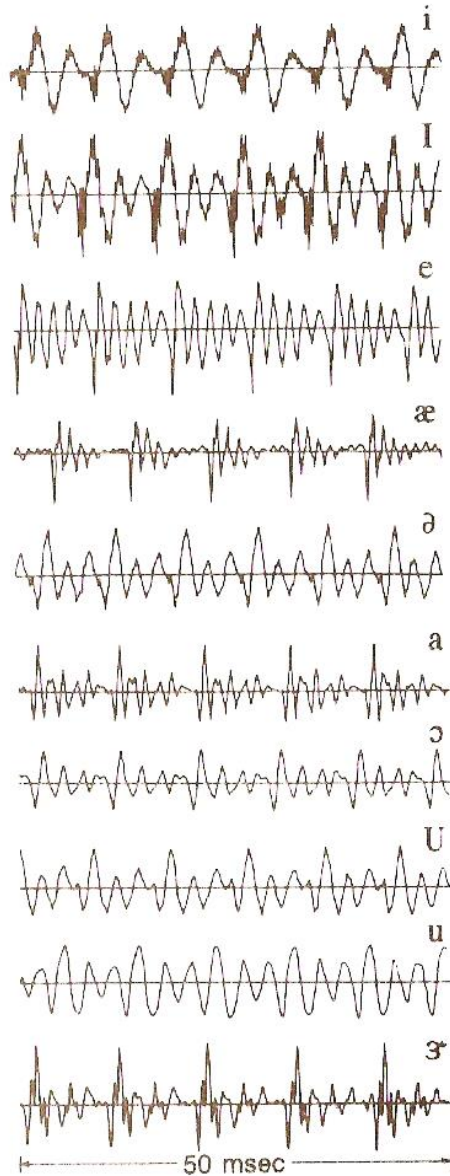


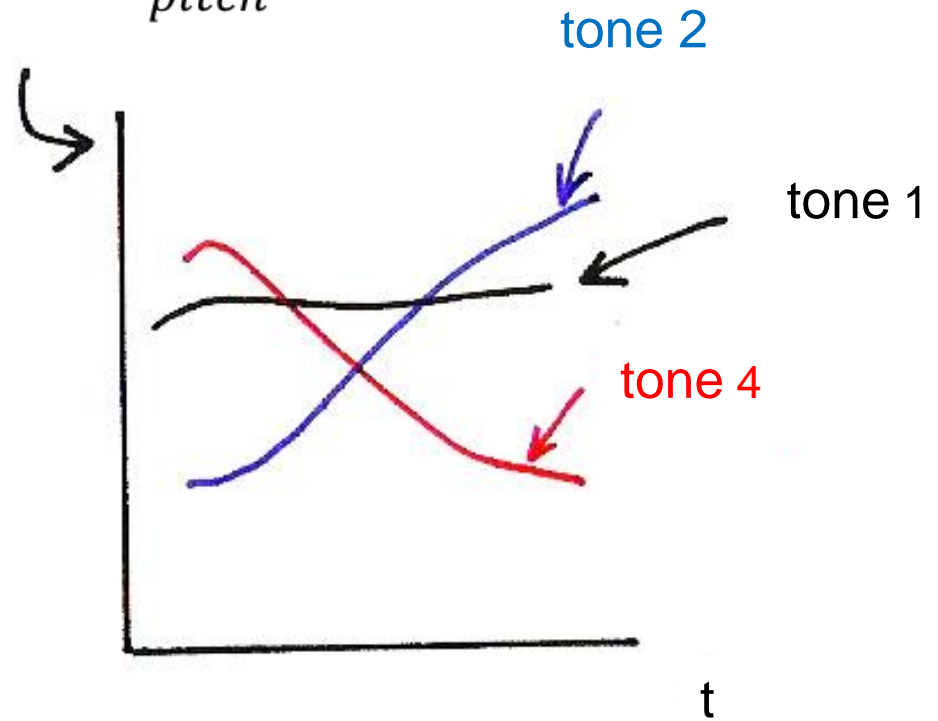
## 7.0 Speech Signals and Front-end Processing

**References:** 1. 3.3, 3.4 of Becchetti  
3. 9.3 of Huang

# Waveform plots of typical vowel sounds - Voiced (濁音)

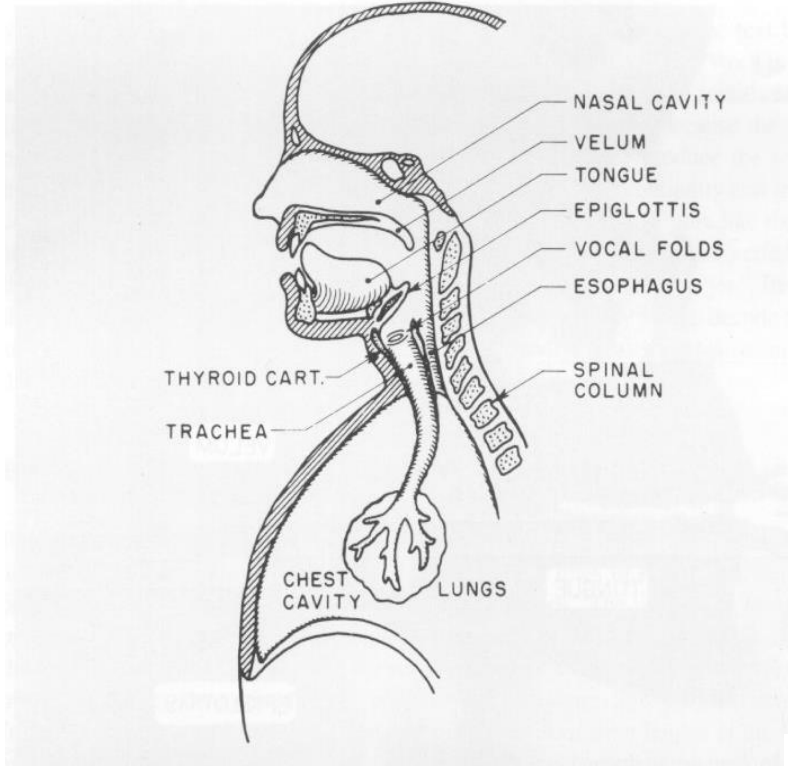


$$F_0 = \frac{1}{pitch} \text{ (音高)}$$

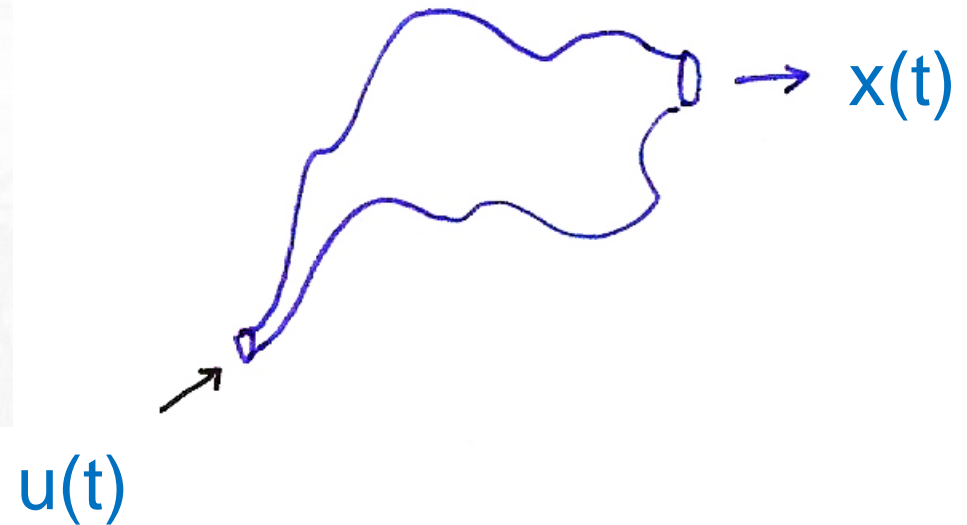
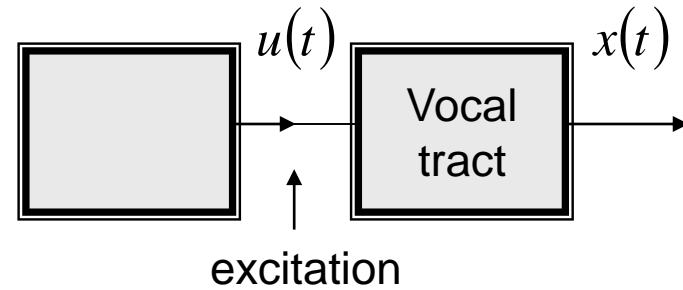


# Speech Production and Source Model

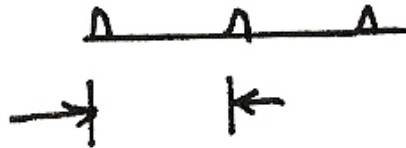
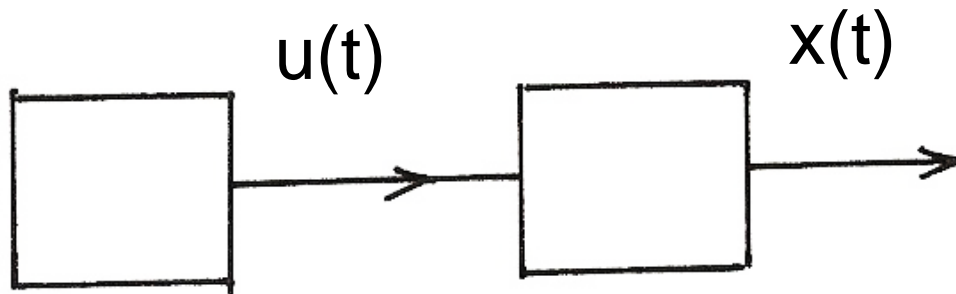
- Human vocal mechanism



- Speech Source Model



# Voiced and Unvoiced Speech



pitch



voiced

pitch

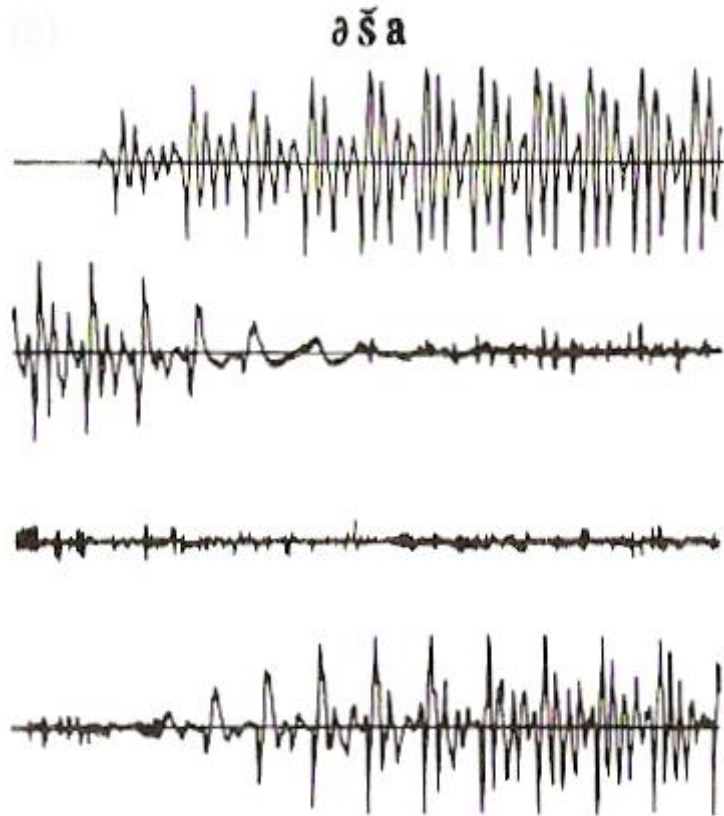


unvoiced

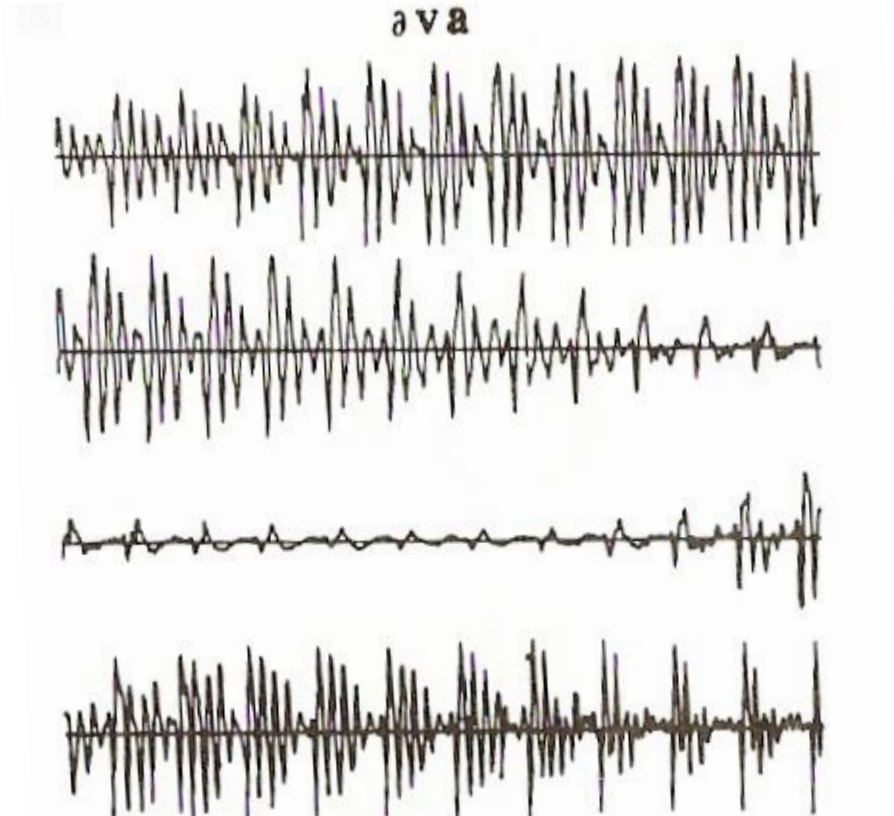


# Waveform plots of typical consonant sounds

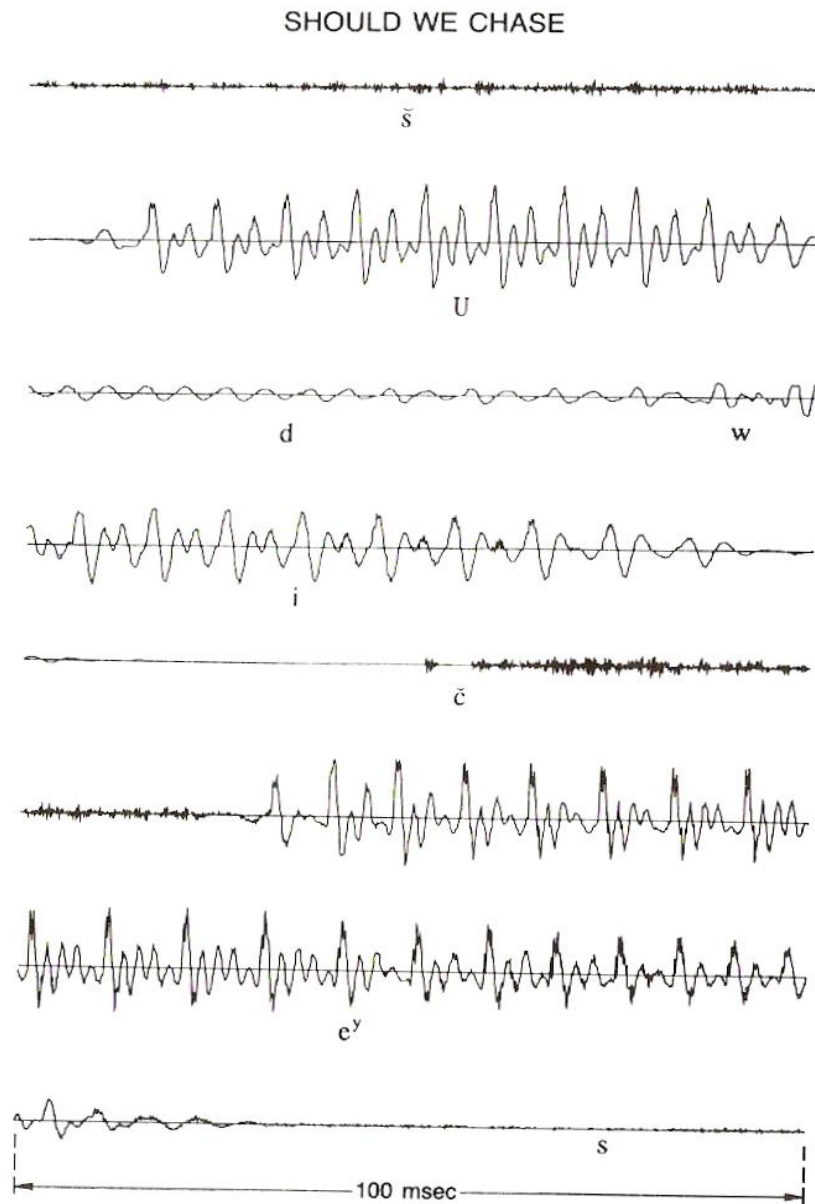
Unvoiced (清音)



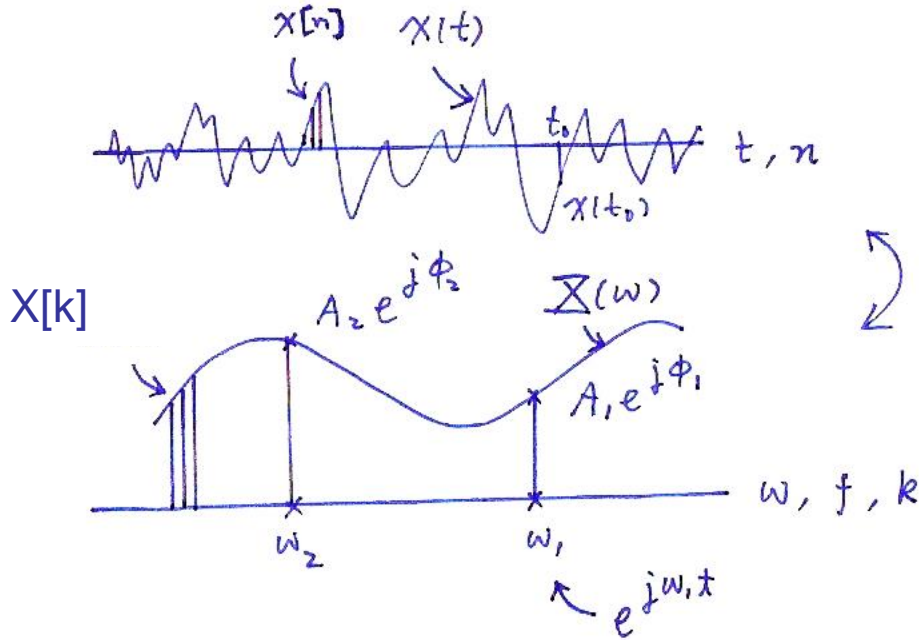
Voiced (濁音)



# Waveform plot of a sentence



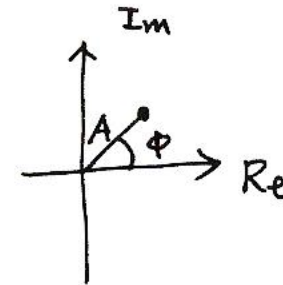
# Time and Frequency Domains (P.12 of 2.0)



time domain

1-1 mapping  
Fourier Transform  
Fast Fourier Transform (FFT)

frequency domain



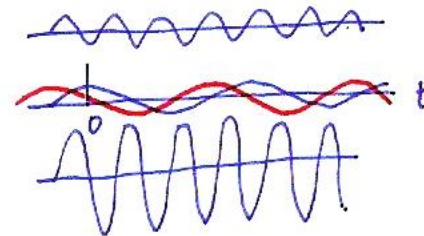
$$\text{Re}\{e^{j\omega_1 t}\} = \cos(\omega_1 t)$$

$$\text{Re}\{(A_1 e^{j\phi_1}) (e^{j\omega_1 t})\} = A_1 \cos(\omega_1 t + \phi_1)$$

$$\vec{X} = a_1 \vec{i} + a_2 \vec{j} + a_3 \vec{k}$$

$$\vec{X} = \sum_i a_i x_i$$

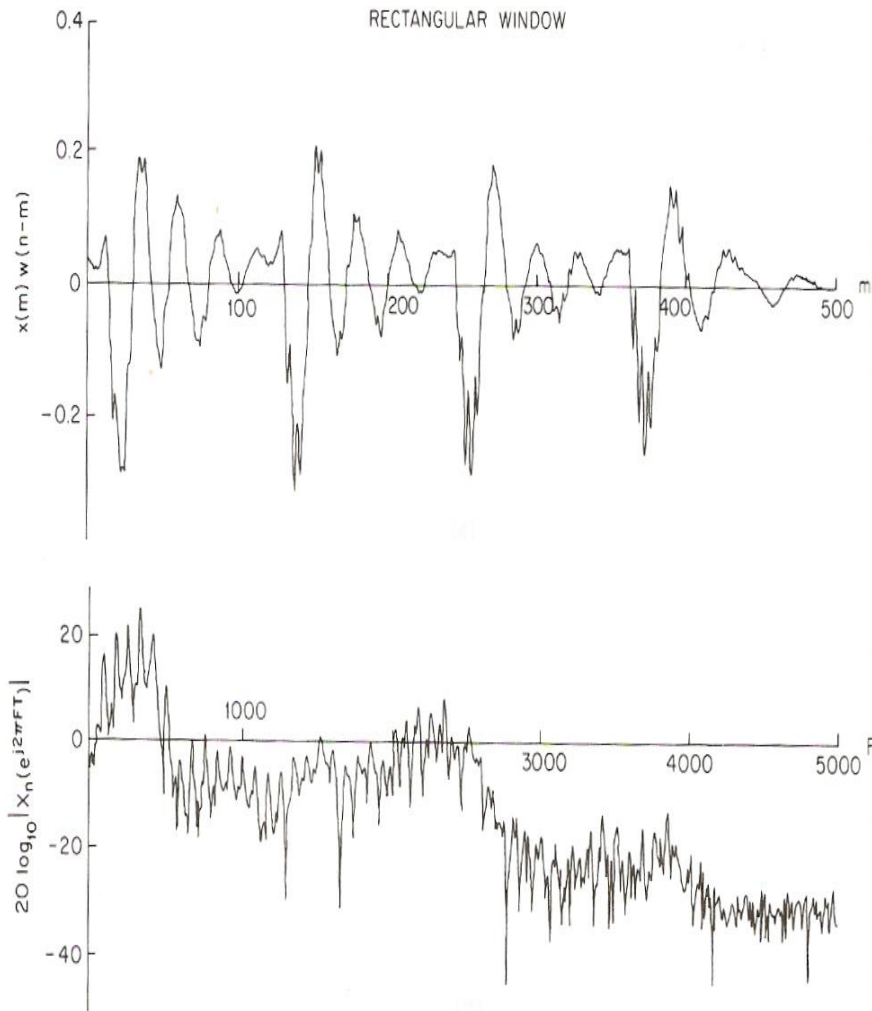
$$x(t) = \sum_i a_i x_i(t)$$



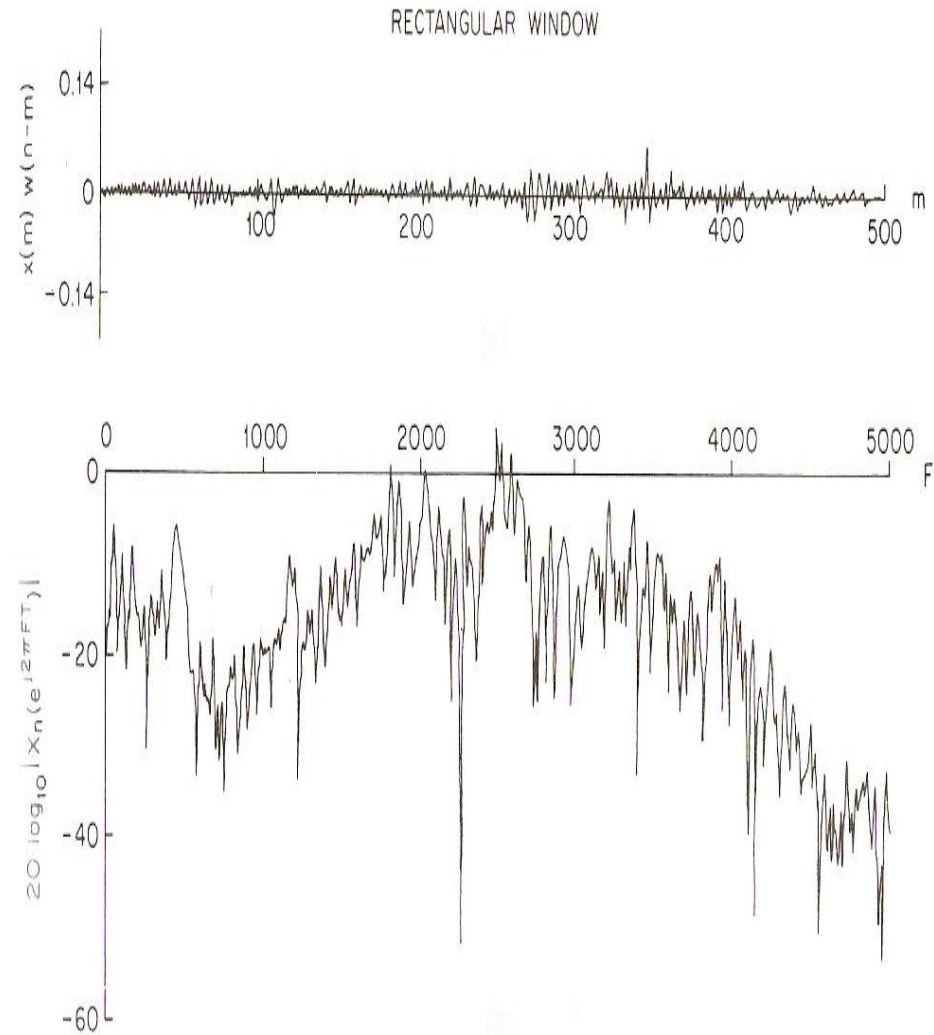


# Frequency domain spectra of speech signals

## Voiced



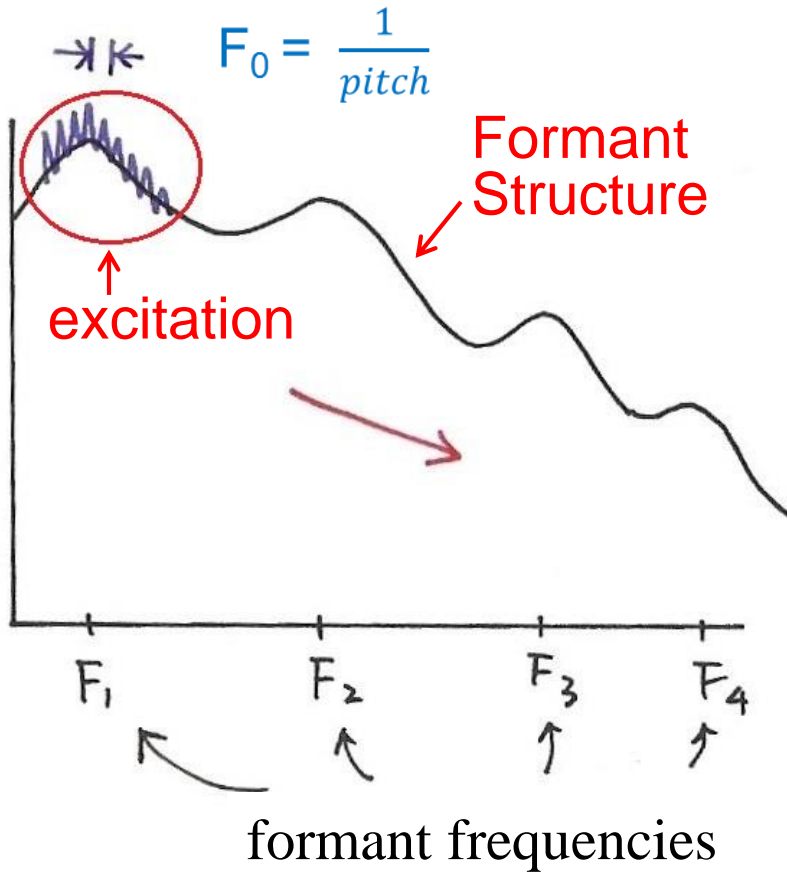
## Unvoiced



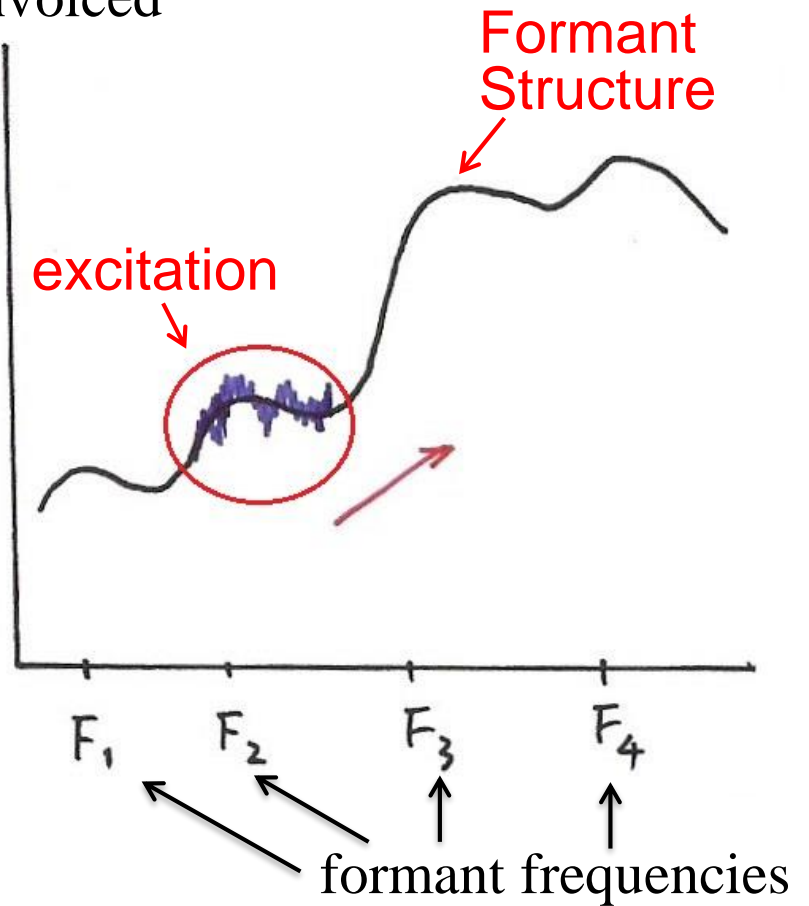


# Frequency Domain

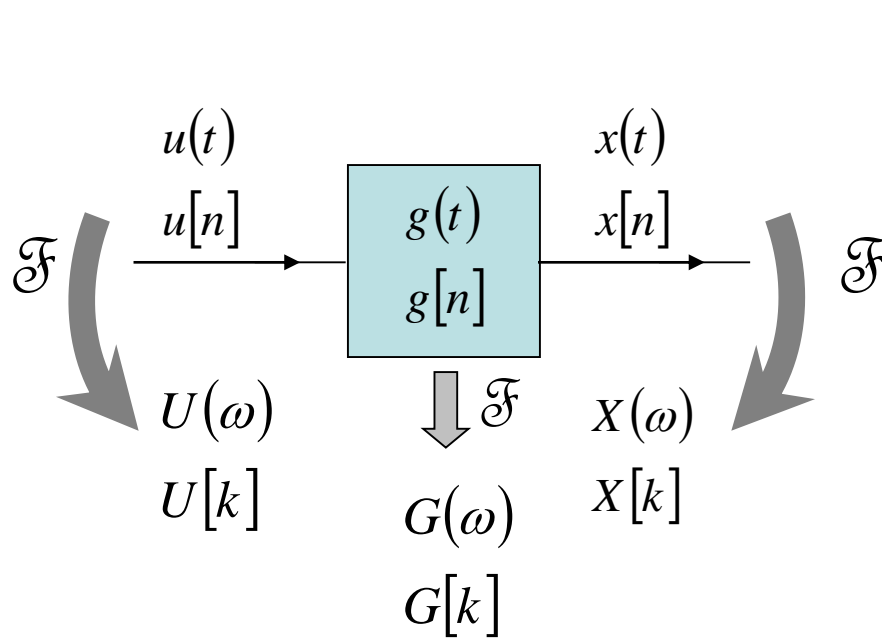
Voiced



Unvoiced



# Input/Output Relationship for Time/Frequency Domains



excitation      formant structure

$$x(t) = u(t) * g(t) = \int_{\tau} u(\tau) g(t - \tau) d\tau$$

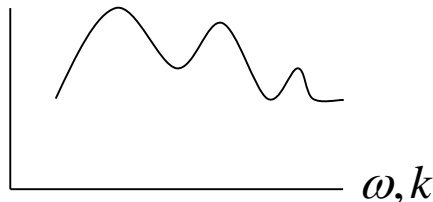
$$x[n] = u[n] * g[n] = \sum_k u[k] g[n - k]$$

time domain: convolution

$$X(\omega) = U(\omega) G(\omega)$$

$$X[k] = U[k] G[k]$$

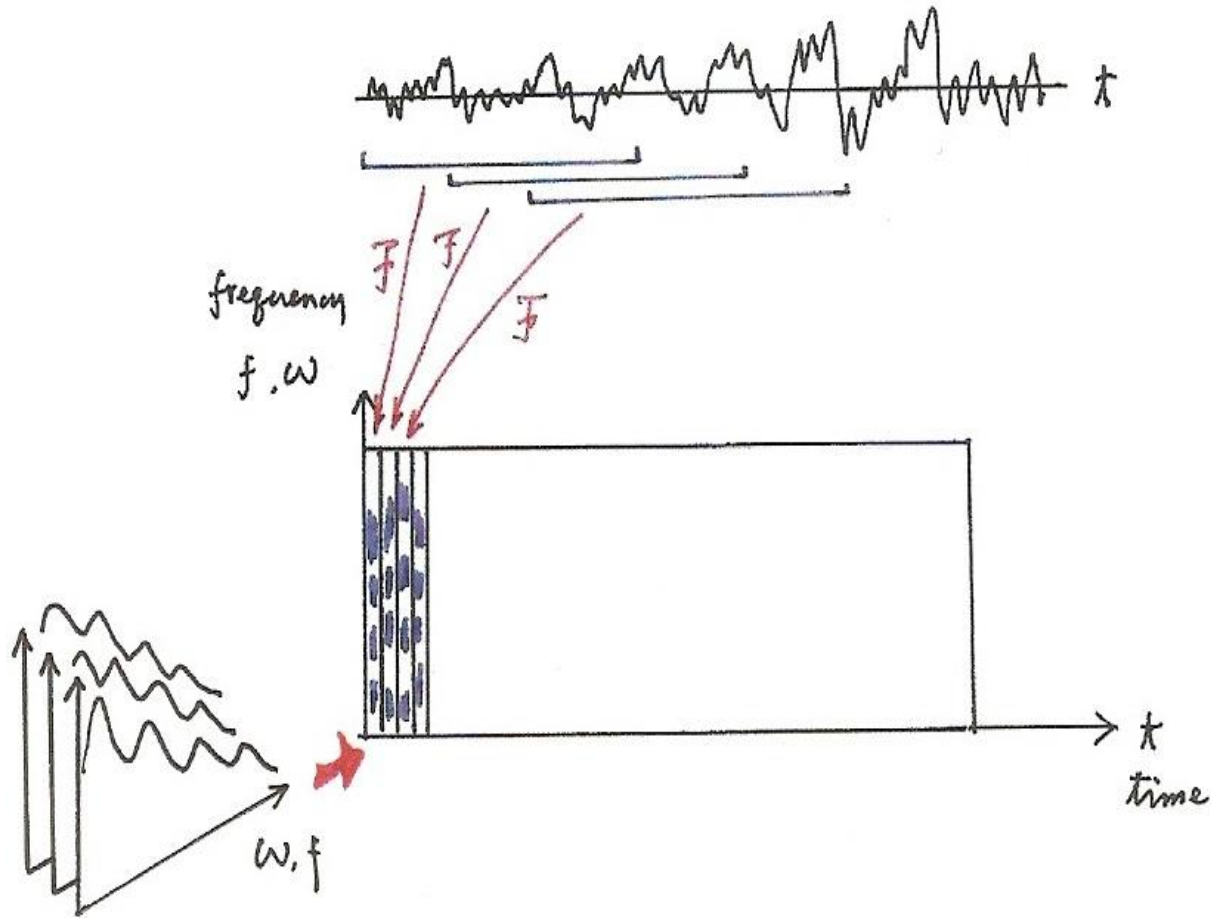
frequency domain: product



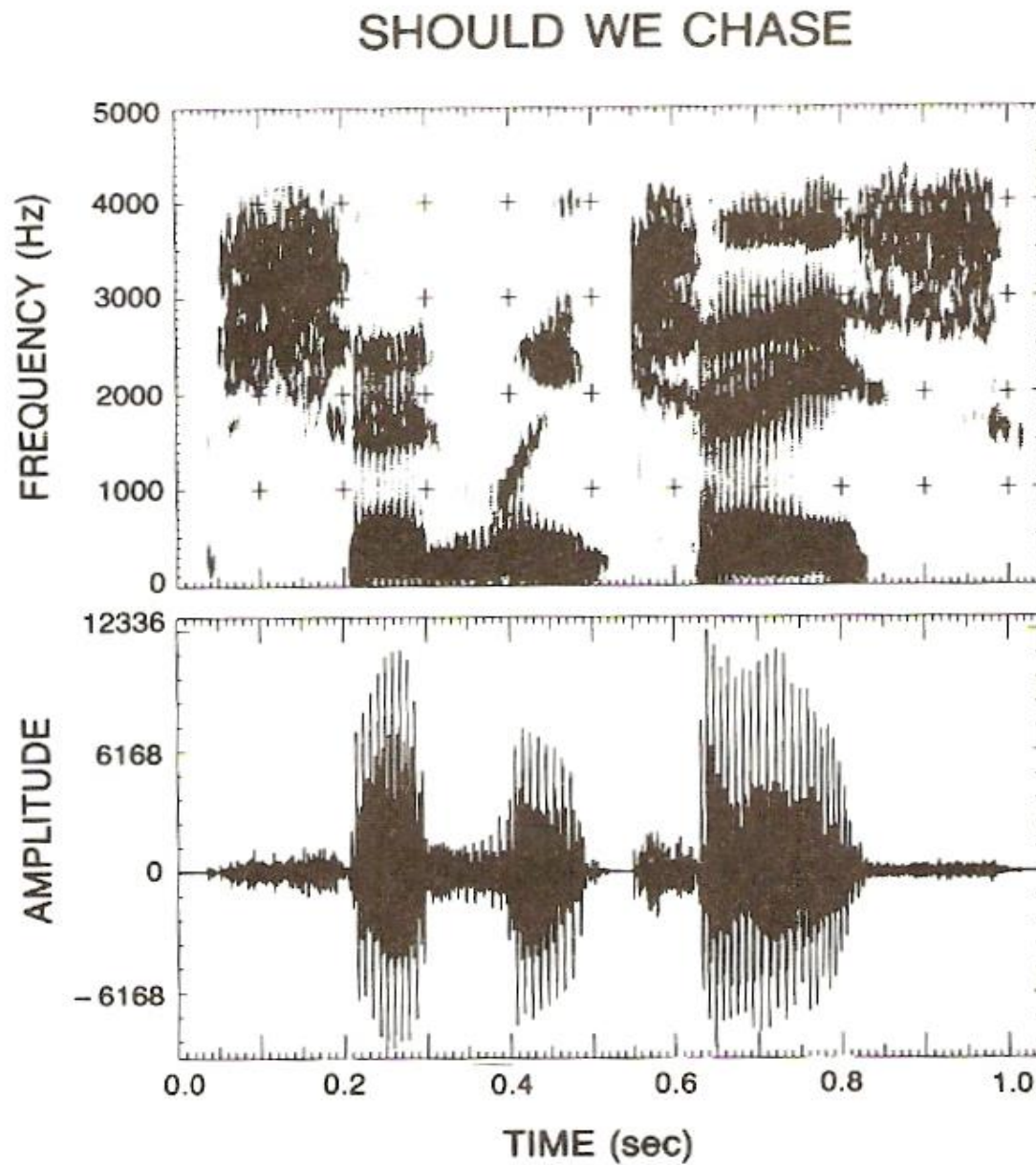
$g(t), G(\omega)$ : Formant structure: differences between phonemes

$u(t), U(\omega)$ : excitation

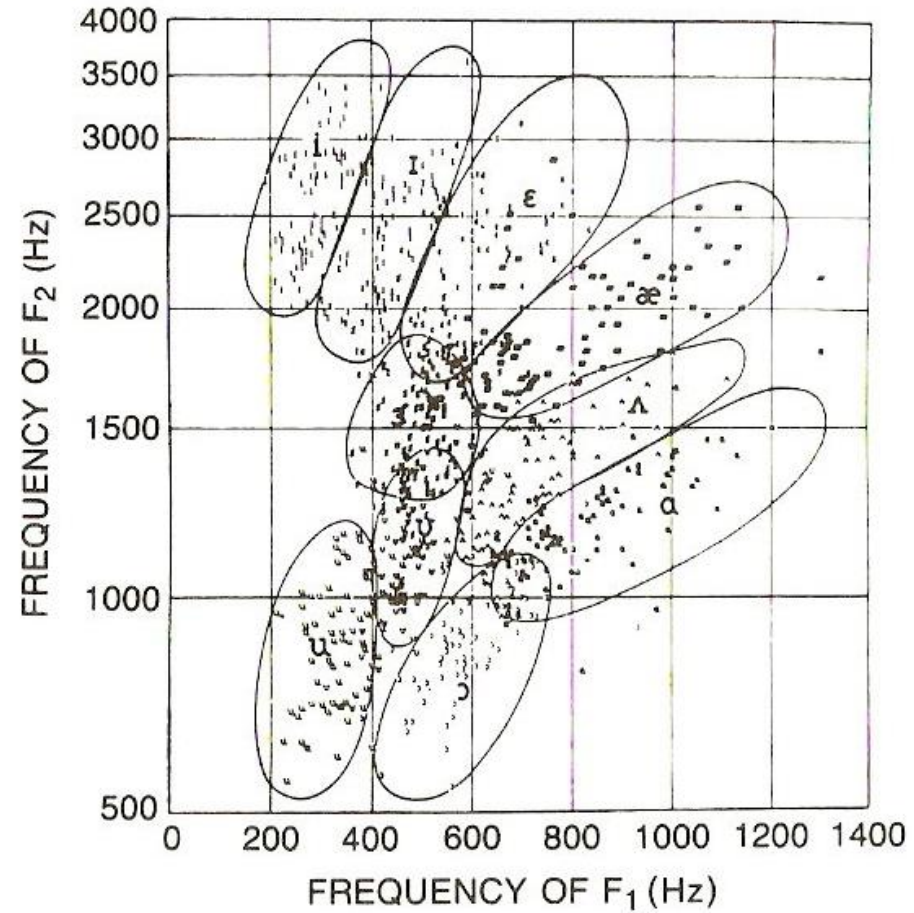
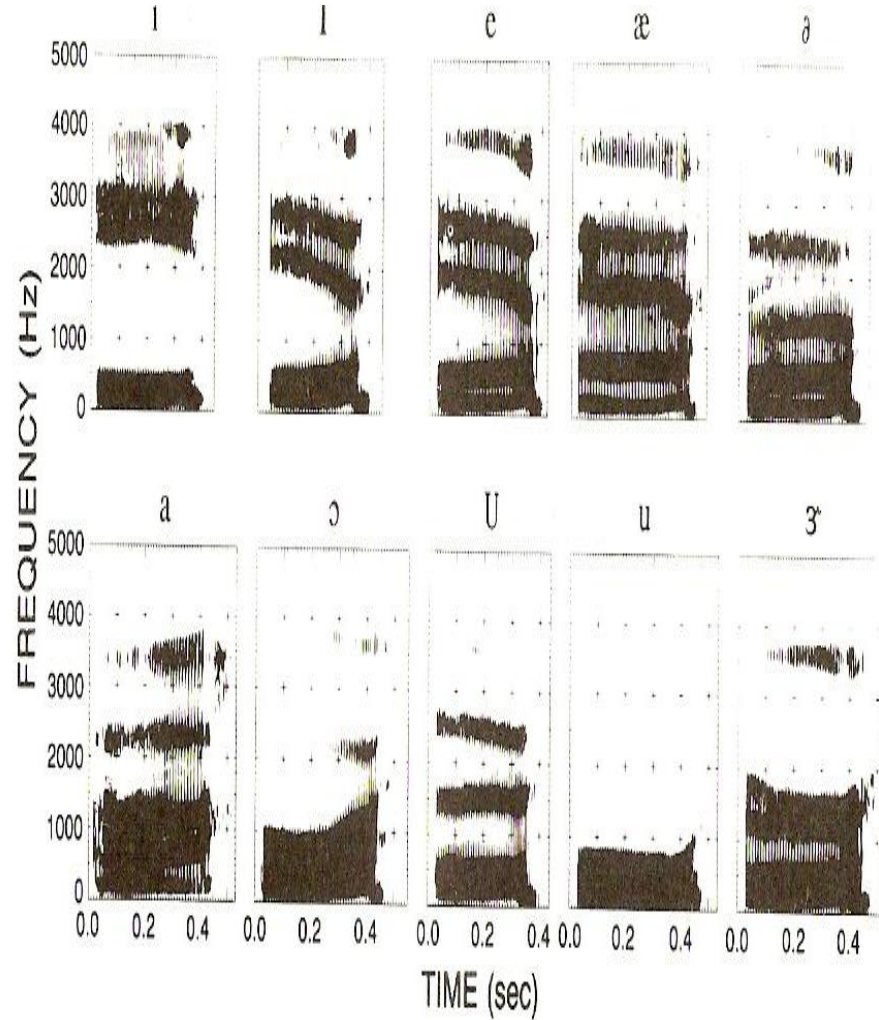
# Spectrogram



# Spectrogram

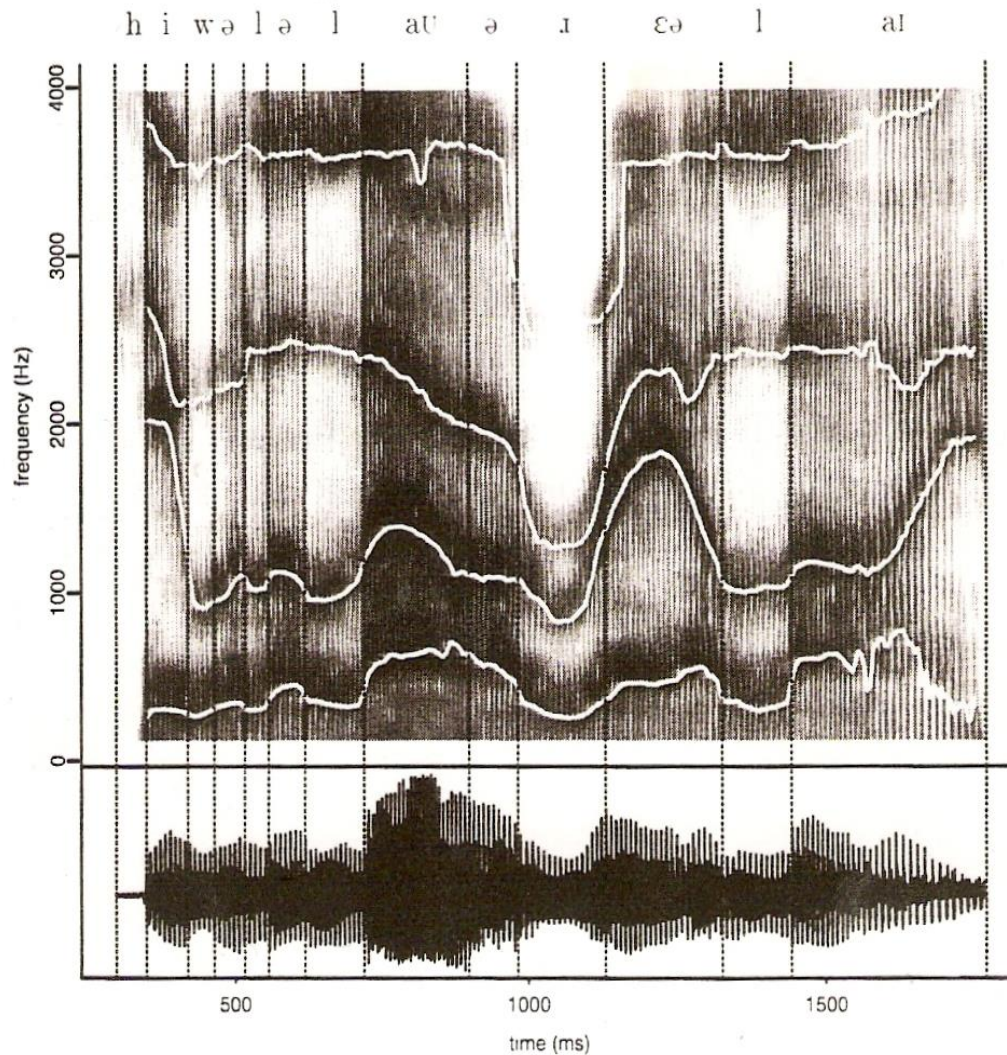


# Formant Frequencies





# Formant frequency contours

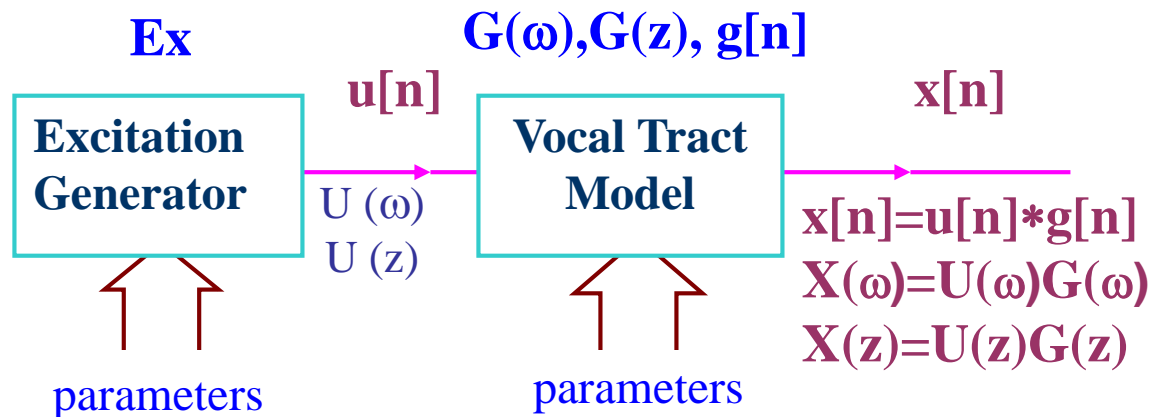


He will allow a rare lie.

Reference: 6.1 of Huang, or 2.2, 2.3 of Rabiner and Juang

# Speech Signals

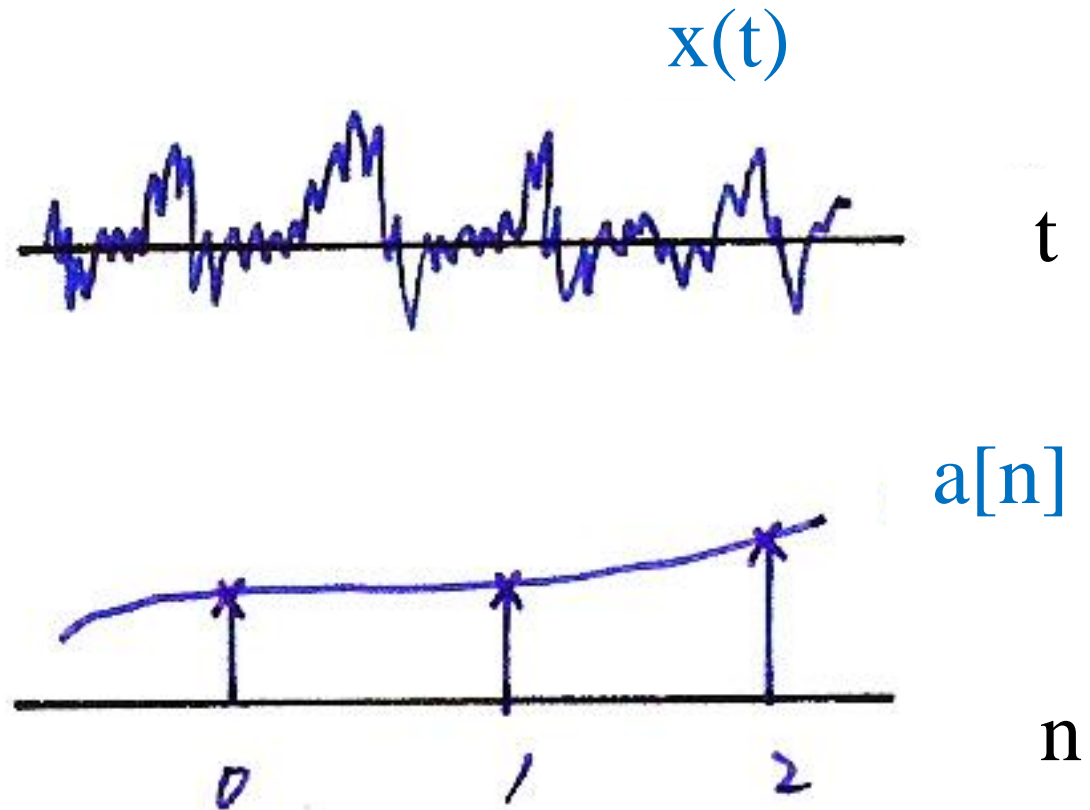
- **Voiced/unvoiced**      濁音、清音
- **Pitch/tone**      音高、聲調
- **Vocal tract**      聲道
- **Frequency domain/formant frequency**
- **Spectrogram representation**
- **Speech Source Model**



- digitization and transmission of the parameters will be adequate
- at receiver the parameters can produce  $x[n]$  with the model
- much less parameters with much slower variation in time lead to much less bits required
- the key for low bit rate speech coding

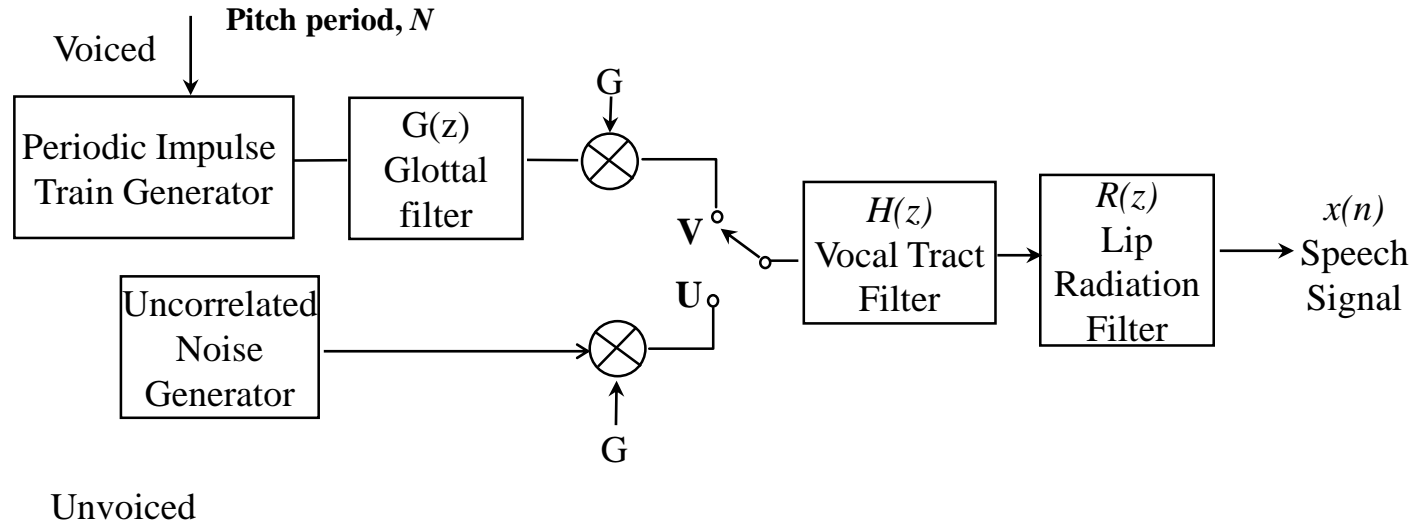


# Speech Source Model

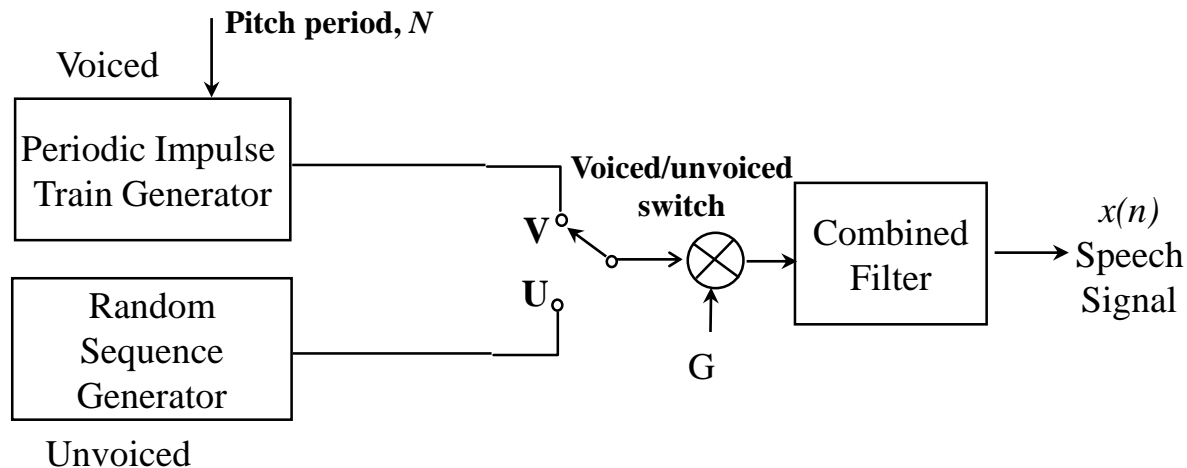


# Speech Source Model

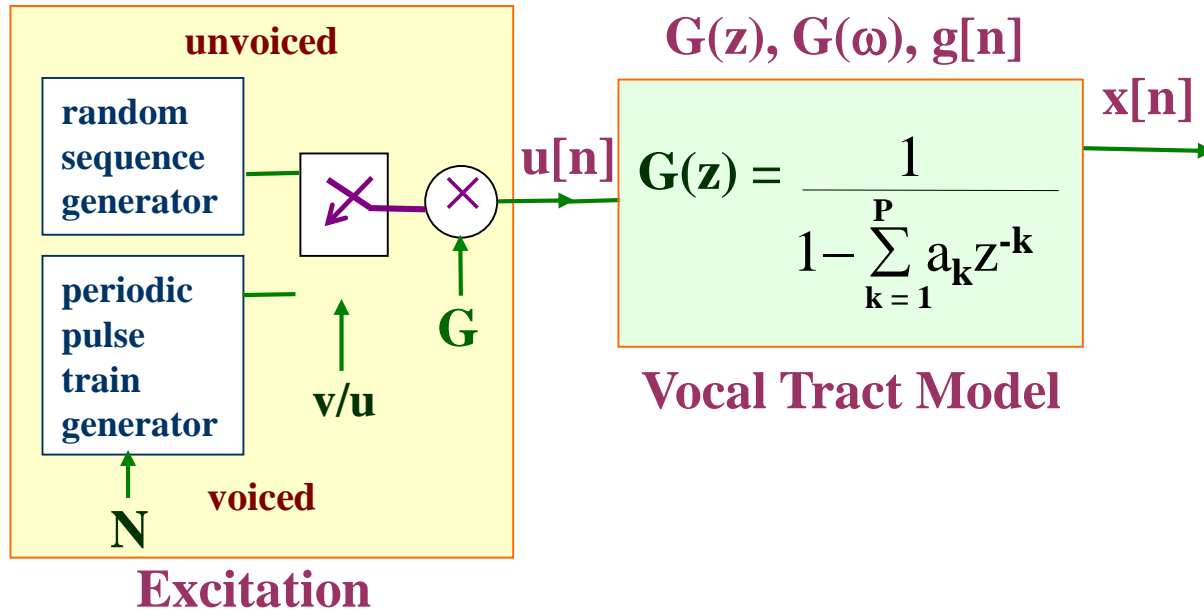
- **Sophisticated model for speech production**



- **Simplified model for speech production**



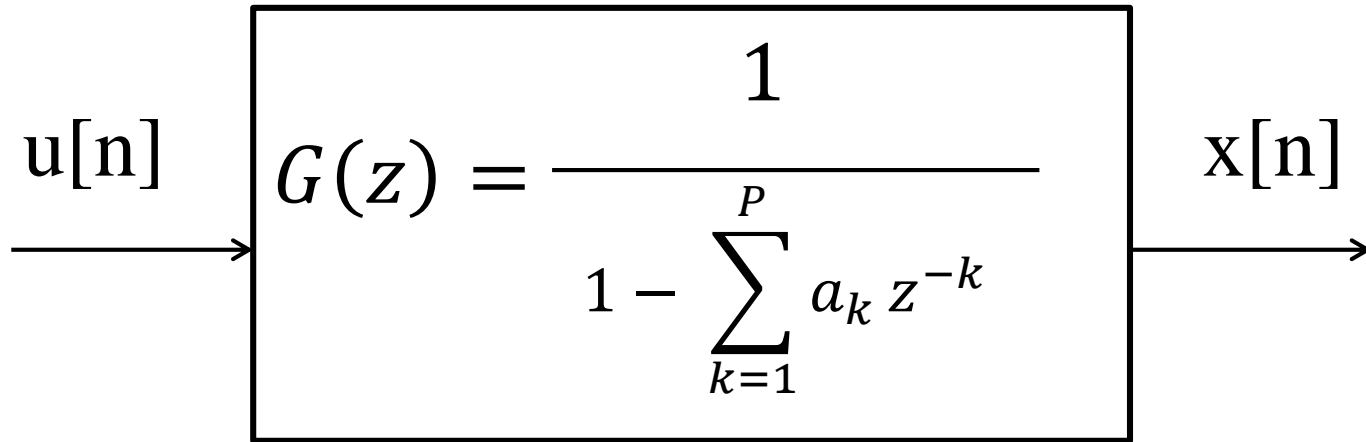
# Simplified Speech Source Model



- Excitation parameters
  - $v/u$  : voiced/ unvoiced
  - $N$  : pitch for voiced
  - $G$  : signal gain
  - excitation signal  $u[n]$
- Vocal Tract parameters
  - $\{a_k\}$  : LPC coefficients
  - formant structure of speech signals
- A good approximation, though not precise enough

Reference: 3.3.1-3.3.6 of Rabiner and Juang, or 6.3 of Huang

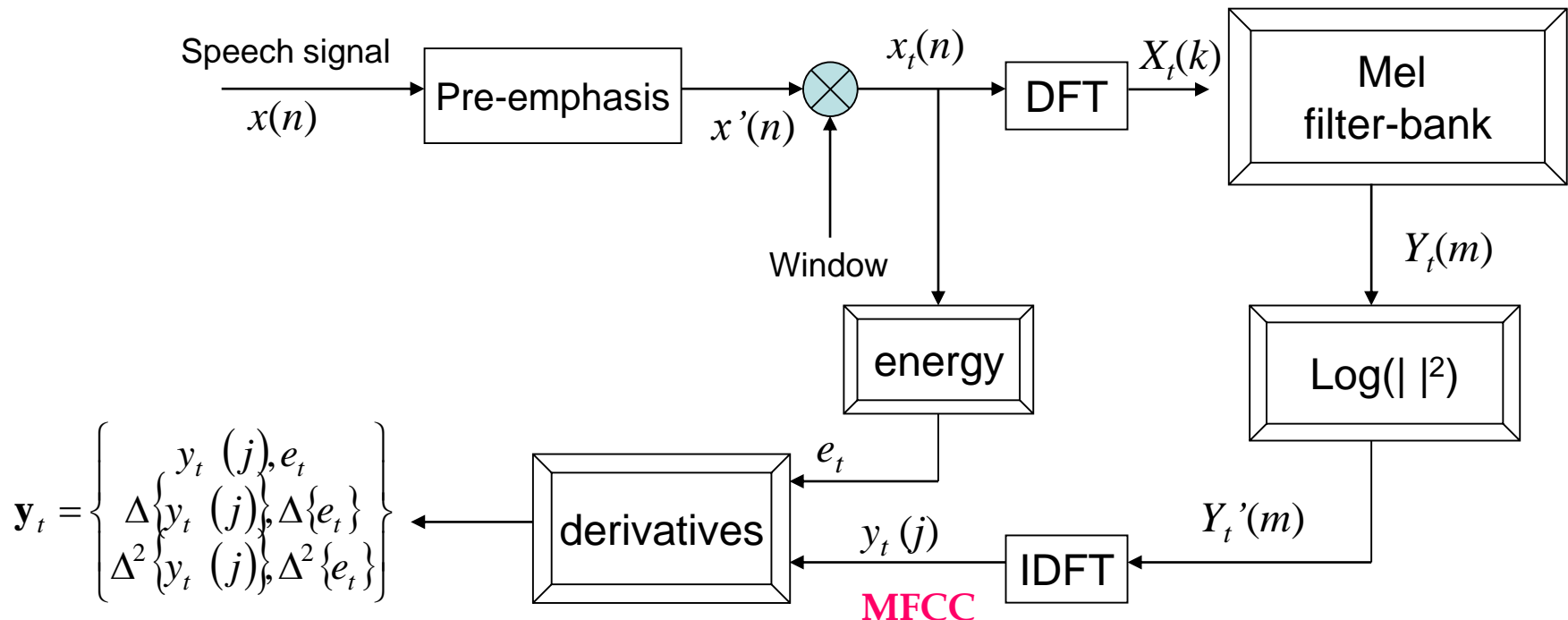
# Speech Source Model



$$x[n] - \sum_{k=1}^P a_k x[n-k] = u[n]$$

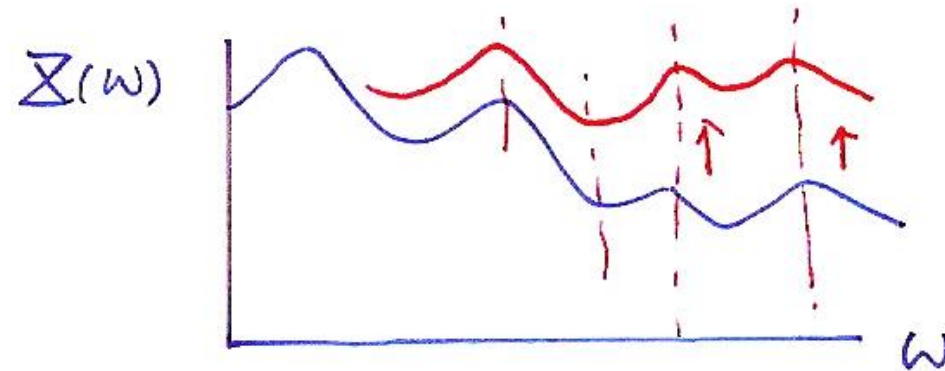
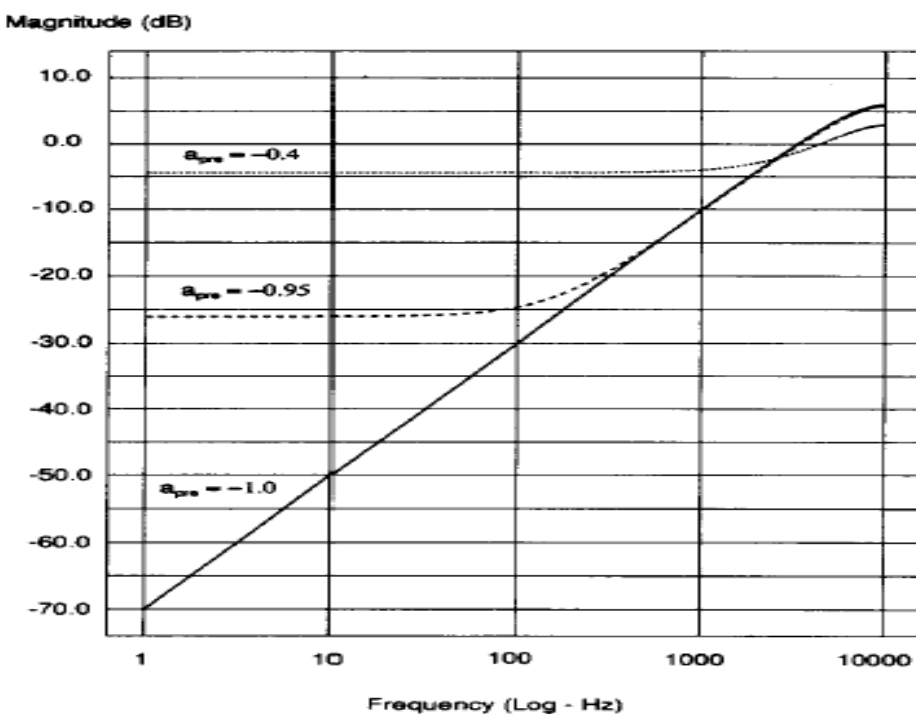
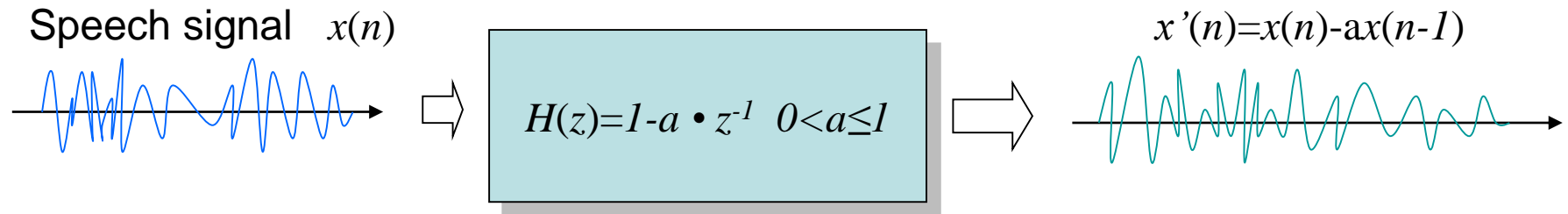
# Feature Extraction - MFCC

- **Mel-Frequency Cepstral Coefficients (MFCC)**
  - Most widely used in the speech recognition
  - Has generally obtained a better accuracy at relatively low computational complexity
  - The process of MFCC extraction :



# Pre-emphasis

- The process of Pre-emphasis :
  - a high-pass filter



# Why pre-emphasis?

---

- **Reason :**
  - Voiced sections of the speech signal naturally have a negative spectral slope (attenuation) of approximately 20 dB per decade due to the physiological characteristics of the speech production system
  - High frequency formants have small amplitude with respect to low frequency formants. A pre-emphasis of high frequencies is therefore helpful to obtain similar amplitude for all formants

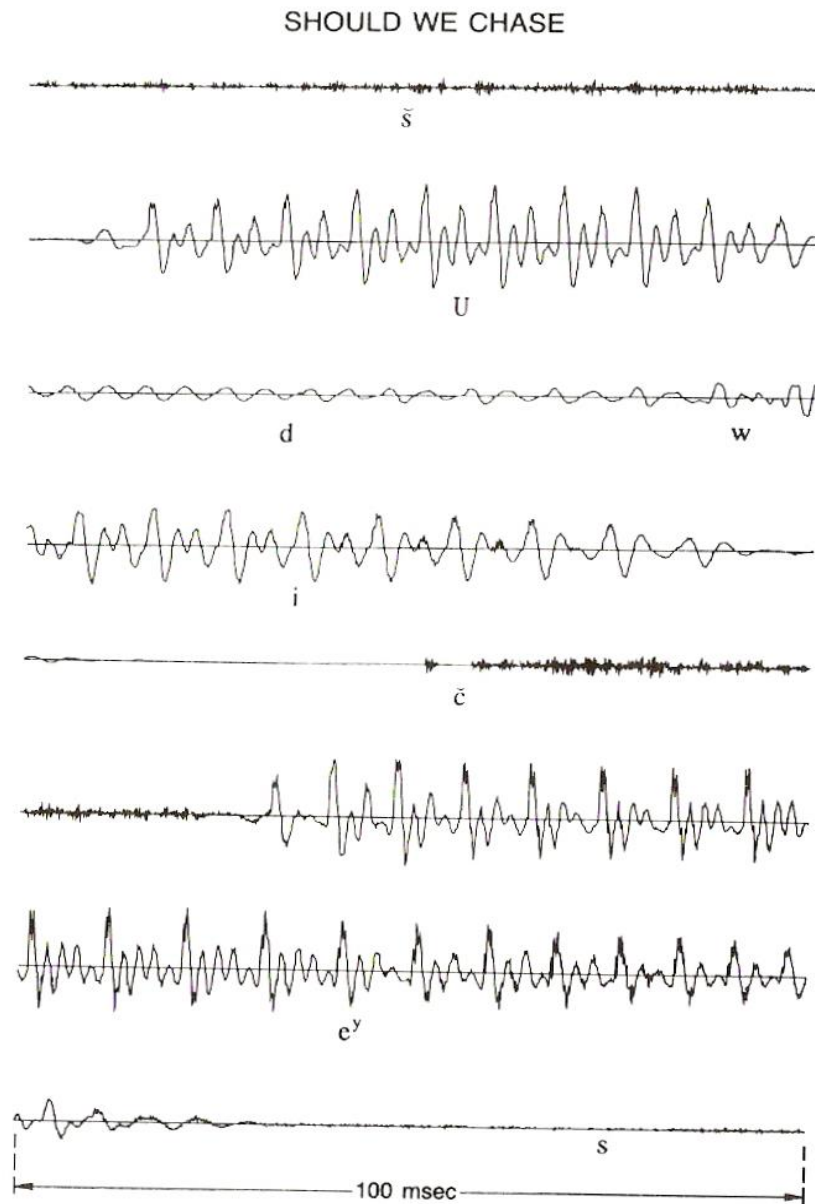


# Why Windowing?

---

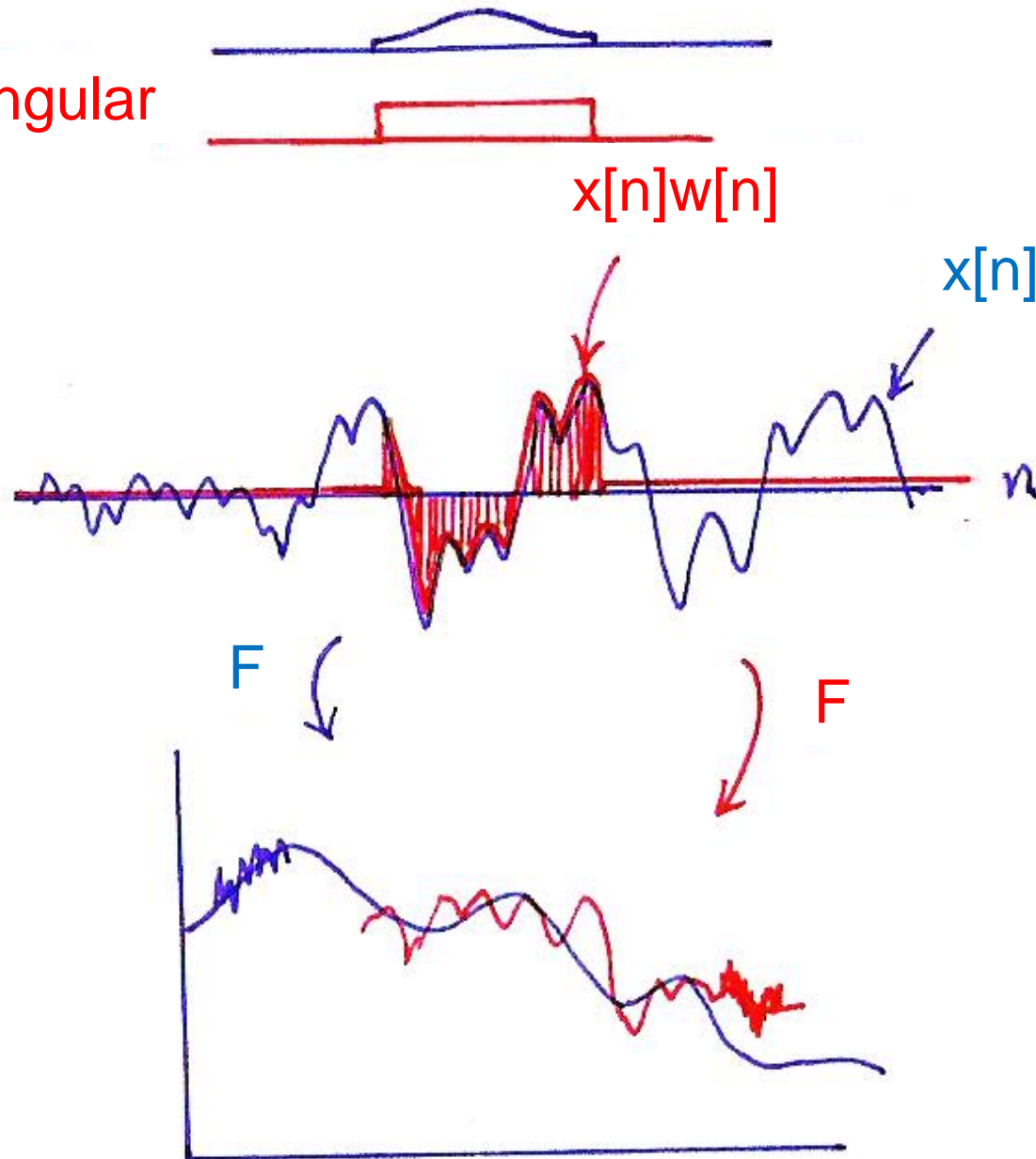
- **Why dividing the speech signal into successive and overlapping frames?**
  - Voice signals change their characteristics from time to time. The characteristics remain unchanged only in short time intervals (short-time stationary, short-time Fourier transform)
- **Frames**
  - **Frame Length** : the length of time over which a set of parameters can be obtained and is valid. Frame length ranges between **20 ~ 10** ms
  - **Frame Shift**: the length of time between successive parameter calculations
  - **Frame Rate**: number of frames per second

# Waveform plot of a sentence



# Hamming

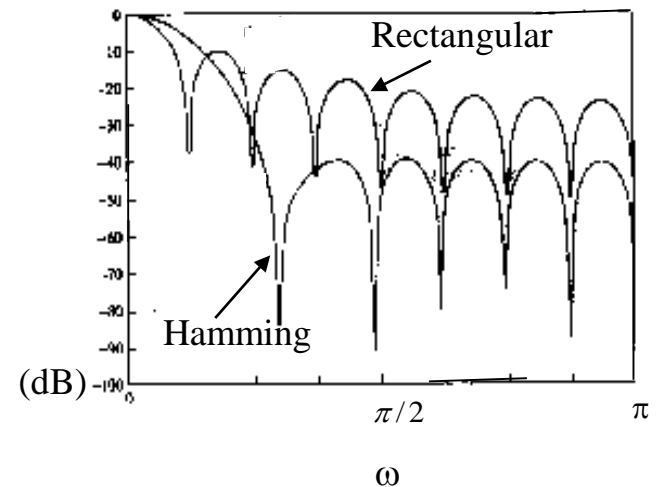
Rectangular



# Effect of Windowing (1)

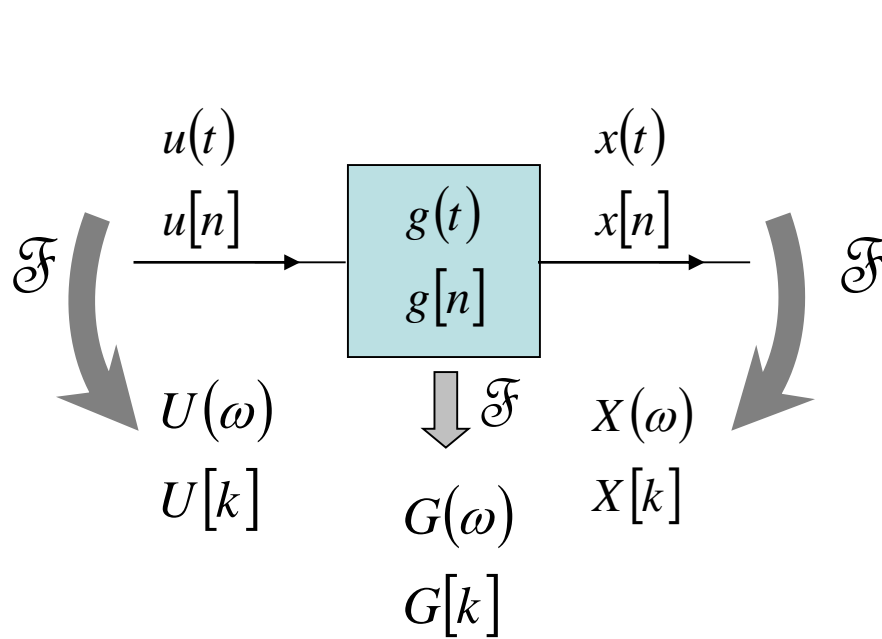
- **Windowing :**

- $x_t(n) = w(n) \cdot x'(n)$ ,  $w(n)$ : the shape of the window (product in time domain)
  - $X_t(\omega) = W(\omega) * X'(\omega)$ ,  $*$ : convolution (convolution in frequency domain)
- Rectangular window ( $w(n) = 1$  for  $0 \leq n \leq L-1$ ):
  - simply extract a segment of the signal
  - whose frequency response has high side lobes
- *Main lobe* : spreads out the narrow band power of the signal (that around the formant frequency) in a wider frequency range, and thus reduces the local frequency resolution in formant allocation
- *Side lobe* : swap energy from different and distant frequencies



# Input/Output Relationship for Time/Frequency Domains

(P.10 of 7.0)



excitation      formant structure

$$x(t) = u(t) * g(t) = \int_{\tau} u(\tau) g(t - \tau) d\tau$$

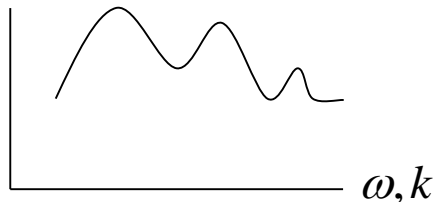
$$x[n] = u[n] * g[n] = \sum_k u[k] g[n - k]$$

time domain: convolution

$$X(\omega) = U(\omega) G(\omega)$$

$$X[k] = U[k] G[k]$$

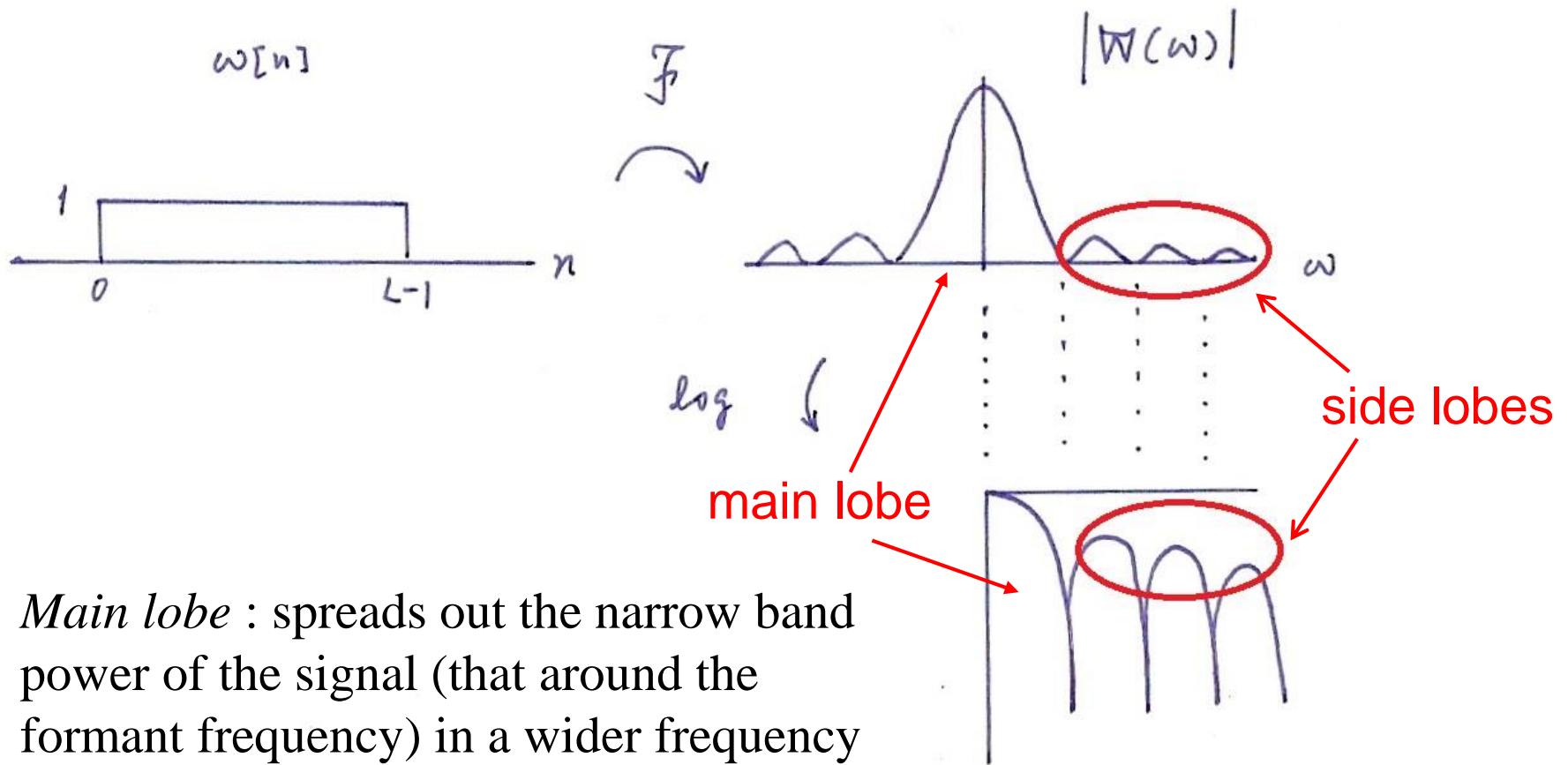
frequency domain: product



$g(t), G(\omega)$ : Formant structure: differences between phonemes

$u(t), U(\omega)$ : excitation

# Windowing



- *Main lobe* : spreads out the narrow band power of the signal (that around the formant frequency) in a wider frequency range, and thus reduces the local frequency resolution in formant allocation
- *Side lobe* : swap energy from different and distant frequencies

# Effect of Windowing (2)

---

- **Windowing (Cont.):**

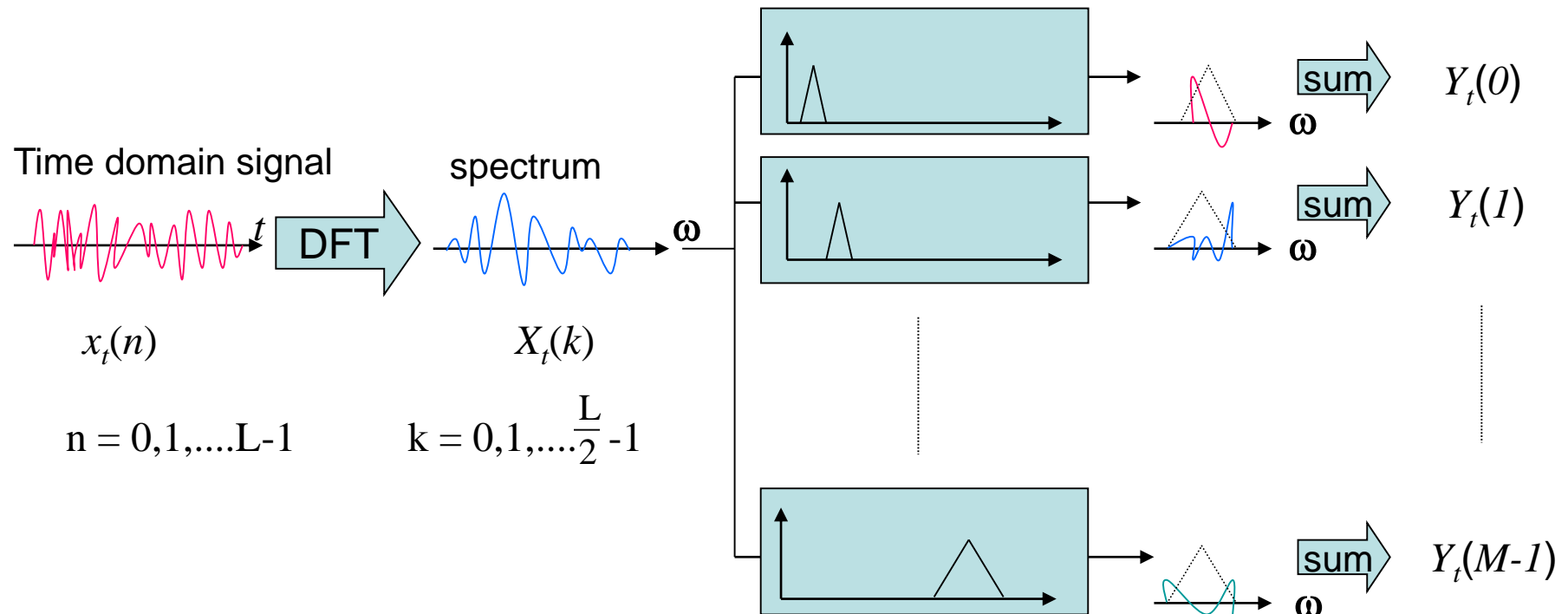
- For a designed window, we wish that
  - the main lobe is as narrow as possible
  - the side lobe is as low as possible
    - However, it is impossible to achieve both simultaneously. Some trade-off is needed
- The most widely used window shape is the Hamming window

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{L-1}\right), & n = 0, 1, \dots, L-1 \\ 0 & \text{otherwise} \end{cases}$$

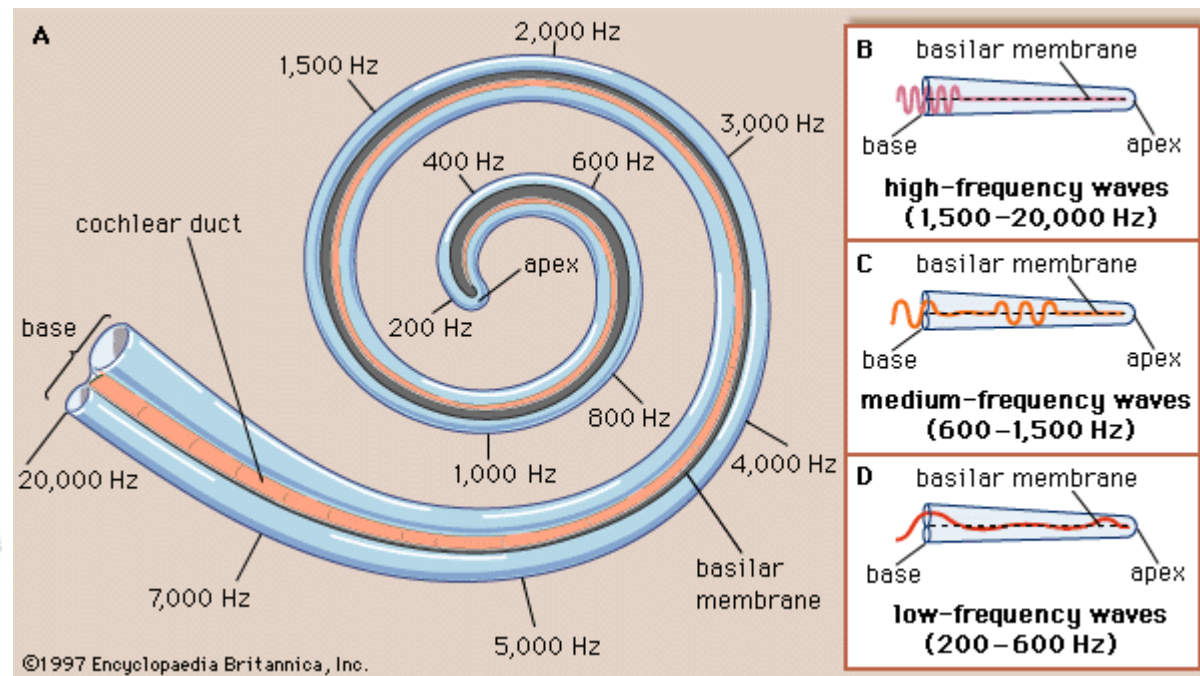
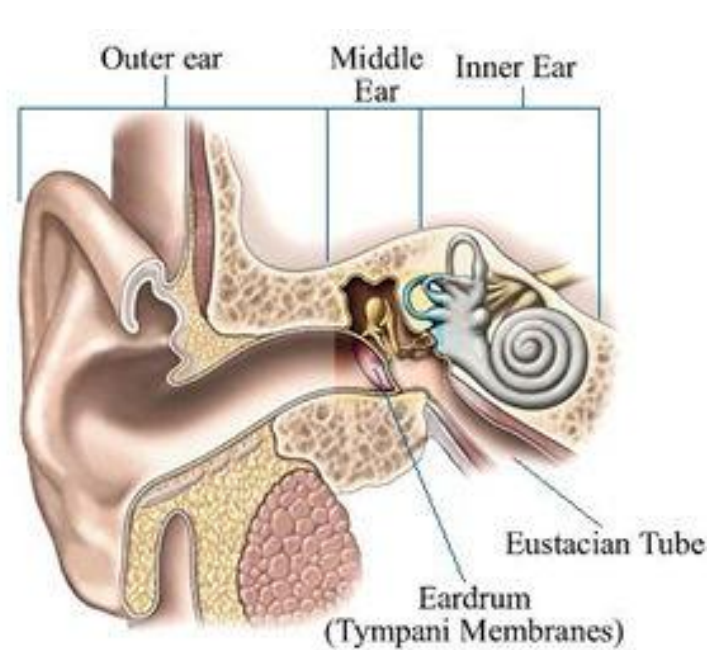


# DFT and Mel-filter-bank Processing

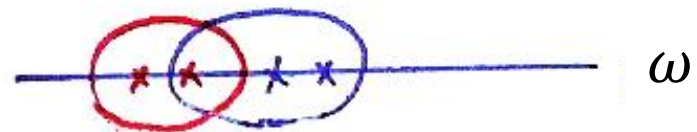
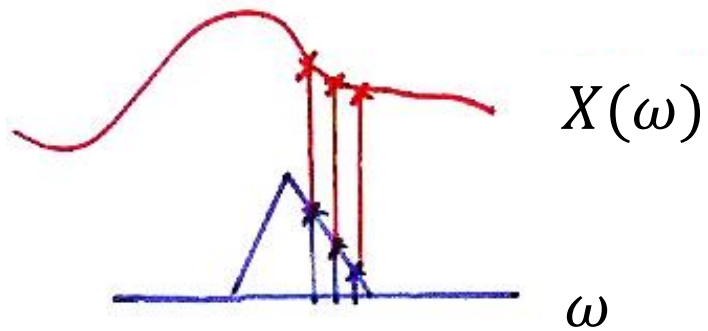
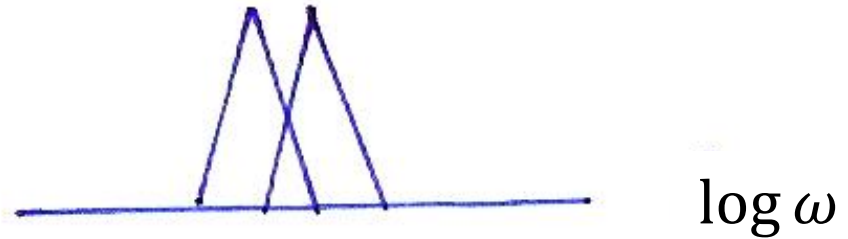
- For each frame of signal ( $L$  points, e.g.,  $L=512$ ),
  - the Discrete Fourier Transform (DFT) is first performed to obtain its spectrum ( $L$  points, for example  $L=512$ )
  - The bank of filters based on Mel scale is then applied, and each filter output is the sum of its filtered spectral components ( $M$  filters, and thus  $M$  outputs, for example  $M=24$ )



# Peripheral Processing for Human Perception

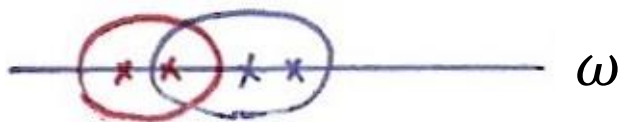


# Mel-scale Filter Bank



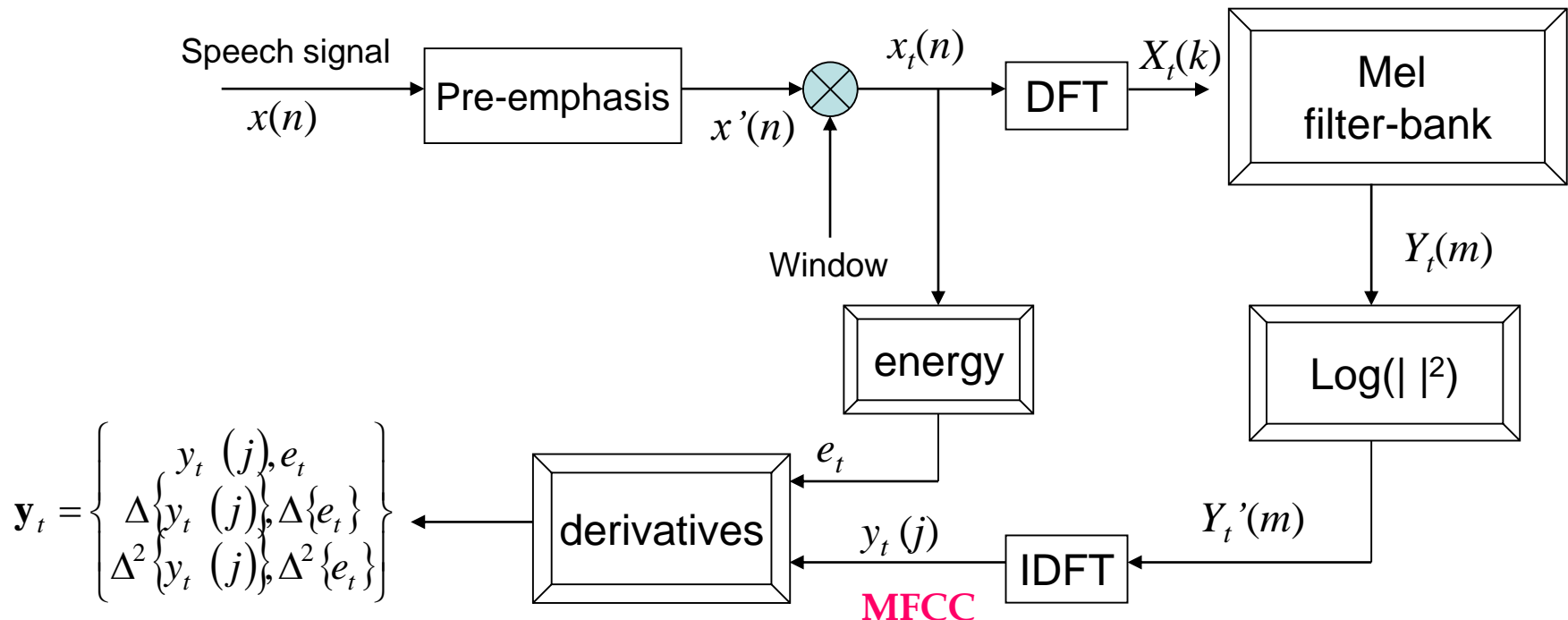
# Why Filter-bank Processing?

- **The filter-bank processing simulates human ear perception**
  - Frequencies of a complex sound within a certain frequency band cannot be individually identified.
  - When one of the components of this sound falls outside this frequency band, it can be individually distinguished.
  - This frequency band is referred to as the critical band.
  - These critical bands somehow overlap with each other.
  - The critical bands are roughly distributed linearly in the logarithm frequency scale (including the center frequencies and the bandwidths), specially at higher frequencies.
  - Human perception for pitch of signals is proportional to the *logarithm* of the frequencies (relative ratios between the frequencies)



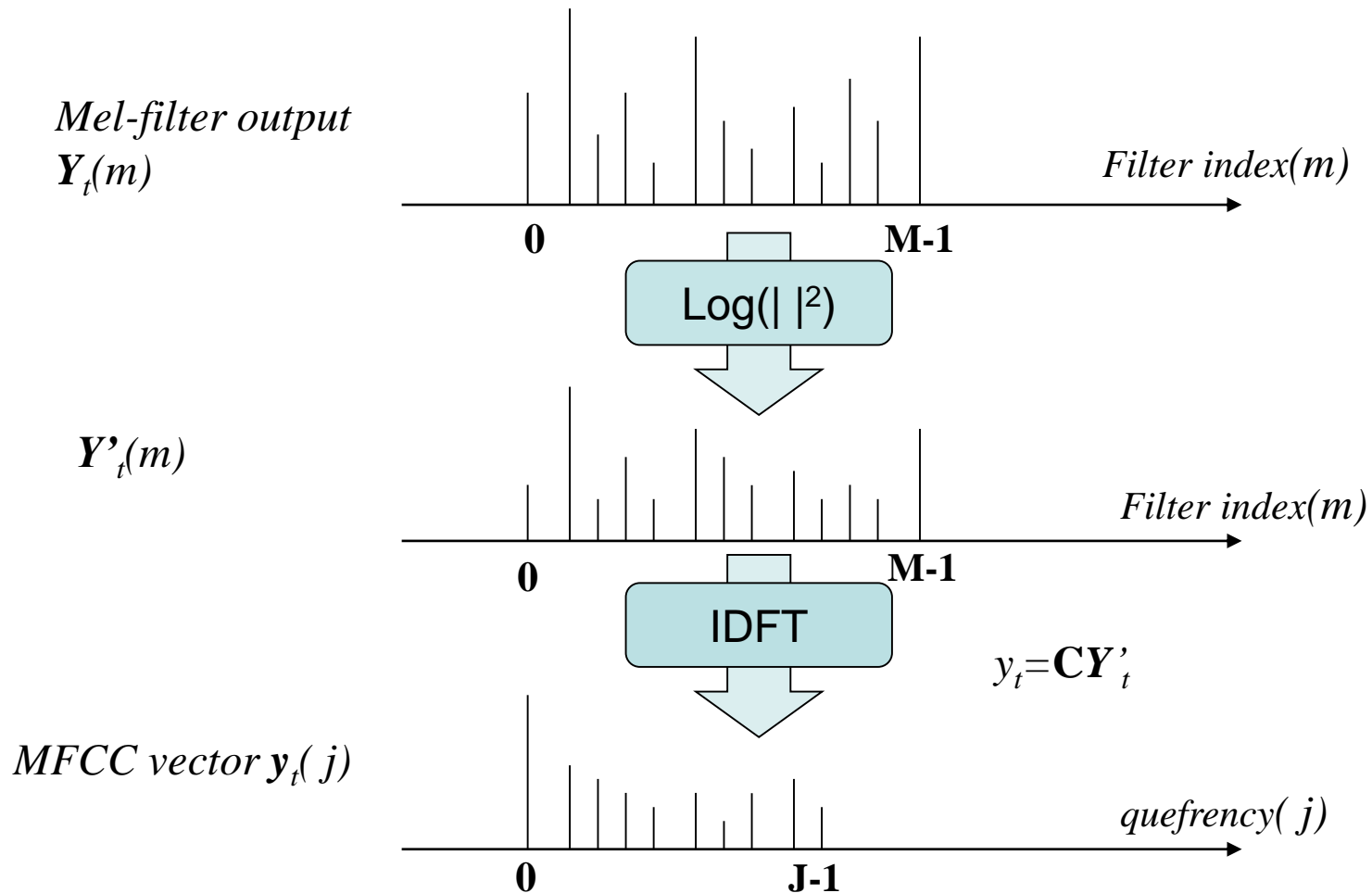
# Feature Extraction - MFCC

- **Mel-Frequency Cepstral Coefficients (MFCC)**
  - Most widely used in the speech recognition
  - Has generally obtained a better accuracy at relatively low computational complexity
  - The process of MFCC extraction :



# Logarithmic Operation and IDFT

- The final process of MFCC evaluation : logarithm operation and IDFT



# Why Log Energy Computation?

- **Using the magnitude (or energy) only**
  - Phase information is not very helpful in speech recognition
    - Replacing the phase part of the original speech signal with continuous random phase usually won't be perceived by human ears
- **Using the Logarithmic operation**
  - Human perception sensitivity is proportional to signal energy in logarithmic scale (relative ratios between signal energy values)
  - The logarithm compresses larger values while expands smaller values, which is a characteristic of the human hearing system
  - The dynamic compression also makes feature extraction less sensitive to variations in signal dynamics
  - To make a convolved noisy process additive
    - Speech signal  $x(n)$ , excitation  $u(n)$  and the impulse response of vocal tract  $g(n)$

$$x(n)=u(n)*g(n) \rightarrow X(\omega)=U(\omega)G(\omega)$$

$$\rightarrow |X(\omega)|=|U(\omega)||G(\omega)| \rightarrow \log|X(\omega)|=\log|U(\omega)|+\log|G(\omega)|$$



# Why Inverse DFT?

---

- **Final procedure for MFCC : performing the inverse DFT on the log-spectral power**

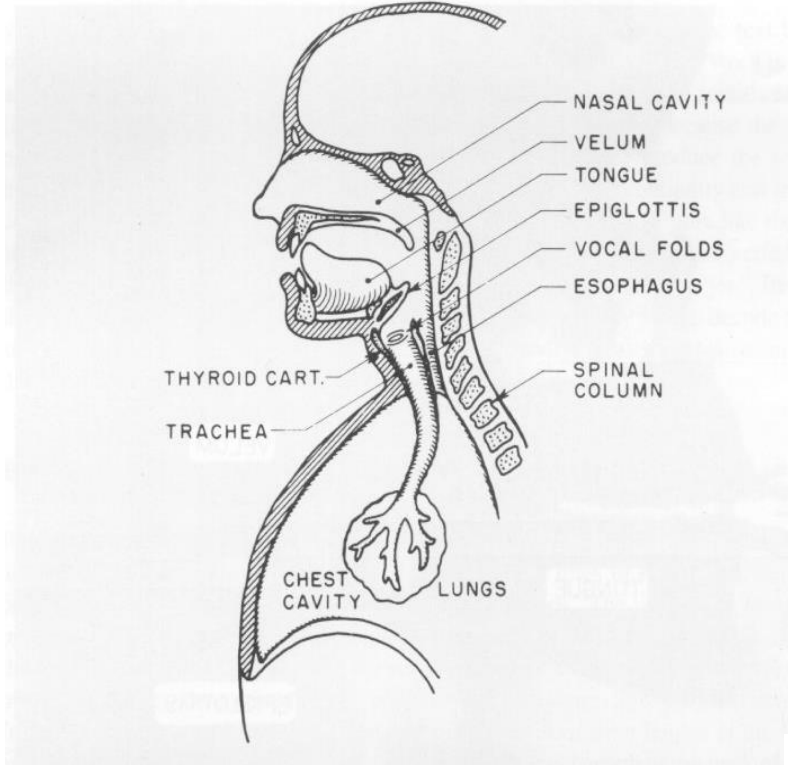
$$y_t(j) = \sum_{m=0}^{M-1} \log(|Y_t(m)|^2) \cos \left[ j \left( m - \frac{1}{2} \right) \frac{\pi}{M} \right], \quad j = 0, 1, \dots, J-1 < M$$

- **Advantages :**

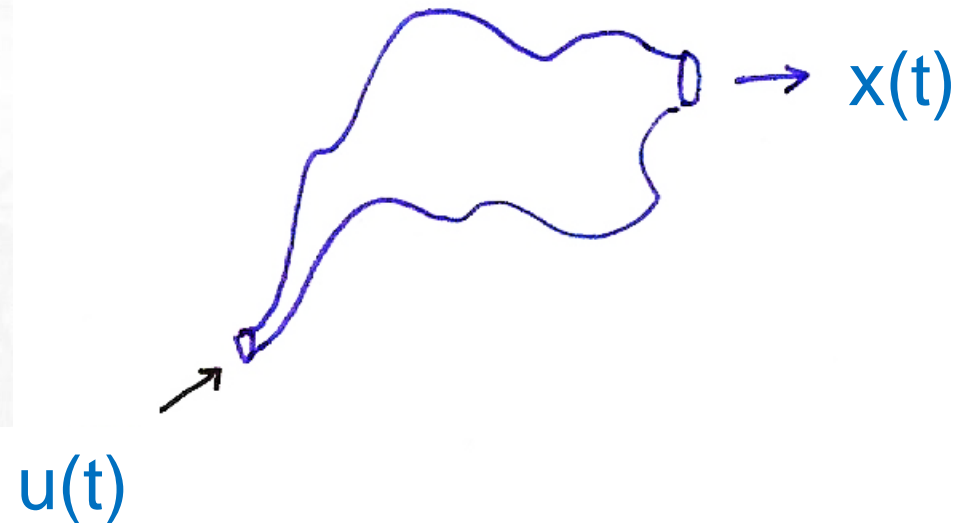
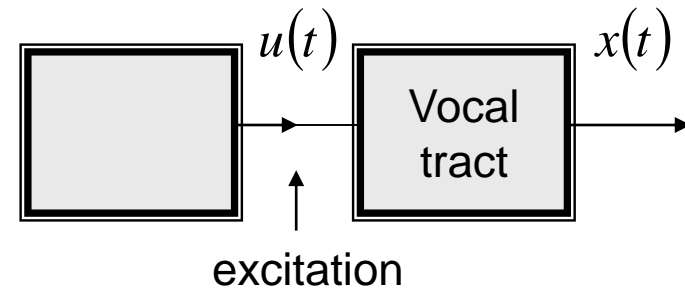
- Since the log-power spectrum is real and symmetric, the inverse DFT reduces to a Discrete Cosine Transform (DCT). The DCT has the property to produce highly uncorrelated features  $y_t$ 
  - diagonal rather than full covariance matrices can be used in the Gaussian distributions in many cases
- Easier to remove the interference of excitation on formant structures
  - the phoneme for a segment of speech signal is primarily based on the formant structure (or vocal tract shape)
  - on the frequency scale the formant structure changes slowly over frequency, while the excitation changes much faster

# Speech Production and Source Model (P.3 of 7.0)

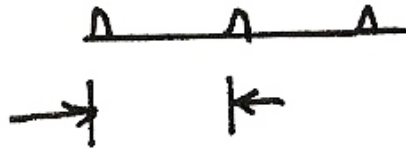
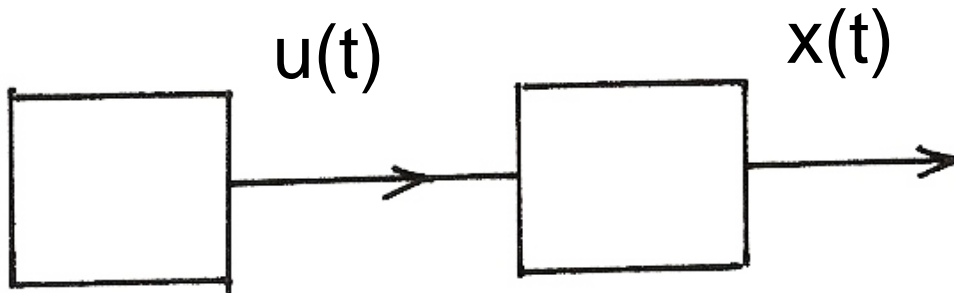
- Human vocal mechanism



- Speech Source Model



# Voiced and Unvoiced Speech (P.4 of 7.0)



pitch



pitch

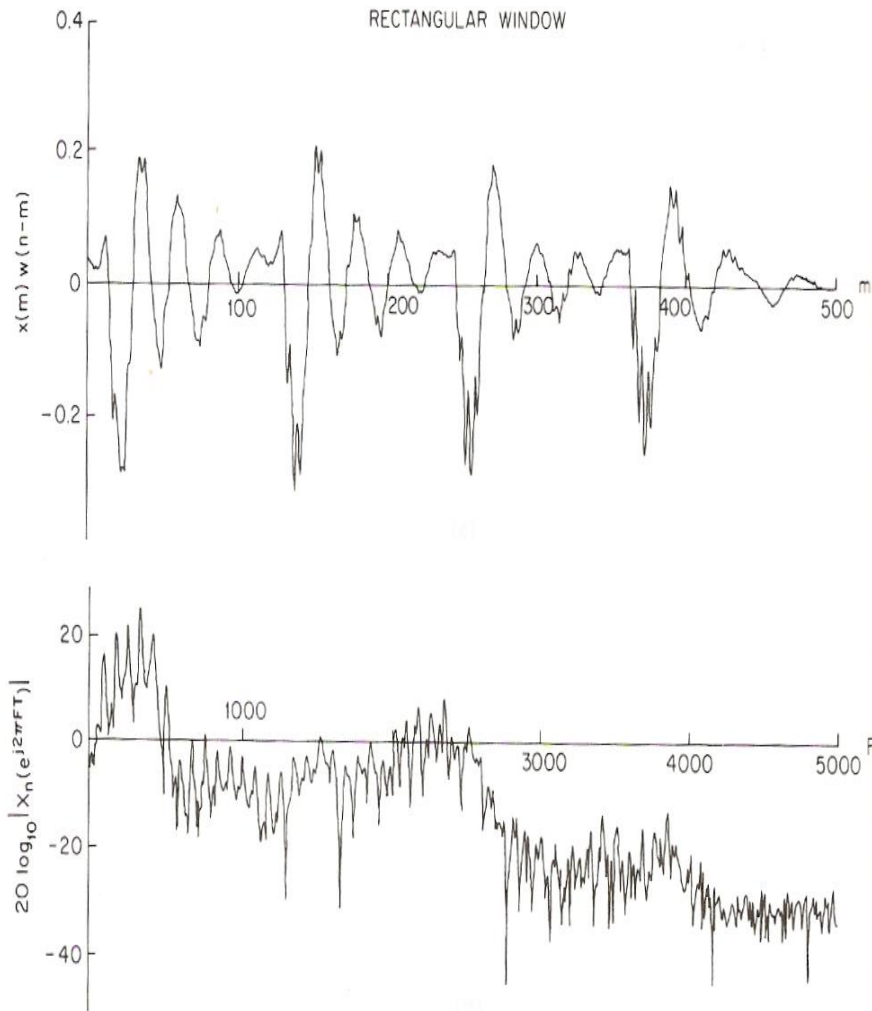
voiced



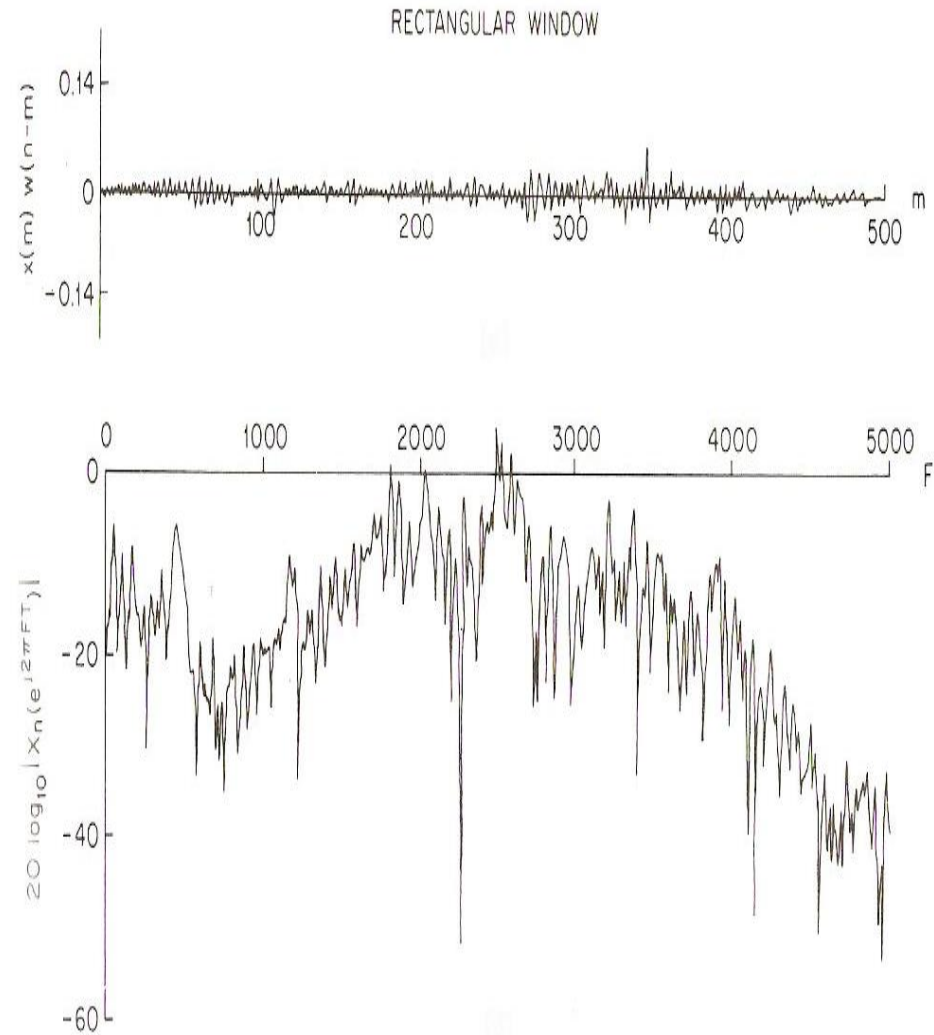
unvoiced

# Frequency domain spectra of speech signals (P.8 of 7.0)

## Voiced

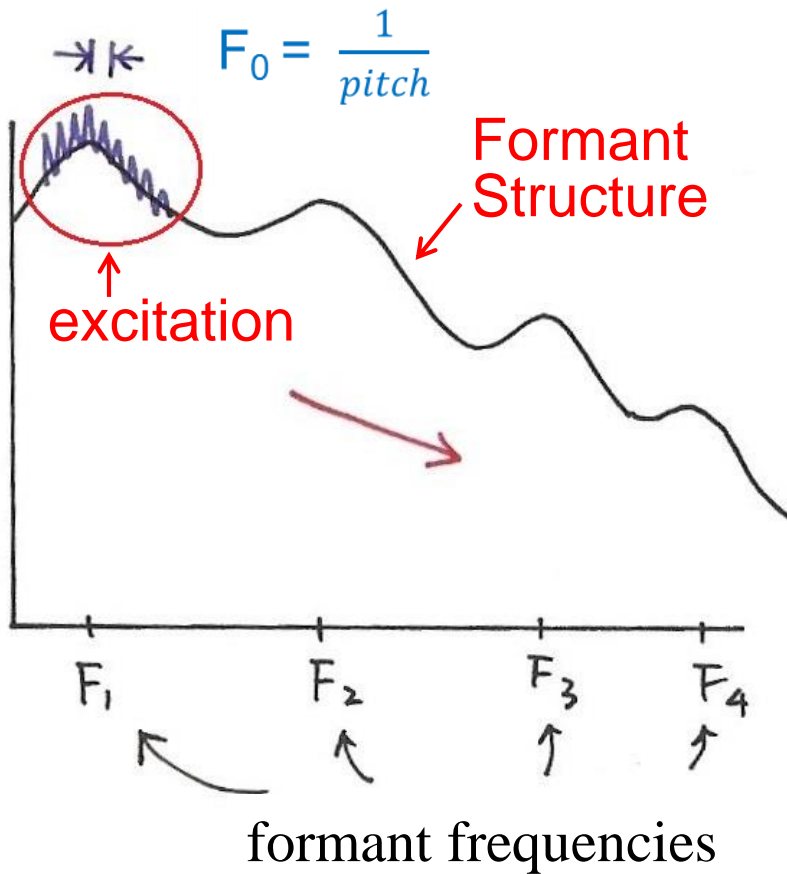


## Unvoiced

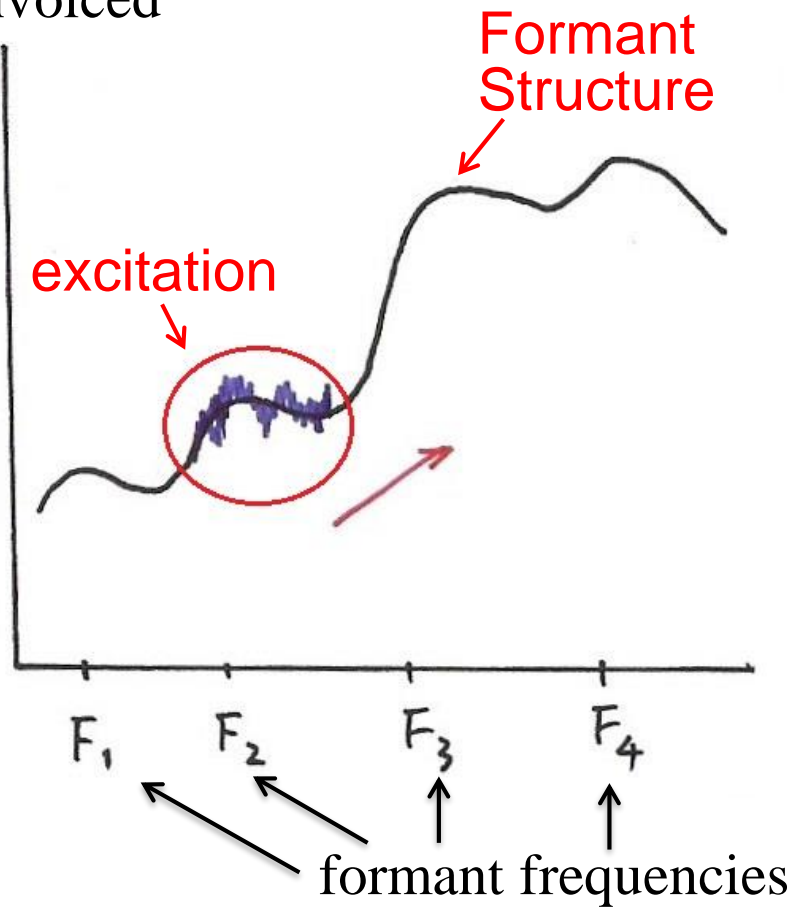


# Frequency Domain (P.9 of 7.0)

Voiced

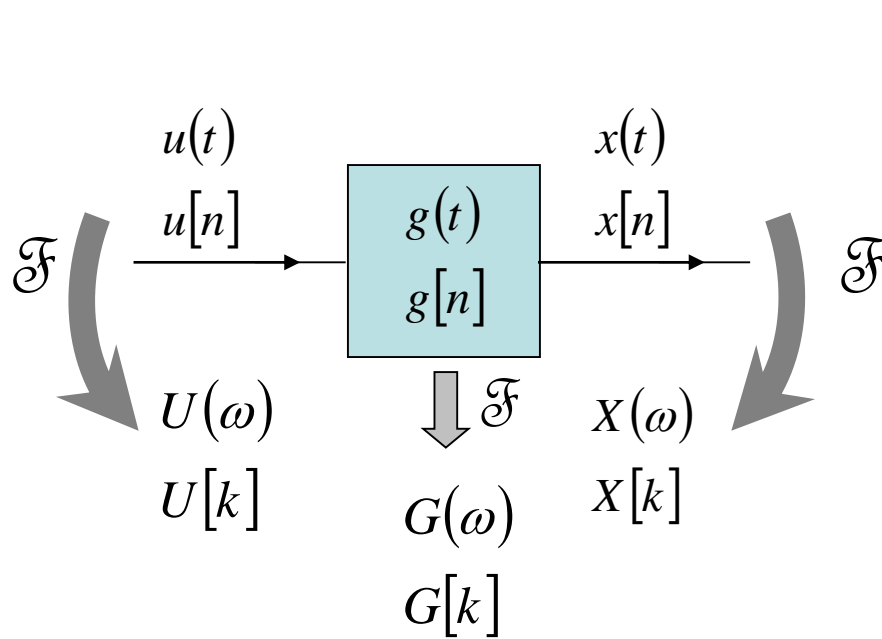


Unvoiced



# Input/Output Relationship for Time/Frequency Domains

(P.10 of 7.0)



excitation      formant structure

$$x(t) = u(t) * g(t) = \int_{\tau} u(\tau) g(t - \tau) d\tau$$

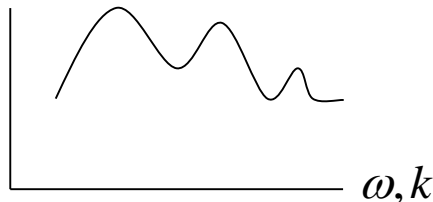
$$x[n] = u[n] * g[n] = \sum_k u[k] g[n - k]$$

time domain: convolution

$$X(\omega) = U(\omega) G(\omega)$$

$$X[k] = U[k] G[k]$$

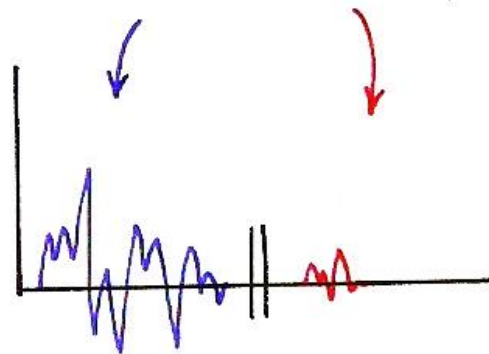
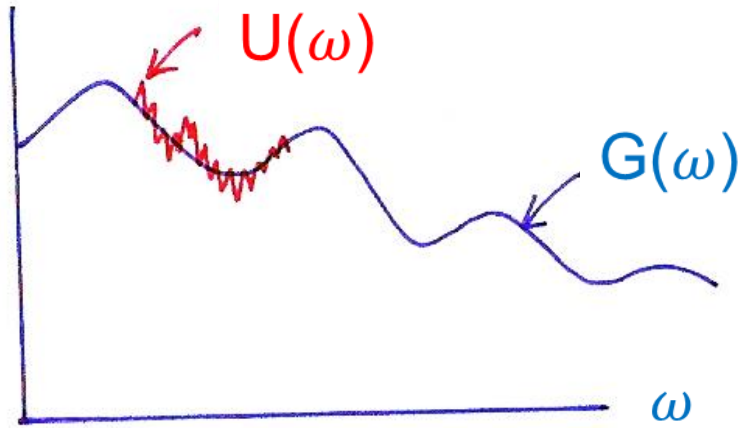
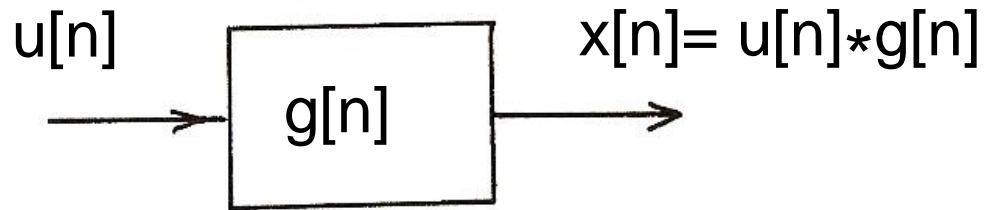
frequency domain: product



$g(t), G(\omega)$ : Formant structure: differences between phonemes

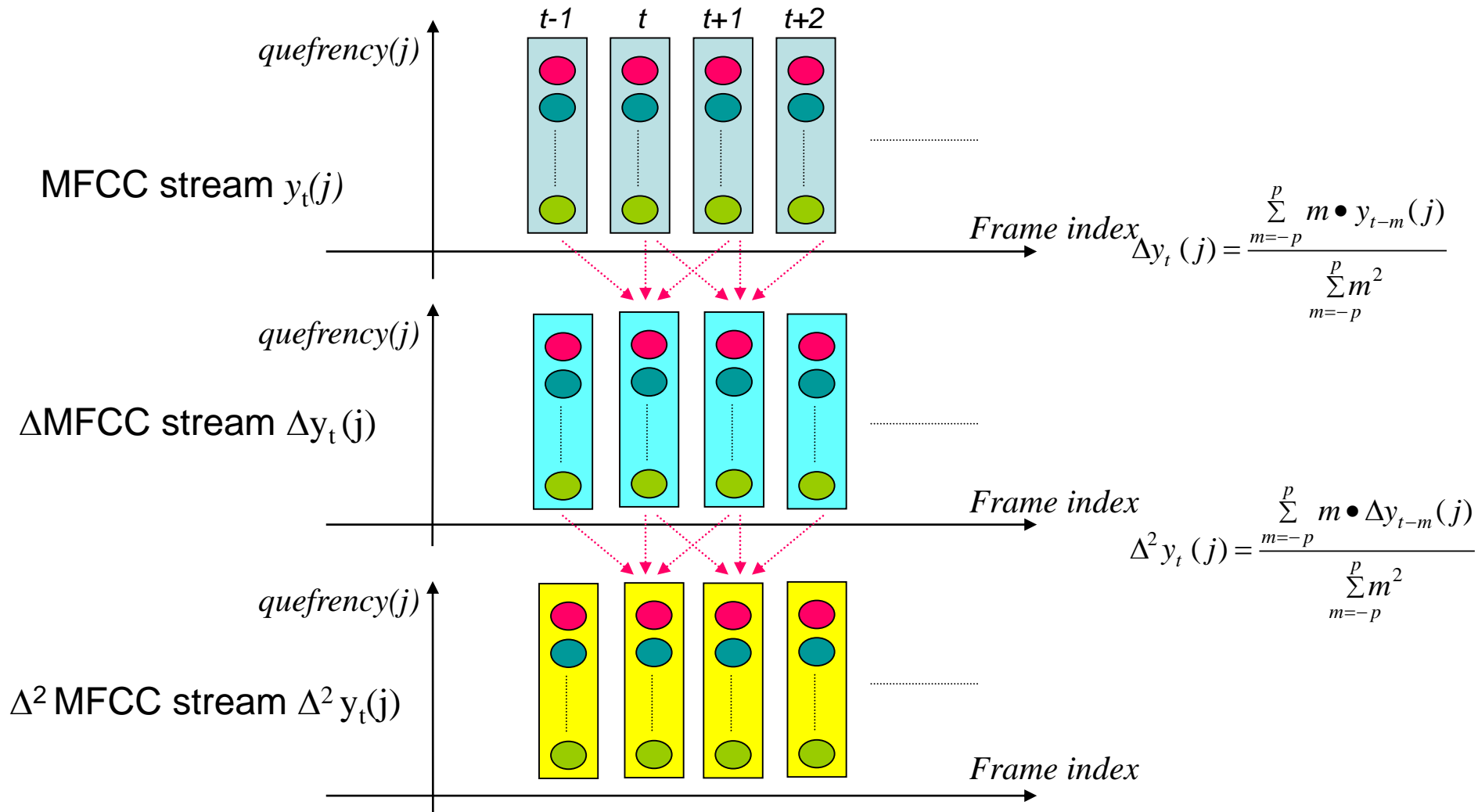
$u(t), U(\omega)$ : excitation

# Logarithmic Operation



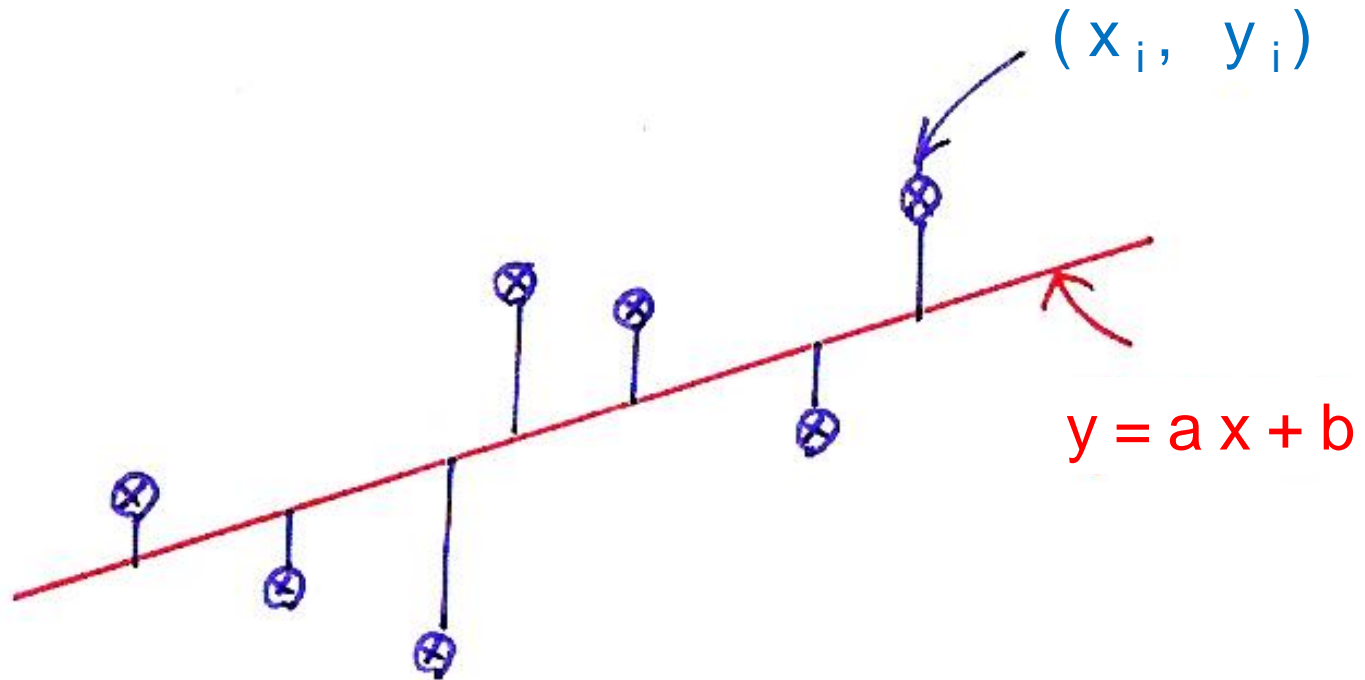
# Derivatives

- Derivative operation : to obtain the change of the feature vectors with time**





# Linear Regression



$$\sum_i (ax_i + b - y_i)^2 = \min$$

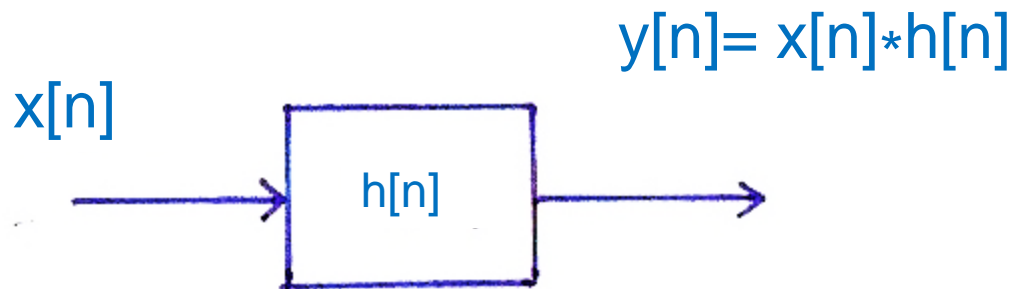
find  $a, b$

# Why Delta Coefficients?

---

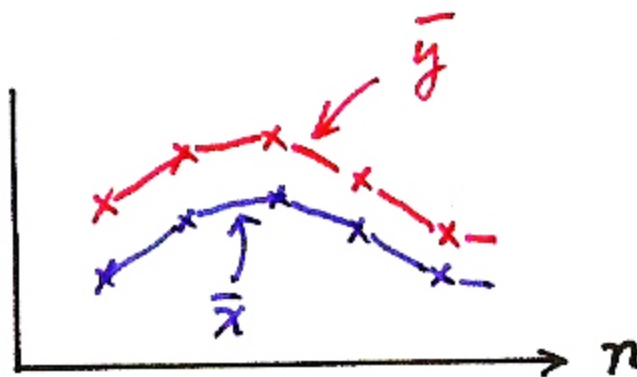
- **To capture the dynamic characters of the speech signal**
  - Such information carries relevant information for speech recognition
  - The value of  $p$  should be properly chosen
    - The dynamic characters may not be properly extracted if  $p$  is too small
    - Too large  $p$  may imply frames too far away
- **To cancel the DC part (channel distortion or convolutional noise) of the MFCC features**
  - Assume, for clean speech, an MFCC parameter stream for an utterance is  $\{\mathbf{y}(t-N), \mathbf{y}(t-N+1), \dots, \mathbf{y}(t), \mathbf{y}(t+1), \mathbf{y}(t+2), \dots\}$ ,  
 $\mathbf{y}(t)$  is an MFCC parameter at time  $t$ ,  
while after channel distortion, the MFCC stream becomes  $\{\mathbf{y}(t-N)+h, \mathbf{y}(t-N+1)+h, \dots, \mathbf{y}(t)+h, \mathbf{y}(t+1)+h, \mathbf{y}(t+2)+h, \dots\}$   
the channel effect  $h$  is eliminated in the delta (difference) coefficients

# Convolutional Noise



MFCC

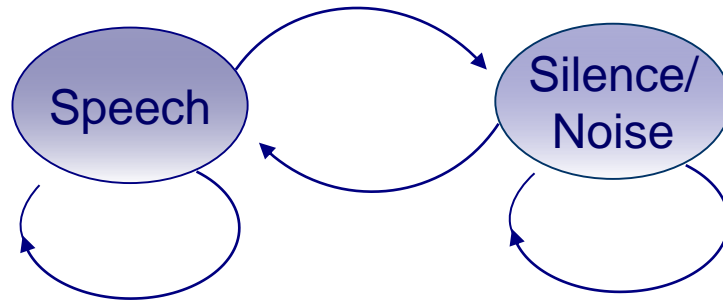
$$\bar{y} = \bar{x} + \bar{h}$$



# End-point Detection

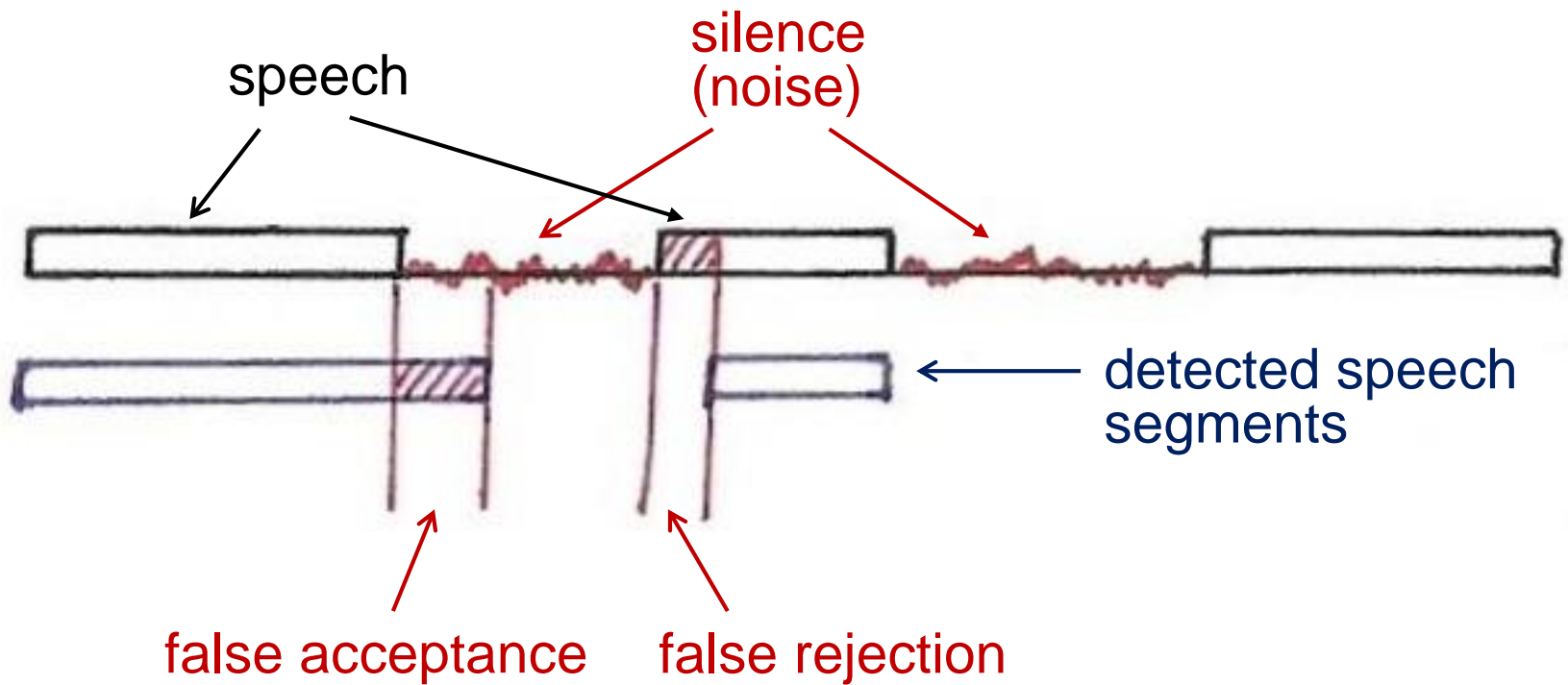
---

- **Push (and Hold) to Talk/Continuously Listening**
- **Keyword Spotting**
- **Adaptive Energy Threshold**
- **Low Rejection Rate**
  - false acceptance may be rescued
- **Vocabulary Words Preceded and Followed by a Silence/Noise Model**
- **Two-class Pattern Classifier**



- Gaussian density functions used to model the two classes
- log-energy, delta log-energy as the feature parameters
- dynamically adapted parameters

# End-point Detection



# 與語音學、訊號波型、頻譜特性有關的網址

17. Three Tutorials on Voicing and Plosives

<http://homepage.ntu.edu.tw/~karchung/intro%20page%2017.htm>

8. Fundamental frequency and harmonics

<http://homepage.ntu.edu.tw/~karchung/phonetics%20II%20page%20eight.htm>

9. Vowels and Formants I: Resonance (with soda bottle demonstration)

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20nine.htm>

10. Vowels and Formants II (with duck call demonstration)

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20ten.htm>

12. Understanding Decibels (A PowerPoint slide show)

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20twelve.htm>

13. The Case of the Missing Fundamental

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20thirteen.htm>

14. Forry, wrong number! I The frequency ranges of speech and hearing

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20fourteen.htm>

19. Vowels and Formants III: Formants for fun and profit (with samples of exotic music)

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20nineteen.htm>

20. Getting into spectrograms: Some useful links

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20twenty.htm>

21. Two other ways to visualize sound signals

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20twentyone.htm>

23. Advanced speech analysis tools II: Praat and more

<http://homepage.ntu.edu.tw/~karchung/Phonetics%20II%20page%20twentythree.htm>

25. Synthesizing vowels online

<http://www.asel.udel.edu/speech/tutorials/synthesis/vowels.html>