Lecture1 Introduction

2102874 Speech Processing

Dr Nisachon Tangsangiumvisai

Course Information

Lecturer **Dr Nisachon Tangsangiumvisai**
ดร. นิศาชล ตั้งเสงี่ยมวิสัย

Contact **office 02-2186909**
E-mail : Nisachon.T@Chula.ac.th
webpage : www.ee.eng.chula.ac.th/~ntang

Course Information (II)

Time	**Wednesday 9.00 – 12.00 pm**	
Schedule	**as given in the course outline**	
Evaluation	**Assignments**	**30 %**
	Mid-term Examination	**40 %**
	Project	**30 %**

Books

- **Speech and Audio Signal Processing : Processing and perception of speech and music**
  B. Gold and N. Morgan
  Wiley, 2000.

- **Digital Speech : Coding for low bit rate communication systems**
  A. M. Kondoz
  Wiley, 2000.

Course Contents

Part I : Speech Processing

- **Fundamentals of the course**
- **Speech Modeling**
- **Speech Analysis and Synthesis**
- **Speech Coding**
- **Human Recognition**
- **Speaker Verification**
- **Text-to-Speech Synthesis**
- **Speech Prosody**

Course Contents (II)

Part II : Speech Enhancement

- **Noise Reduction Techniques**

Overview

Speech Communication

- **What is speech communication ?**
  The transfer of information from one person to another via *speech*, which consists of variations in pressure coming from the mouth of a speaker.

- **What is sound ?**
  Sound is an acoustic *wave* that results when a vibrating source (e.g vocal cords) disturbs an elastic medium (e.g air).

Speech Communication (II)

- When a sound wave reaches a listener's ear drum, the vibrations are transmitted to the inner ear (or cochlea), where mechanical displacements are converted to neural pulses that are sent to the brain and result in the sensation of sound.
- ***Speech chain*** consists of speech mechanism in the speaker, transmission through medium, and a speech perception process in the ear and brain of the listener.
- In many applications of speech processing, part of the speech chain is implemented by a simulation device.

Speech Processing

- **Speech Analysis**
  - Speech Recognition
  - Speaker Verification
  - Speaker Identification
- **Speech Synthesis**
  - Text-to-speech (TTS)
- **Speech Codec**

Factors

- **Analogue** form of Speech for Telecommunication
  - transmission power
  - spectral utilization
- **Digital** Transmission of Speech
  - low cost
  - consistent quality
  - security
  - spectral efficiency

Digitization of Speech Signals

- **Sampling :** Nyquist Criterion
- **Quantization :** Number of quantizer levels is proportional to reconstruction at the receiver
- **Speech Transmission Systems**
  In mobile communication systems
  ⇨ for reduction in PCM bit rate
  (64 kbps; restricts the desired spectral efficiency)

Digitization of Speech Signals (II)

- **Digital Coding of Speech Signals :**
  The parameters for speech production (sensitive to corruption) are encoded and transmitted through degraded channels.
  → problem in maintaining speech quality
- **Bit Rate Reduction :**
  - difficult to maintain speech quality as the bit rate falls
  - the algorithm becomes complex; not practical for real-time implementation
  - results in excessive delay → echo control problem

Digitization of Speech Signals (III)

- **With the use of DSP chips (faster & more reliable)**
  → possible for real-time implementation of highly complex algorithms

- **Digitally encoded Speech :**
  - condense down to a binary sequence
  - advantages of digital systems can be exploited
  (flexibility, security, integration into Integrated Service Digital Networks (ISDN))
  BUT extra BW for transmission → signal compression

Speech Coding

- **Low bit rate (LBR) speech coders**

speech information → compressed into few parameters → coded for transmission → some degraded channels → distortion in the recovered speech

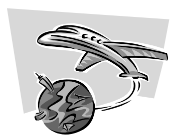
- To identify the sensitivity of each speech parameters for error control at the decoder (For power efficiency : only the most sensitive bits are **error protected**)

Audio Coding for Wireless Communication

Criteria to be considered :

- High quality digital voice
- Bit rate as low as 4 kbps (due to low-rate applications e.g. cellular phones)
- Low delay speech coding at 16 kbps (standard CCITT G.728)
- Implementational Complexity

Let's Begin!



Fundamentals of the course

- Digital Signals
- Spectrum Analysis of Digital Signals
  - *Fourier Transform*
  - *Z-transform*
- Digital Systems
- Digital Filters

Digital Signals

continuous-time signal, $x_a(t)$ → ADC → discrete-time signal, $x(n)$

Analog-to-Digital Converter (ADC) performs :
- *sampling* of the amplitudes of the analog signal $x_a(t)$ on an equidistant grid along the horizontal time axis

- *quantization* of the amplitudes to fixed samples represented by numbers $x(n)$ along the vertical amplitude axis

Sampling

sampling time

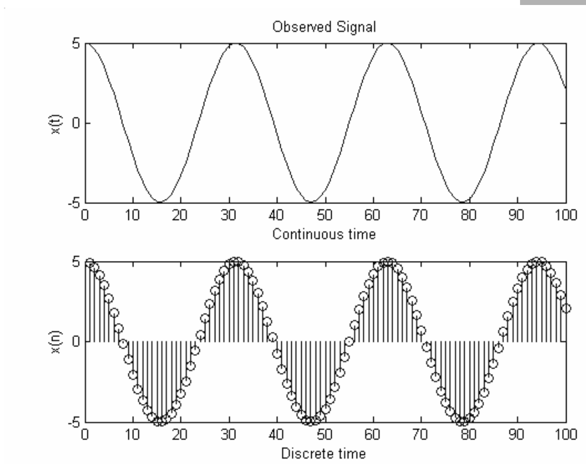
$$x(n) = x_a(n\boxed{T}), \quad -\infty < n < \infty \qquad \text{...(1)}$$

analogue signal

Integer (discrete time index)

To avoid aliasing
- input has to be band-limited
- sampling frequency must satisfy the nyquist criterion

Example of signals



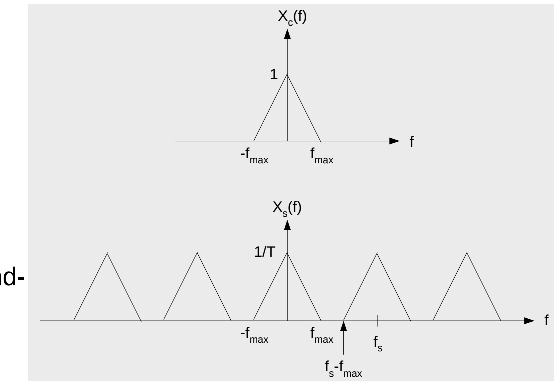
Sampling (II)

$f_s \geq 2 \quad f_{\max}$ ...(2)

$\left(T \leq \frac{1}{2} \quad T_{\max}\right)$ ...(3)

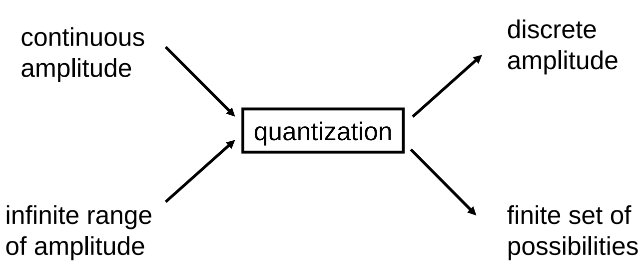
In telecommunication networks, analogue speech signals are band-limited to 300-3400 Hz,

$f_s = 8 \quad kHz$



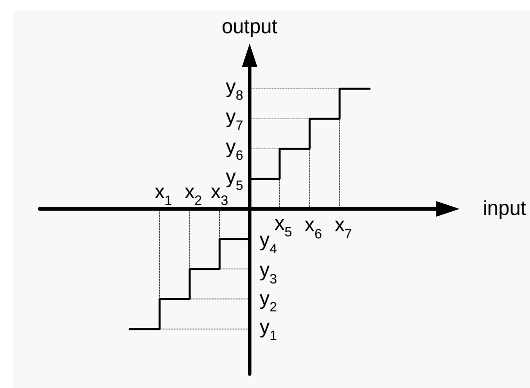
Effect of sampling

Quantization



Scalar Quantization

Each of a set of discrete values is quantized separately.



Characteristics of uniform quantization

Scalar Quantization (II)

- **Quantisation step-size (Δ)**
  : distance between the finite sets of amplitude levels
- **Each Δ_n is represented by a <u>code-word</u> for transmission.**
- **At digital receiver → De-quantizer**
  : indicate which discrete amplitude to be used.
- **Channel transmission bit rate :**

$$\boxed{T_c = B\, f_s} \qquad \text{(bit/sec)} \qquad ...(4)$$

B : number of bits representing the discrete amplitudes. (length of $c(n)$)

Scalar Quantization (III)

- From Eq.(4), if f_s is fixed,
  T_c can be reduced by reducing B
- $\downarrow B \Rightarrow \uparrow \Delta \Rightarrow \downarrow$ quality of reconstructed signal
- In the i^{th} interval,

$$x_i = \frac{\Delta_i}{2} \le x_a(nT) < x_i + \frac{\Delta_i}{2} \qquad ...(5)$$

$x_a(nT)$ is represented by the quantized amplitude.

Scalar Quantization (IV)

- The quantization of the amplitudes to fixed numbers in the range between $-32768 \ldots 32767$ is based on a 16-bit representation of the sample amplitudes.
- It allows 2^{16} quantized values in the range of -2^{15} to $2^{15}-1$.
- Thus, for *B*-bit representation, the number range is $-2^{(B-1)}$ to $2^{(B-1)}-1$.

Scalar Quantization (V)

- Instantaneous squared error is

$$\left(x_a(nT) - x_i\right)^2 \qquad ...(6)$$

- Mean Squared Error of the signal $x = x_a(nT)$

$$E_i^2 = \int_{x_i - \frac{\Delta i}{2}}^{x_i + \frac{\Delta i}{2}} \left(\boxed{x} - x_i\right)^2 p(x)\, dx \qquad ...(7)$$

p.d.f. of x

- Assume fine resolution (small Δ_i), $p(x)$ is flat within $\left[x_i - \frac{\Delta}{2}, x_i + \frac{\Delta}{2}\right] \Rightarrow$ use its center value $p(x_i)$

Scalar Quantization (VI)

- Hence, MSE becomes

$$E_i^2 = \frac{\Delta_i^3}{12} p(x_i) \qquad \text{...(8)}$$

- Prob. of signal falling in this interval is

$$\Gamma_i = \int_{x_i - \frac{\Delta_i}{2}}^{x_i + \frac{\Delta_i}{2}} p(x)dx = p(x_i)\Delta_i \qquad \text{...(9)}$$

- Substituting for $p(x_i)$ in Eq.(8), $E_i^2 = \frac{\Delta_i^2}{12} \Gamma_i$ ...(10)
- Hence, total MSE is $\boxed{E^2 = \frac{1}{12} \sum_{i=1}^{N} \Gamma_i \Delta_i^2}$ ...(11)

(N = no. of levels in the quantizer)

Scalar Quantization (VII)

- Assume uniform quantisation step-size

$$\boxed{E^2 = \frac{\Delta^2}{12}} \qquad \text{...(12)}$$

- For B-bit binary code words

$$N = 2^B \qquad \text{...(13)}$$

- Assuming $|x| \le X_{\max}$ and $p(x)$ is symmetrical

$$2X_{\max} = \Delta 2^B \qquad \text{...(14)}$$

Scalar Quantization (VIII)

- Thus, the step-size can be found from

$$\frac{2X_{\max}}{2^B} = \Delta \qquad \text{...(15)}$$

- Quantization error $e_q(n)$ is bounded by

$$-\tfrac{\Delta}{2} \le e_q(n) \le \tfrac{\Delta}{2} \qquad \text{...(16)}$$

- Uniform quantizer assumes uniform pdf. of constant height $\frac{1}{2X_{\max}}$.

Scalar Quantization (IX)

- Hence, input power is given by

$$P_x = \int_{-X_{\max}}^{X_{\max}} \frac{x^2}{2X_{\max}} p(x)dx \qquad \text{...(17)}$$

- SNR

$$\frac{P_x}{P_n} = \frac{X_{\max}^2/3}{\Delta^2/12} = 2^{2B} \Rightarrow 6.02B \text{ (dB)} \qquad \text{...(18)}$$

Vector Quantization

To characterize the spectrum of a speech signal (**spectral analysis**), consider the following example :

<u>Uncompressed Signal</u>

- 10 kHz sampled speech with 16-bit speech amplitudes
- Information rate is 160,000 bps (required for storage of speech analysis)

<u>Compressed Signal</u>

- spectral vector $\mathbf{v}_l, \quad l = 1, 2, \ldots, L$
- vectors of dimension p=10 using 100 spectral vectors/s.
- by representing each spectral component to 16-bit precision, the required storage is 100x10x16 bps

➡ **10 times reduction**

Vector Quantization (II)

It is needed a **single** spectral representation for each basic speech unit.

➡ A finite number of **unique** spectral vectors.

(each vector corresponds to one of the basic speech units.)

➡ impossible, due to time-varying properties of the spectra of the signal.

➡ build a codebook of **distinct** analysis vectors.

(common techniques for vector quantization (VQ) methods)

Vector Quantization (III)

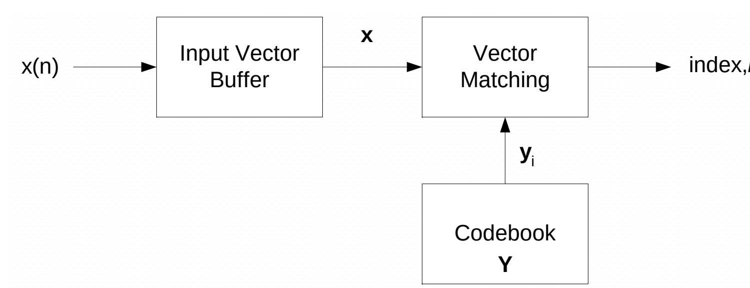
- So, if we require to a codebook with 1024 unique spectral vectors, then we need a 10-bit number.

- Assume a rate of 100 spectral vectors/s, a total bit rate of <u>1 kbps</u> is required to represent the spectral vectors of a speech signal. (this is 1/16 time the rate required by the continuous spectral vectors.)

- So, the VQ method is an **efficient** representation of the spectral information in the speech signal.

Vector Quantization (IV)

- Block Quantization / Pattern Matching Quantization
- Signals are quantized as a single vector.



Block diagram showing a vector quantization

Vector Quantization (V)

Advantages

- reduced storage
- reduced computation for determining similarity of spectral analysis vectors (use of a table lookup)
- discrete representation of speech sounds (use of codebook)

Vector Quantization (VI)

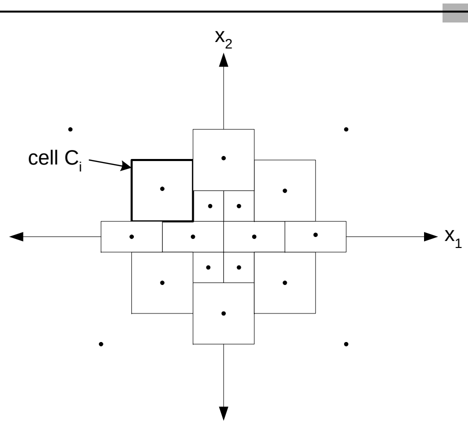
Disadvantages

- since there is a *finite* number of codebook vectors, the **“best”** representation of a given spectral vector is obtained with a certain level (non-zero) of quantization error.

(As the size of the codebook increases, the size of the quantization error decreases.)

- To reduce the quantization error, larger codebook is needed ➡ more storage!!

Vector Quantization (VII)



Partitioning a 2-dimensional space into 18 cells

Comparison

- **Scalar Quantization**
  - different cells have same shape
- **Vector Quantization**
  - different cells have different shapes
  - at very low bit rate, performs better than the scalar method, but at computational and storage costs.

Comparison (II)

Vector quantization systems tend to be less robust to random channel errors than scalar method.

E.g. 10-bit vector quantizer

10 1-bit scalar quantizer

Assume channel error rate that causes 1-bit error, thus

scalar quantizer : 1-bit in error → affect 1 value in 1 dimension

vector quantizer : 1-bit in error → affect 10-bit vector

Signal Reconstruction

discrete-time signal, $x(n)$ → DAC → continuous-time signal, $x_a(t)$

Digital-to-Analog Converter is used for reconstruction of analog signal.

Spectrum Analysis of Digital Signals

- Discrete-time Fourier Transform (DTFT)

$$X(e^{j\omega}) = F\left[x(n)\right] = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \quad \text{... (19)}$$

a function of continuous ω

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{jn\omega}\,d\omega \quad \text{... (20)}$$

Spectrum Analysis (II)

- Discrete Fourier Transform (DFT)

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}, \quad k = 0,1,\ldots,N-1 \quad \text{... (21)}$$

for digital computation

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N}, \quad n = 0,1,\ldots,N-1 \quad \text{... (22)}$$

Spectrum Analysis (III)

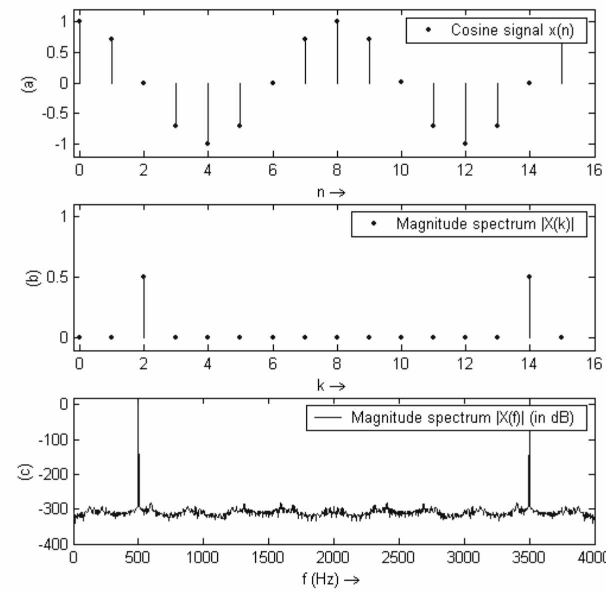
- The fast version of the DFT is called the Fast Fourier Transform (FFT).
  >> Xmag = abs(fft(x,N)); Xphase = angle(fft(x,N));

- From Eq.(21),

$$\boxed{X(k) = X_R(k) + jX_I(k), \quad k = 0,1,\ldots,N-1}$$

Magnitude spectrum : $|X(k)| = \sqrt{X_R^2(k) + X_I^2(k)}$ ... (23)

Phase : $\varphi(k) = \arctan\left(\dfrac{X_I(k)}{X_R(k)}\right)$ ... (24)

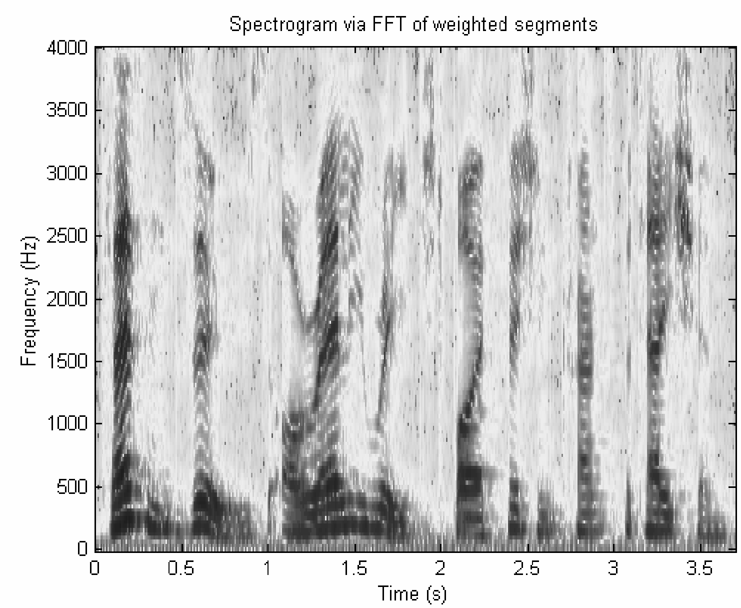


Spectrum Analysis (V)

<u>Time-frequency representation</u> ➔ Spectrogram

- It is an estimate of the short-time, time-localized frequency content of the signal.
- The signal is split into segments of length N
- and are multiplied by a window
- and an FFT is performed
- An overlap of the weighted segment is used to increase the time-localization of the short-time spectra.

>> [B,F,T] = specgram(x,NFFT,Fs,WINDOW,NOVERLAP);



Spectrum Analysis (VII)

- **z-transform**

$$X(z) = Z\left[x(n)\right] \quad \text{... (25)}$$

$$= \sum_{n=-\infty}^{\infty} x(n)\, z^{-n} \quad \text{... (26)}$$

therefore

$$x(n) = \frac{1}{2\pi j}\oint_C X(z)\, z^{n-1} dz \quad \text{... (27)}$$

$$\boxed{X(z)\big|_{z=e^{j\omega}} = X\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n}} \quad \text{... (28)}$$

Digital Systems

$$x(n) \longrightarrow \boxed{T\{\cdot\}} \longrightarrow y(n)$$

$$\boxed{y(n) = T\{x(n)\}} \quad \text{...(29)}$$

Properties of discrete-time systems:

- linearity
- shift-invariance
- causality
- invertibility

Digital Systems (II)

- **Linearity**

$$\mathrm{T}\left\{a x_1(n) + b x_2(n)\right\} = a\,\mathrm{T}\{x_1(n)\} + b\,\mathrm{T}\{x_2(n)\} \quad \text{...(30)}$$

a, b are constants.

- **Shift-invariance**

If $\quad x(n) \rightarrow y(n) \quad$ ...(31)

then $\quad x(n-n_0) \rightarrow y(n-n_0) \quad$ ...(32)

Digital Systems (III)

- **Causality**

For any n_0, the response of the system at time n_0 depends only upon the values of the input for $n < n_0$.

Examples :

$$y(n) = x(n) + x(n-1) \quad \text{...(33)}$$

$$y(n) = x(n) + x(n+1) \quad \text{...(34)}$$

- **Invertibility**

The input to the system maybe uniquely determined by observing the output.

Digital Systems (IV)

- **Invertibility (contd.)**

$$x_1(n) \leftrightarrow y_1(n)$$
$$x_2(n) \leftrightarrow y_2(n)$$

***Examples* :**

when $x_1(n) \neq x_2(n)$ and $y_1(n) \neq y_2(n)$

If $$y(n) = x(n)\,h(n) \quad ...(35)$$

➡ $$x(n) = y(n)/h(n) \quad ...(36)$$

Digital Systems (V)

- Linear-time-invariant (LTI) systems
- Discrete Convolution

  x(*n*) : input signal

  h(*n*) : impulse response of a digital system

  y(*n*) : output signal

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \quad ...(37)$$

```
>> y = conv(x,h);
```


Digital Systems (VI)

- FIR system : a system with a finite impulse response

$$H(z) = \frac{Y(z)}{X(z)} = 1 + b_1 z^{-1} + b_2 z^{-2} \quad ...(38)$$

- IIR system : a system with an infinite impulse response

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \quad ...(39)$$

```
>> y = filter(B,A,x);
>> [H,W] = freqz(B,A,N);
```


Digital Filters

- **Lowpass filter (LPF)**
  - select low frequencies up to the cutoff frequency f_c
  - Attenuate frequencies higher than f_c
- **Highpass filter (HPF)**
  - select frequencies higher than f_c
  - attenuate frequencies below f_c
- **Bandpass filter (BPF)**
  - select frequencies between a lower f_{c1} and a higher f_{c2}
  - attenuate frequencies outside this range

Digital Filters (II)

- **Bandreject filter**
  - attenuate frequencies between a lower f_{c1} and a higher f_{c2}
  - frequencies outside this range are passed
- **Notch filter**
  - attenuate frequencies in a narrow bandwidth around the cutoff frequency f_c
- **Allpass filter**
  - pass all frequencies but modify the phase of the input signal

Summary

- **Introduction to the course**
- **Fundamentals of the course**

- Next Lecture

  <u>Lecture 2</u> : Speech Modeling