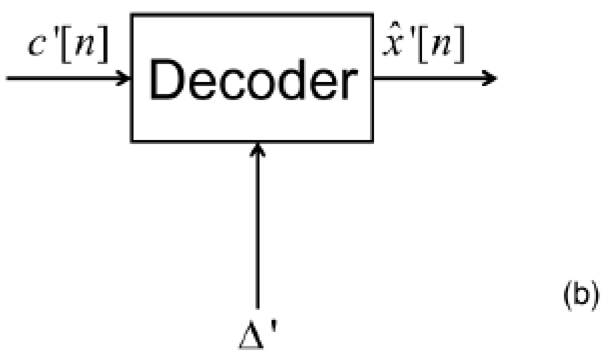
Waveform Coding

Instantaneous Quantization

- separating the processes of sampling and quantization
- assume $x(n)$ obtained by sampling a bandlimited signal at a rate at or above the Nyquist rate
- assume $x(n)$ is known to infinite precision in amplitude
- need to quantize $x(n)$ in some suitable manner

Quantization and Coding





assume Δ’=Δ

- Coding is a two-stage process
  - quantization process:
    $$x(n) \rightarrow \hat{x}(n)$$
  - encoding process:
    $$\hat{x}(n) \rightarrow c(n)$$
    where Δ is the (assumed fixed) quantization step size
- Decoding is a single-stage process
  - decoding process:
    $$c'(n) \rightarrow \hat{x}'(n)$$
- if $c'(n) = c(n)$, then $\hat{x}'(n) = \hat{x}(n)$
- $\hat{x}'(n) \neq x(n)$ => quantization loses information

B-bit Quantization

- use B-bit binary numbers to represent the quantized samples => 2^B quantization levels
- Information Rate of Coder: $I=B\ F_S$= total bit rate in bits/second
  - $B=16$, $F_S= 8$ kHz => $I=128$ Kbps
  - $B=8$, $F_S= 8$ kHz => $I=64$ Kbps
  - $B=4$, $F_S= 8$ kHz => $I=32$ Kbps
- goal of waveform coding is to get the highest quality at a fixed value of I (Kbps), or equivalently to get the lowest value of I for a fixed quality
- since F_S is fixed, need most efficient quantization methods to minimize *the errors between the reconstructed wave and the original wave or increase the* SNR for quantized speech

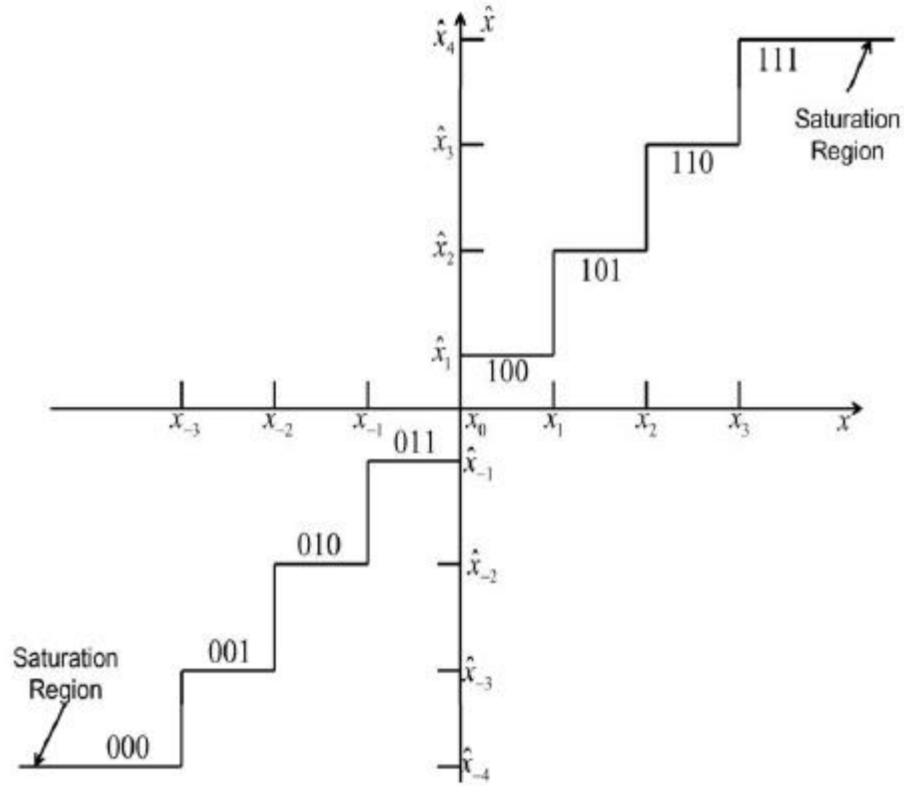
Quantization Process

- quantization => dividing amplitude range into a finite set of ranges and assigning the same amplitude value to all samples in a given range

$$0 = x_0 < x(n) \le x_1 \Rightarrow \hat{x}_1 \ (100)$$
$$x_1 < x(n) \le x_2 \Rightarrow \hat{x}_2 \ (101)$$
$$x_2 < x(n) \le x_3 \Rightarrow \hat{x}_3 \ (110)$$
$$x_3 < x(n) < \infty \Rightarrow \hat{x}_4 \ (111)$$
$$x_{-1} < x(n) \le x_0 = 0 \Rightarrow \hat{x}_{-1} \ (011)$$
$$x_{-2} < x(n) \le x_{-1} \Rightarrow \hat{x}_{-2} \ (010)$$
$$x_{-3} < x(n) \le x_{-2} \Rightarrow \hat{x}_{-3} \ (001)$$
$$-\infty < x(n) \le x_{-3} \Rightarrow \hat{x}_{-4} \ (000)$$

<u>range</u> <u>level</u> <u>codeword</u>

3-bit quantizer => 8 levels



Uniform Quantization

Uniform Quantization

- choice of quantization ranges and levels so that signal can easily be processed digitally

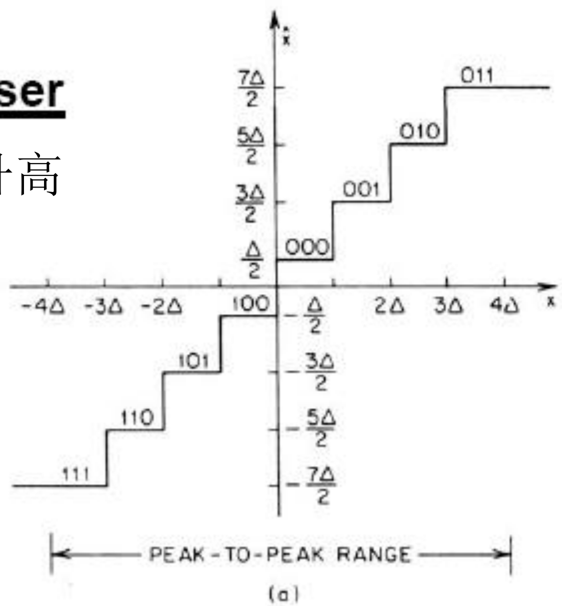
$$x_i - x_{i-1} = \Delta$$

$$\hat{x}_i - \hat{x}_{i-1} = \Delta$$

$$\Delta = \text{quantization step size}$$

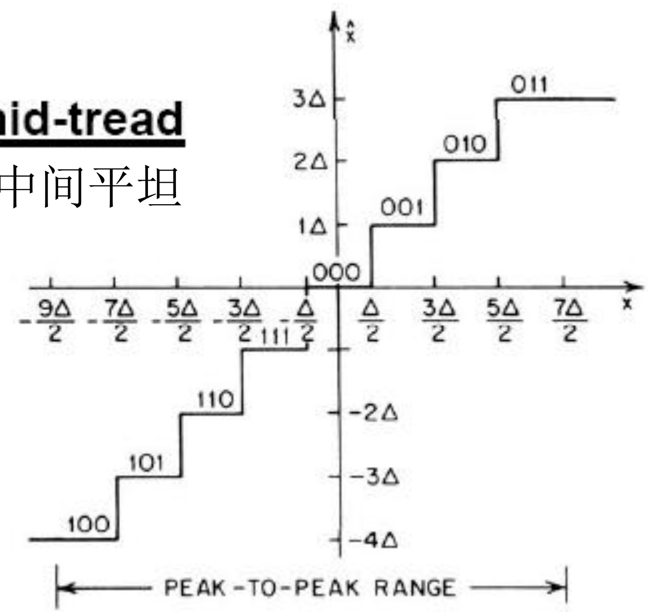
mid-riser

中间升高

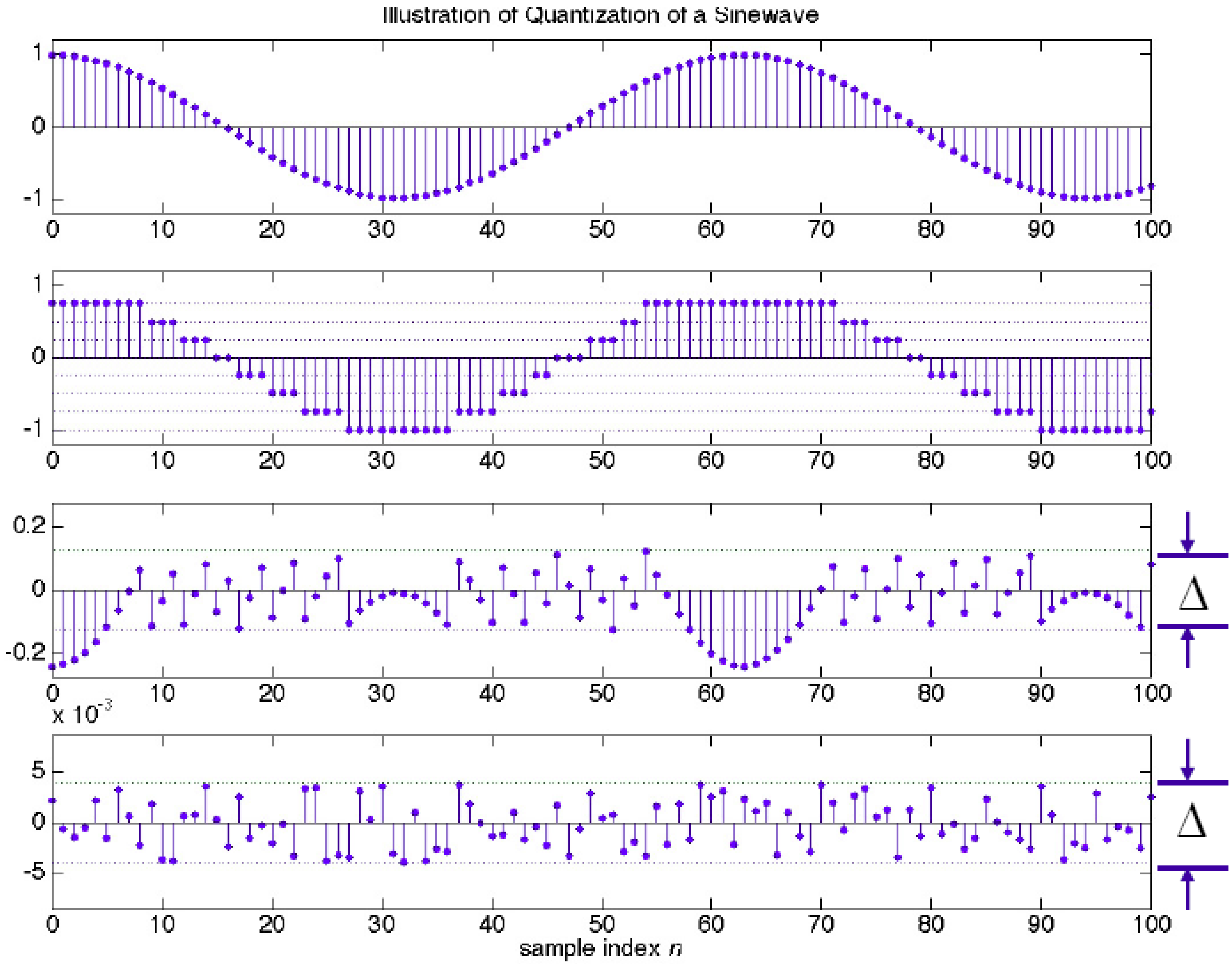


mid-tread

中间平坦



Quantization of a Sine Wave



Unquantized sinewave

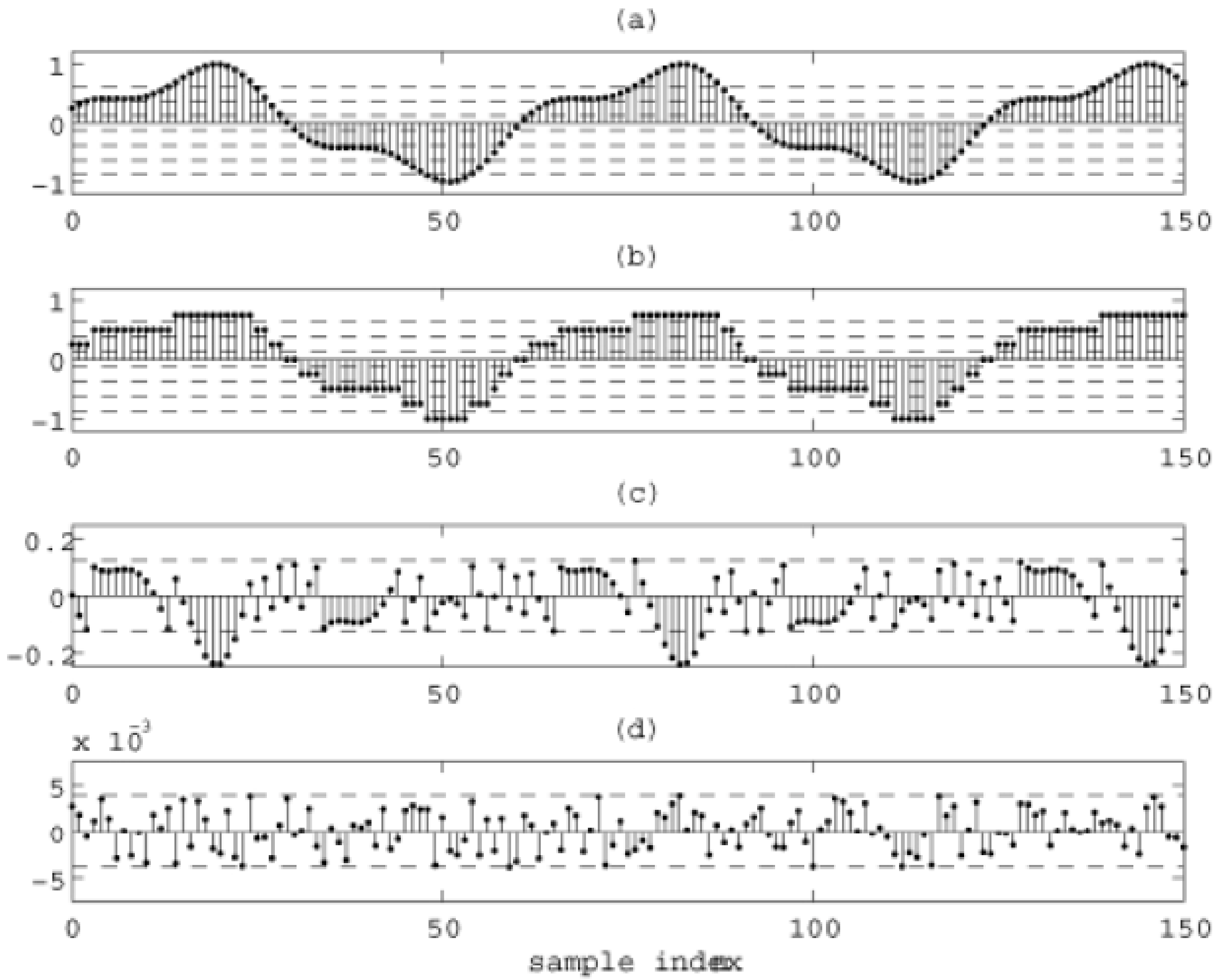
3-bit quantization waveform

3-bit quantization error

8-bit quantization error

Quantization of Complex Signal

$$x[n] = \sin(0.1n) + 0.3\cos(0.3n)$$



Unquantized waveform

3-bit quantization waveform

3-bit quantization error

8-bit quantization error

Uniform Quantizers

- Uniform Quantizers characterized by:
  - number of levels—2^B (B bits)
  - quantization step size-Δ
- if $|x(n)| \leq X_{max}$ and $x(n)$ is a symmetric density, then

$$\Delta 2^B = 2\, X_{max}$$
$$\Delta = 2\, X_{max} / 2^B$$

- if we let

$$\hat{x}(n) = x(n) + e(n)$$

with $x(n)$ the unquantized speech sample, and the $e(n)$ quantization error (noise), then

$$-\frac{\Delta}{2} \leq e(n) \leq \frac{\Delta}{2}$$

(except for last quantization level which can exceed *Xmax* and thus the error can exceed $\Delta/2$)

SNR for Quantization

- can determine SNR for quantized speech as

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\left(x^2(n)\right)}{E\left(e^2(n)\right)} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)}$$

$$\Delta = \frac{2X_{\max}}{2^B} \text{ (uniform quantizer step size)}$$

- assume $p(e) = \frac{1}{\Delta} \quad -\frac{\Delta}{2} \le e \le \frac{\Delta}{2}$ (uniform distribution)

$$= 0 \quad otherwise$$

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

SNR for Quantization

- can determine SNR for quantized speech as

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\left(x^2(n)\right)}{E\left(e^2(n)\right)} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)}$$

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

$$SNR = \frac{(3)2^{2B}}{\left[\frac{X_{\max}}{\sigma_x}\right]^2}; \; SNR(dB) = 10\log_{10}\left[\frac{\sigma_x^2}{\sigma_e^2}\right] = 6B + 4.77 - 20\log_{10}\left[\frac{X_{\max}}{\sigma_x}\right]$$

- If we choose $X_{\max} = 4\sigma_x$, then $SNR = 6B - 7.2$

$$B = 16, \quad SNR = 88.8 \; dB$$

$$B = 8, \quad SNR = 40.8 \; dB$$

$$B = 3, \quad SNR = 10.8 \; dB$$

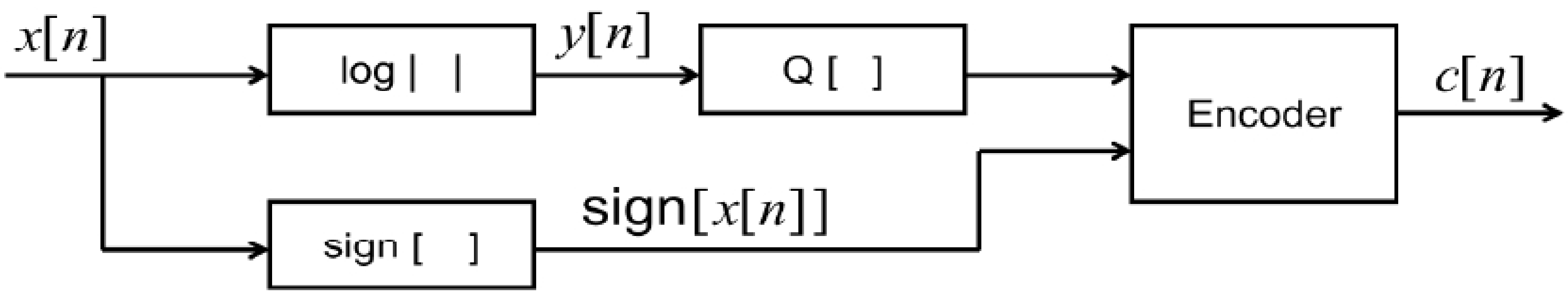
Uniform Quantizer SNR Issues

- to get an SNR of at least 30 dB, need at least $B \geq 6$ bits (assuming $X_{max}=4\ \sigma_x$) $SNR = 6B - 7.2$
  - this assumption is weak across speakers and different transmission environments since σ_x varies so much (order of 40 dB) across sounds, speakers, and input conditions
  - SNR(dB) predictions can be off by significant amounts if full quantizer range is not used; e.g., for unvoiced segments => need more than 6 bits for real communication systems, more like 11-12 bits
  - need a quantizing system where the SNR is independent of the signal level => need non-uniform quantization

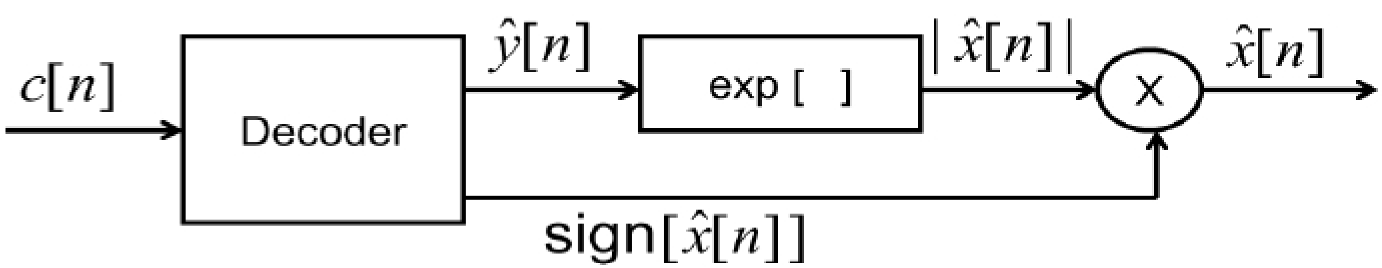
Instantaneous Companding

Instantaneous Companding (Compression/Expansion)

- Non-uniform quantization
  - quantize logarithm of input signal rather than input signal itself



(a)



(b)

Logarithmic Quantizer

$$y(n) = \ln|x(n)|$$

$$x(n) = \exp[y(n)] \cdot sign[x(n)]$$

- where $\quad sign[x(n)] = +1 \quad x(n) \geq 0$

$$= -1 \quad x(n) < 0$$

- the quantized log magnitude is

$$\hat{y}(n) = Q[\log|x(n)|]$$

$$= \log|x(n)| + \varepsilon(n) \quad \underline{\text{new error signal}}$$

Logarithmic Quantizer

- assume that $\varepsilon(n)$ is independent of $\log|x(n)|$. The inverse is

$$\begin{aligned}\hat{x}(n) &= \exp\left[\hat{y}(n)\right]\cdot sign\left[x(n)\right]\\ &=|x(n)|\cdot sign\left[x(n)\right]\exp\left[\varepsilon(n)\right]\\ &= x(n)\cdot\exp\left[\varepsilon(n)\right]\end{aligned}$$

- assume $\varepsilon(n)$ is small, then $\exp\left[\varepsilon(n)\right]\approx 1+\varepsilon(n)+\dots$

$$\hat{x}(n) = x(n)\left[1+\varepsilon(n)\right] = x(n)+\varepsilon(n)x(n) = x(n)+f(n)$$

- since we assume $x(n)$ and $\varepsilon(n)$ are independent, then

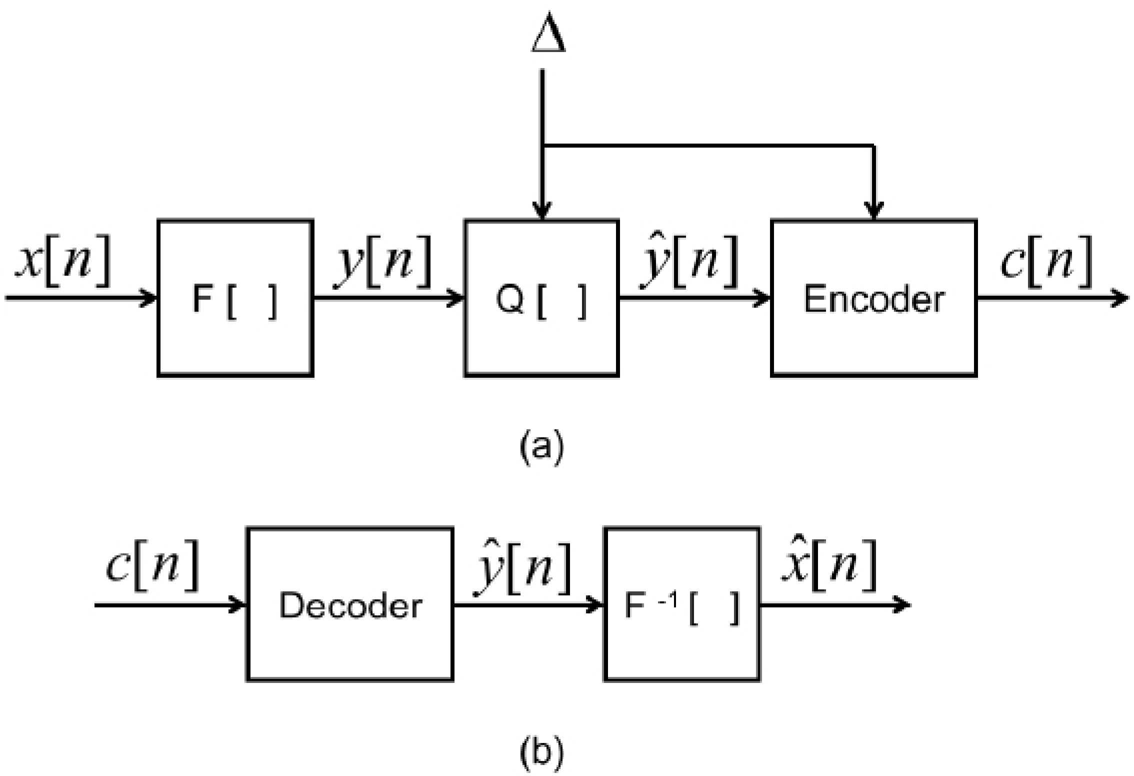
$$\sigma_f^2 = \sigma_x^2\cdot\sigma_\varepsilon^2$$

$$SNR = \frac{\sigma_x^2}{\sigma_f^2} = \frac{1}{\sigma_\varepsilon^2}$$

- SNR is independent of σ_x^2, it depends only on stepsize

Pseudo-Logarithmic Compression

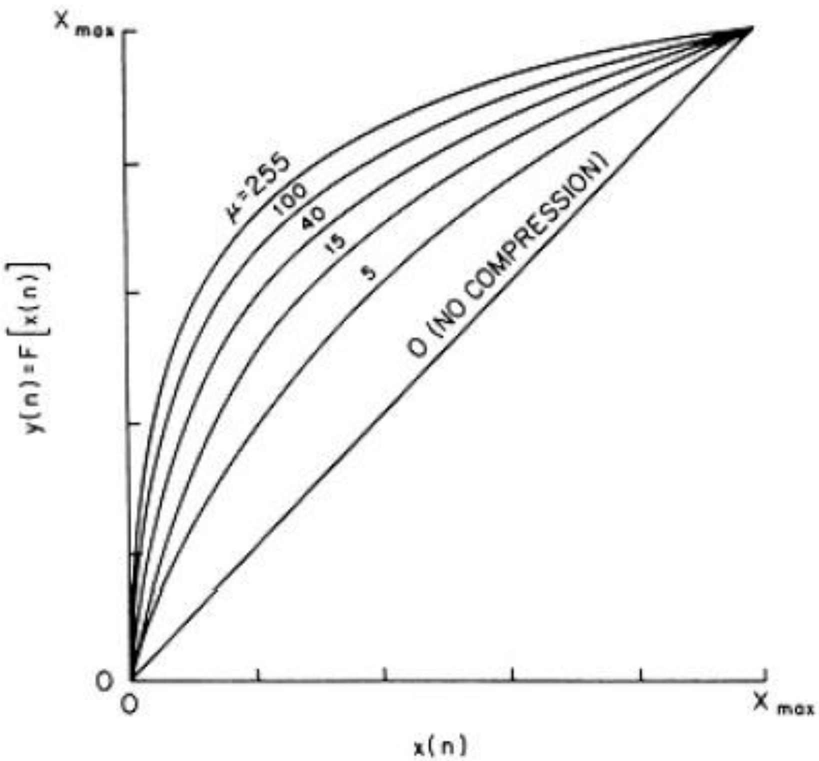
- unfortunately true logarithmic compression is not practical, since the dynamic range (ratio between the largest and smallest values) is infinite => need an infinite number of quantization levels
- need an approximation to logarithmic compression => μ-law/A-law compression



μ- law Compression

$$y(n) = F[x(n)]$$

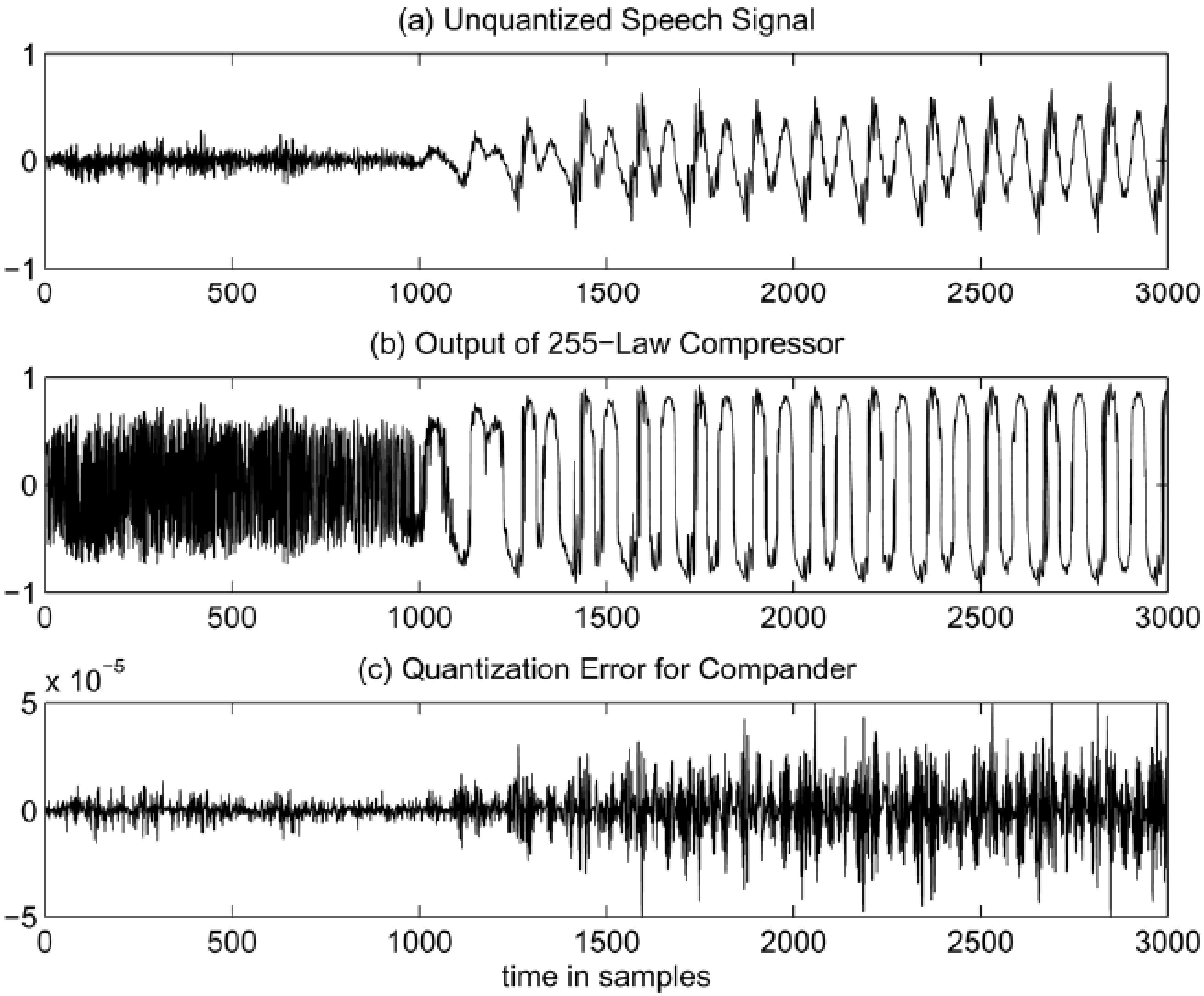
$$= X_{\max} \frac{\log\left[1 + \mu \frac{|x(n)|}{X_{\max}}\right]}{\log(1+\mu)} \cdot sign[x(n)]$$



- when $x(n) = 0 \Rightarrow y(n) = 0$
- when $\mu = 0$, $y(n) = x(n) \Rightarrow$ linear compression
- when μ is large, and for large $|x(n)|$

$$|y(n)| \approx \frac{X_{\max}}{\log \mu} \cdot \log\left[\frac{\mu |x(n)|}{X_{\max}}\right]$$

μ-Law Companding



Mu-law compressed signal utilizes almost the full dynamic range (±1) much more effectively than the original speech signal

SNR for μ-law Quantizer

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

- 6*B* dependence on *B* => good
- for large μ, SNR is less sensitive to changes in $\frac{X_{\max}}{\sigma_x}$ => good
- μ-law quantizer used in wireline telephony for more than 40 years

ITU-T G.711 Standard

- μ-law
  - 8bit, 8kHz
  - 64kbps log-PCM
- A-law
  - Compression function

$$F(x) = \operatorname{sgn}(x) \begin{cases} \frac{A|x|}{1+\ln(A)}, & |x| < \frac{1}{A} \\ \frac{1+\ln(A|x|)}{1+\ln(A)}, & \frac{1}{A} \le |x| \le 1, \end{cases}$$

  - A-law with A = 87.56 is similar to μ-law with μ = 255

Demos

- 8kHz 16-bit linear PCM
- 8kHz 8-bit linear PCM
- 8kHz 8-bit μ -law PCM
- 8kHz 8-bit A-law PCM