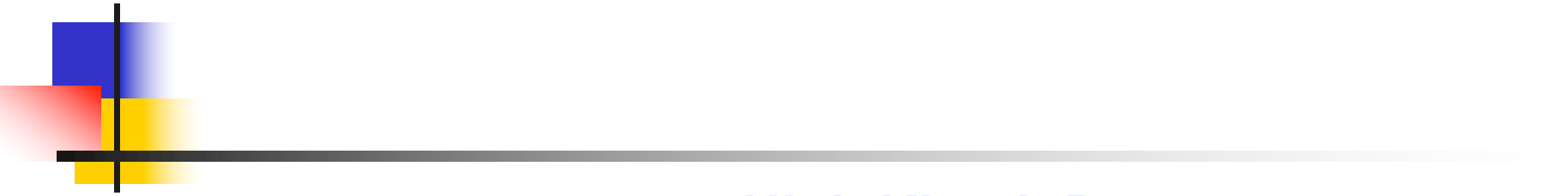
# COURSE:

# DIGITAL SIGNAL PROCESSING



Instructor: Ninh Khanh Duy

# CHAPTER 6:
# SPEECH SIGNAL PROCESSING

**Lecture 6.1:** Introduction to speech signals

**Lecture 6.2:** Time-domain features and applications

**Lecture 6.3:** Frequency-domain features and applications

**Duration:** 6 periods

# Lecture 6.1

# Introduction to speech signals

- **Outline:**

  **1. Overview of speech signals**

  **2.** Basic properties of speech signals

# Overview of speech signals

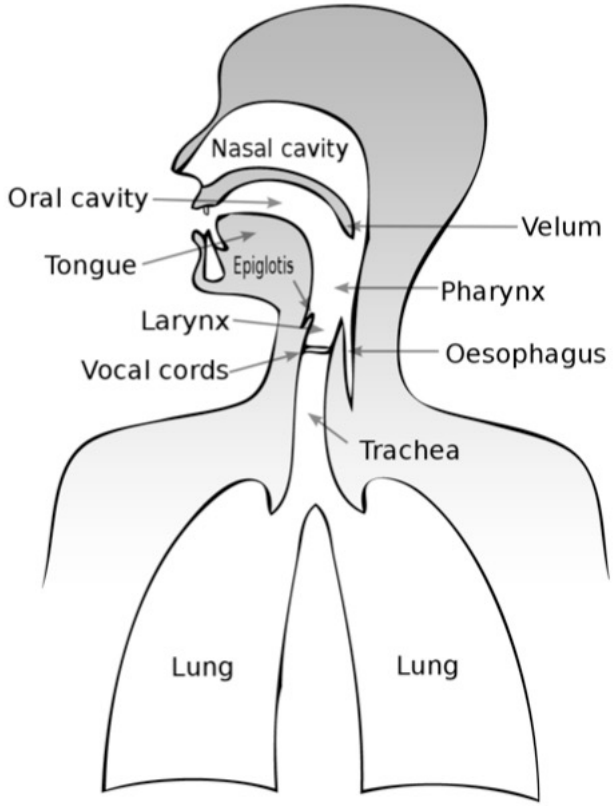
- Speech signals are obtained by a digital recording process (sampling, quantizing, coding) of acoustic waves

"Các bạn trẻ …"

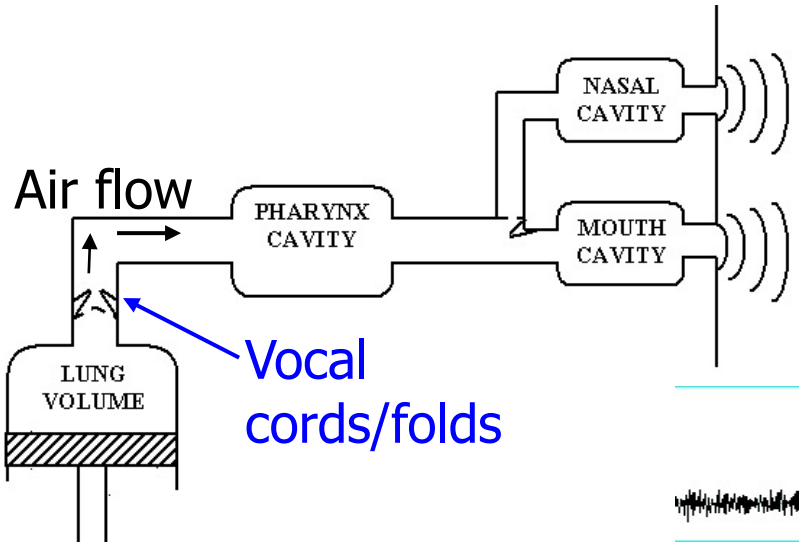
- Speech signals encode messages of speakers, which include linguistic information such as phonemes, sentence types, etc

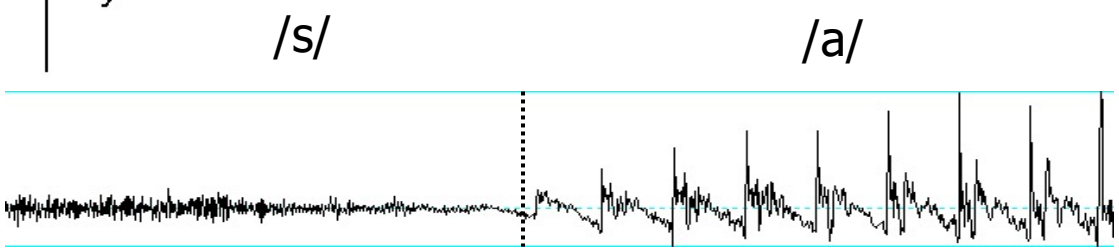
# Overview of speech signals

- Acoustic wave at mouth and nose is the output of the air low going from lung through human vocal tract



# Mechanisms of phones and voicing




- Speech (Output signal): include different phones and voicing
- Resonance cavities (System) ⇒ diff. phones: /a/, /m/, /s/, /z/
- Air flow after vocal cords (Input signal) ⇒ diff. voicing:
  - Vocal cords vibrate: Quasi-periodic pulses ⇒ voiced phones: /a/, /m/
  - Vocal cords close: Turbulence ⇒ unvoiced phones: /s/, /z/, /p/, /k/

# Lecture 6.1

# Introduction to speech signals

- **Outline:**

  1. Overview of speech signals

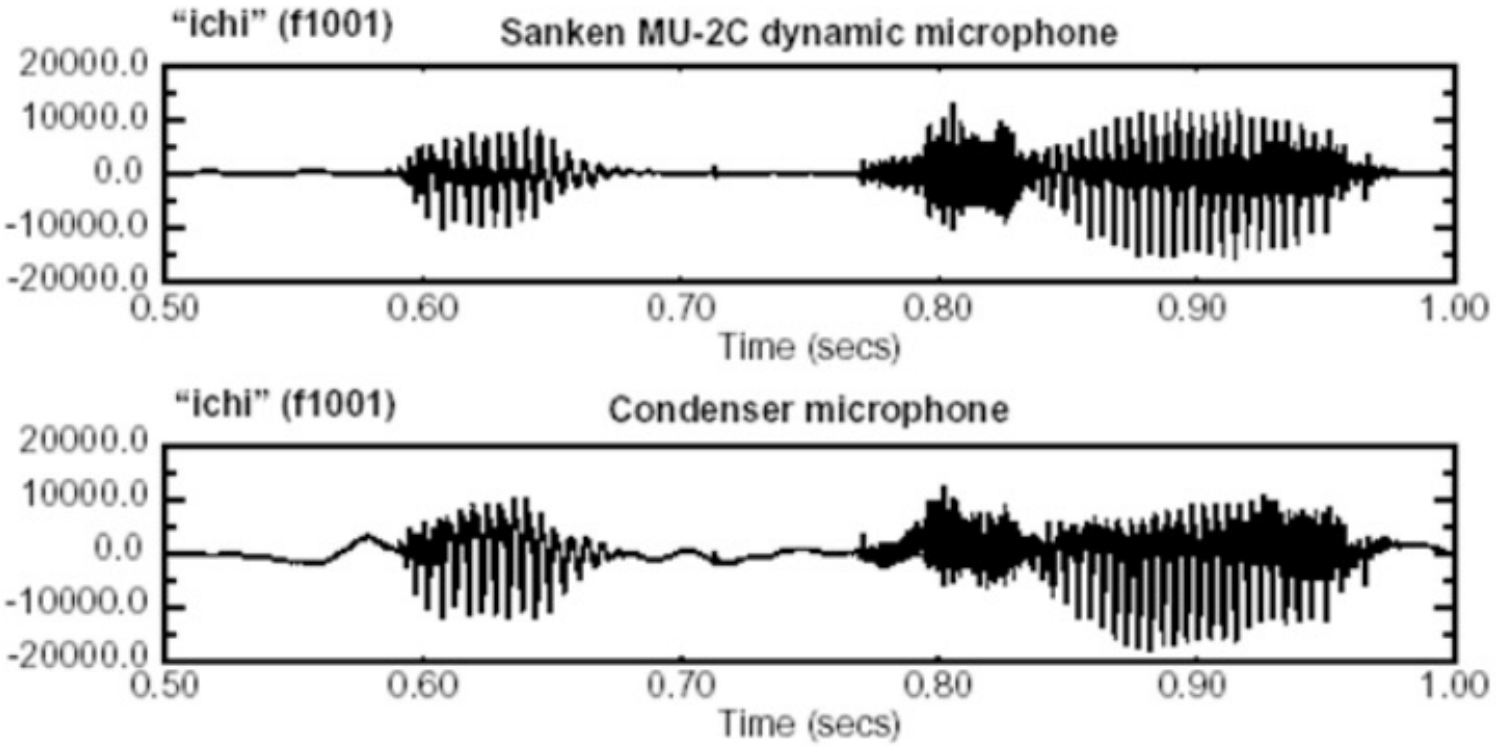
  2. Basic properties of speech signals

# Basic properties of speech signals

- Randomness
  - Speech (like most real-world signals) is random: impossible to predict with certainty their future values from past values
    - *Deterministic signal*: for each value of time we have a rule which enables us to determine the precise value of the signal
  - The value of a signal at any instant of time x(t) is a *random variable*
    - The actual value of a signal is only known after observation
  - A signal is assumed to be generated by a *random process* with a structure that can be characterized and described

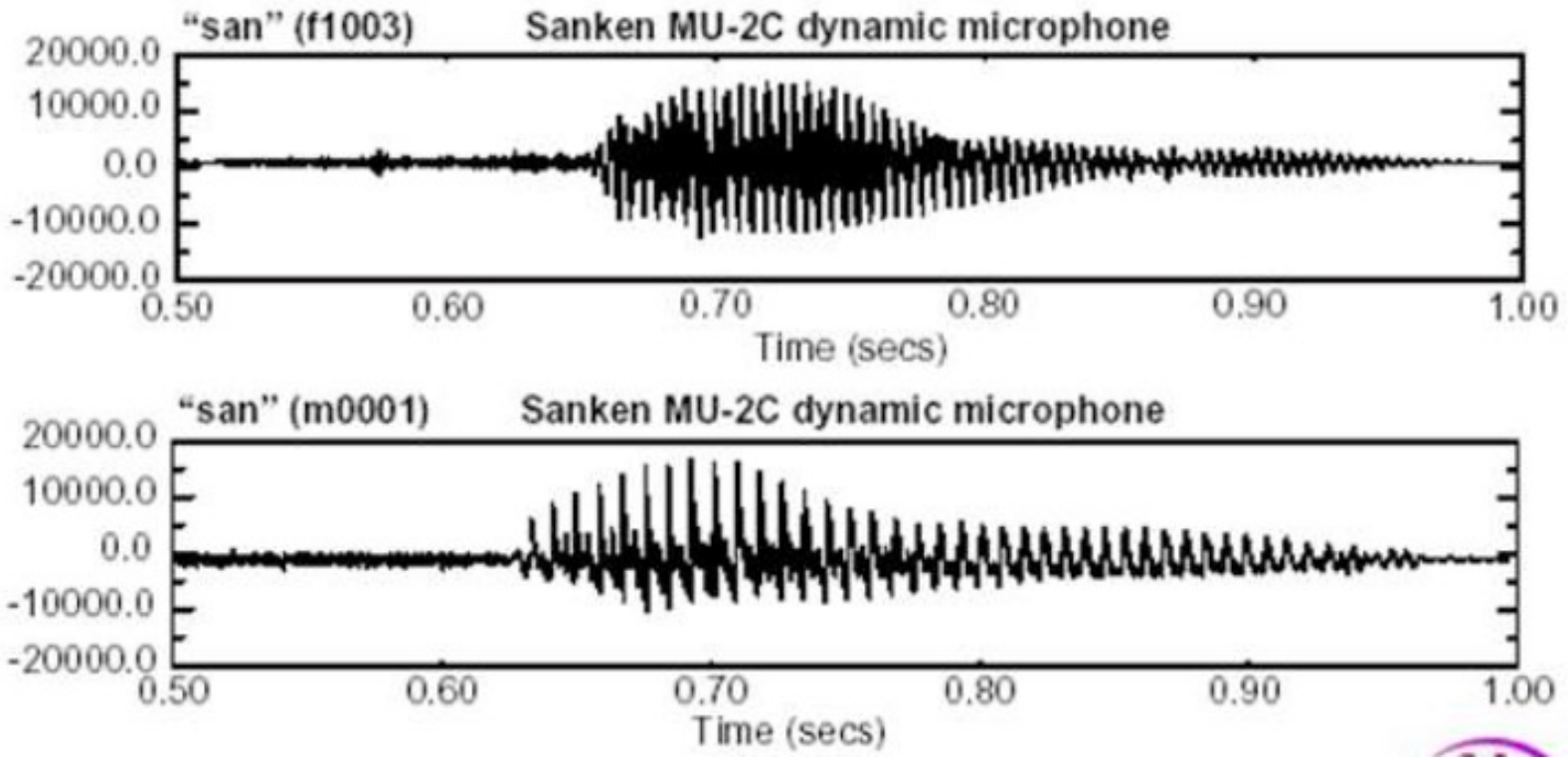
# Basic properties of speech signals

- Variability
  - Depend on different microphones



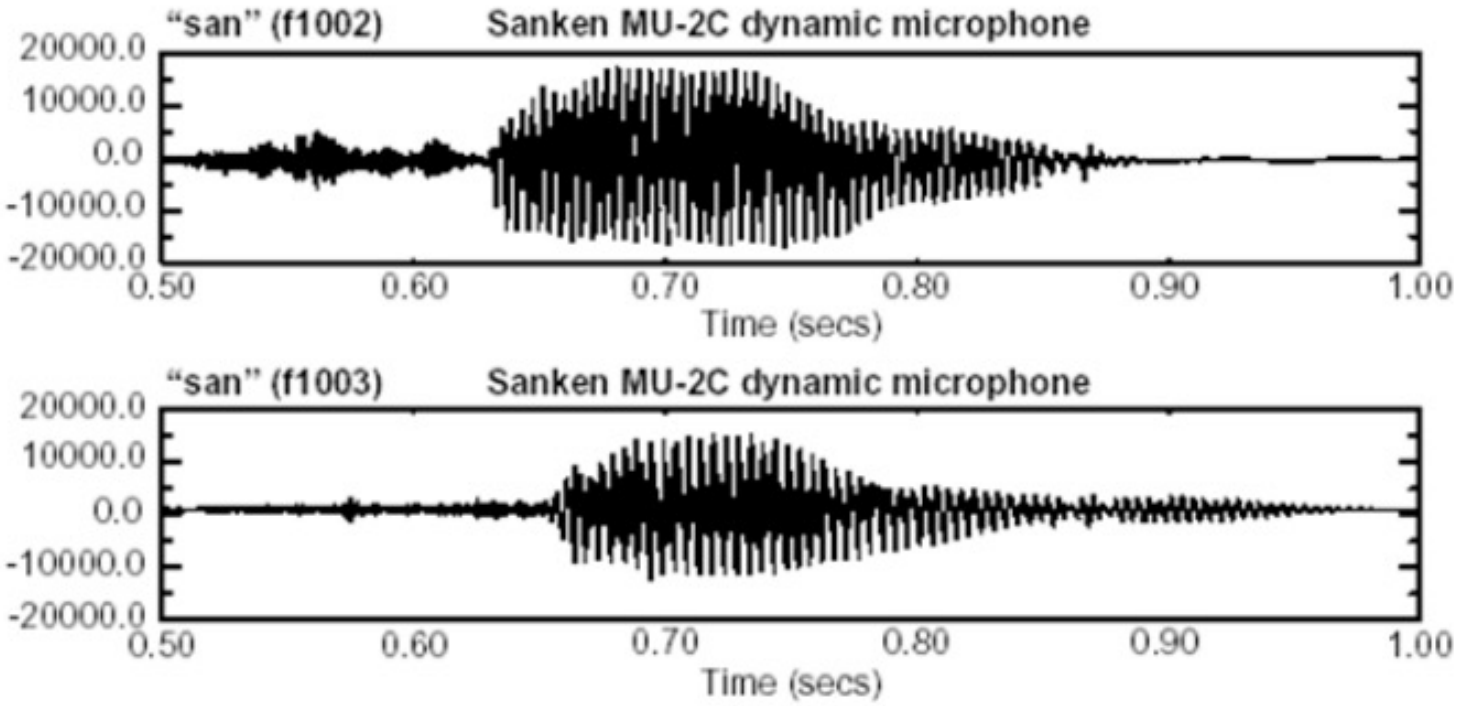
# Basic properties of speech signals

- Variability
  - Depend on different speakers (voices)



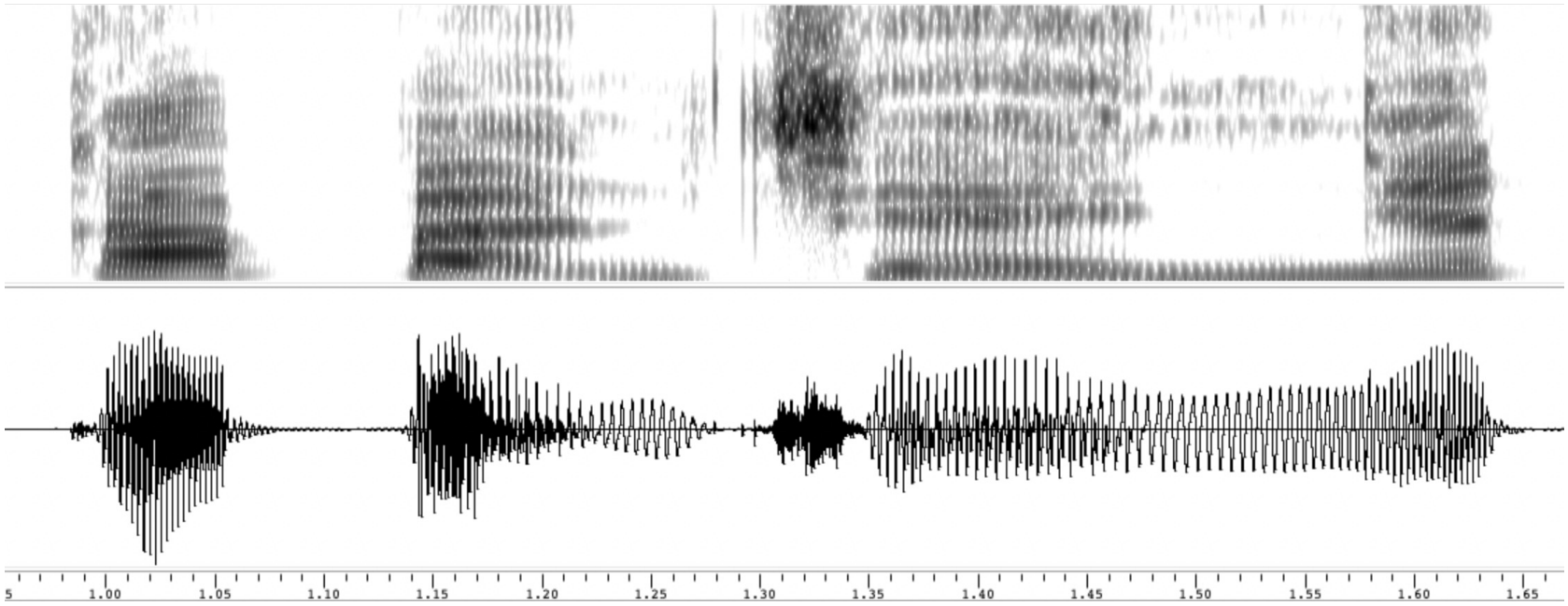
# Basic properties of speech signals

- Variability
  - Depend on dif. physical/emotional states of the same speaker



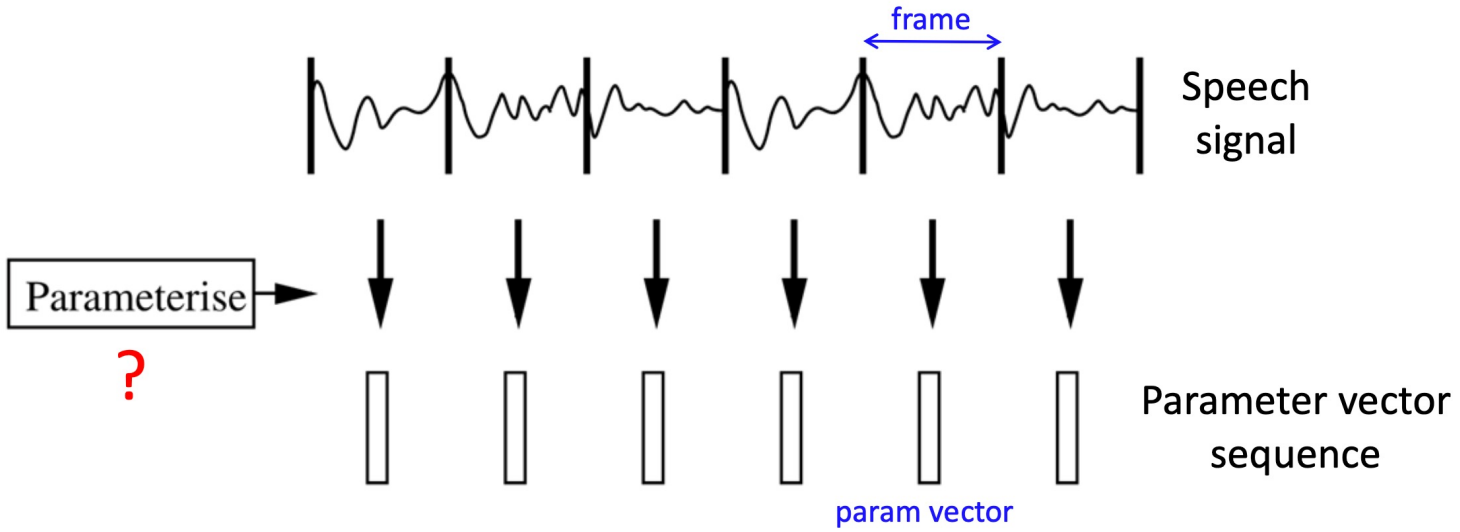
# Basic properties of speech signals

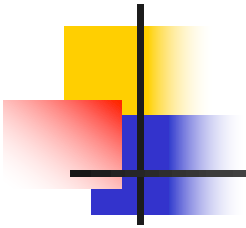
- Characteristics are slowly varying in time
  - Time/Frequency related features are quite stable within short segments of 10-50 ms (duration to pronounce a phoneme)



# Short-time processing technique

- Divide a signal into consecutive frames, each having a fixed duration (e.g., 25 ms)
- Extract features frame-by-frame
- Combine extracted features into feature sequence (time axis is now frame index)





# Homework

1. Read Section 2 & 3 of "CS425 Audio and Speech Processing_Hodgkinson_2012"

2. Write a program to compute the energy and power of a recorded signal following the formulas (2.1) & (2.2) in page 25 of the textbook "Applied Digital Signal Processing -Theory and Practice_Manolakis-Ingle_2011"

# CHAPTER 6:
# SPEECH SIGNAL PROCESSING

**Lecture 6.1:** Introduction to speech signals

**Lecture 6.2:** Time-domain features and applications

**Lecture 6.3:** Frequency-domain features and applications

**Duration:** 6 periods

# Lecture 6.2

# Time-domain features and applications

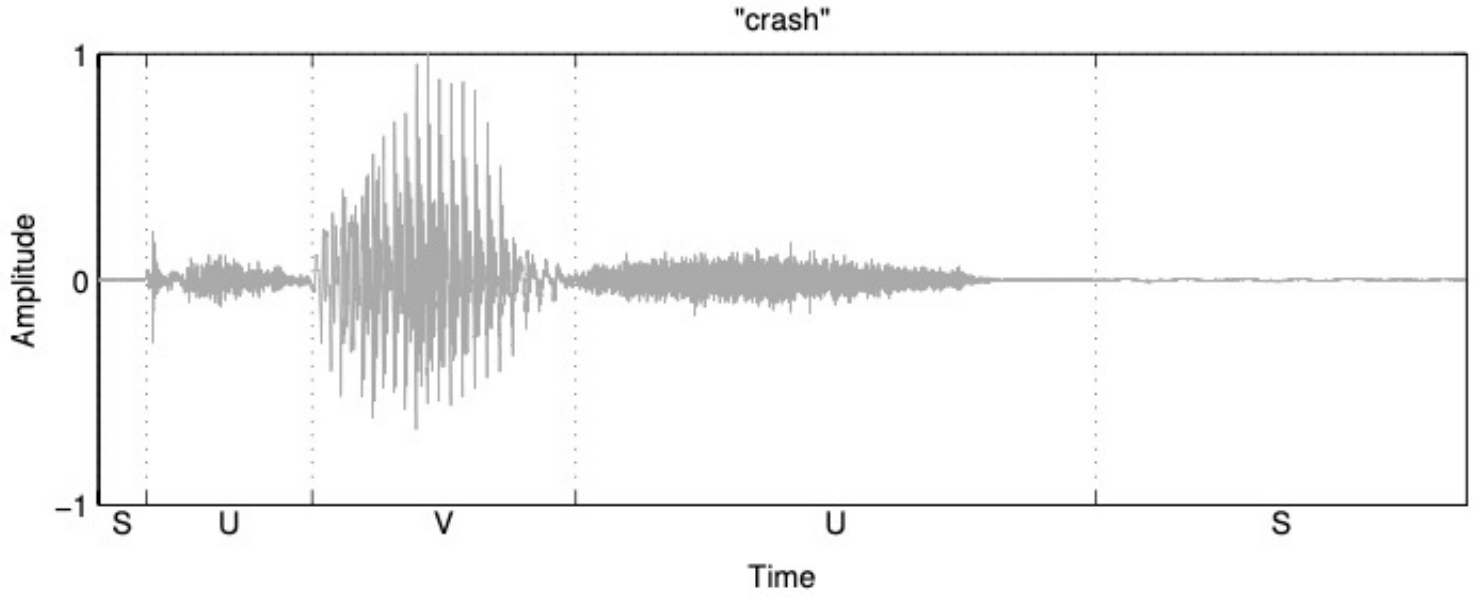
- **Outline:**

  **1. Voiced/Unvoiced/Silence segmentation**

  **2.** Time-domain pitch estimation

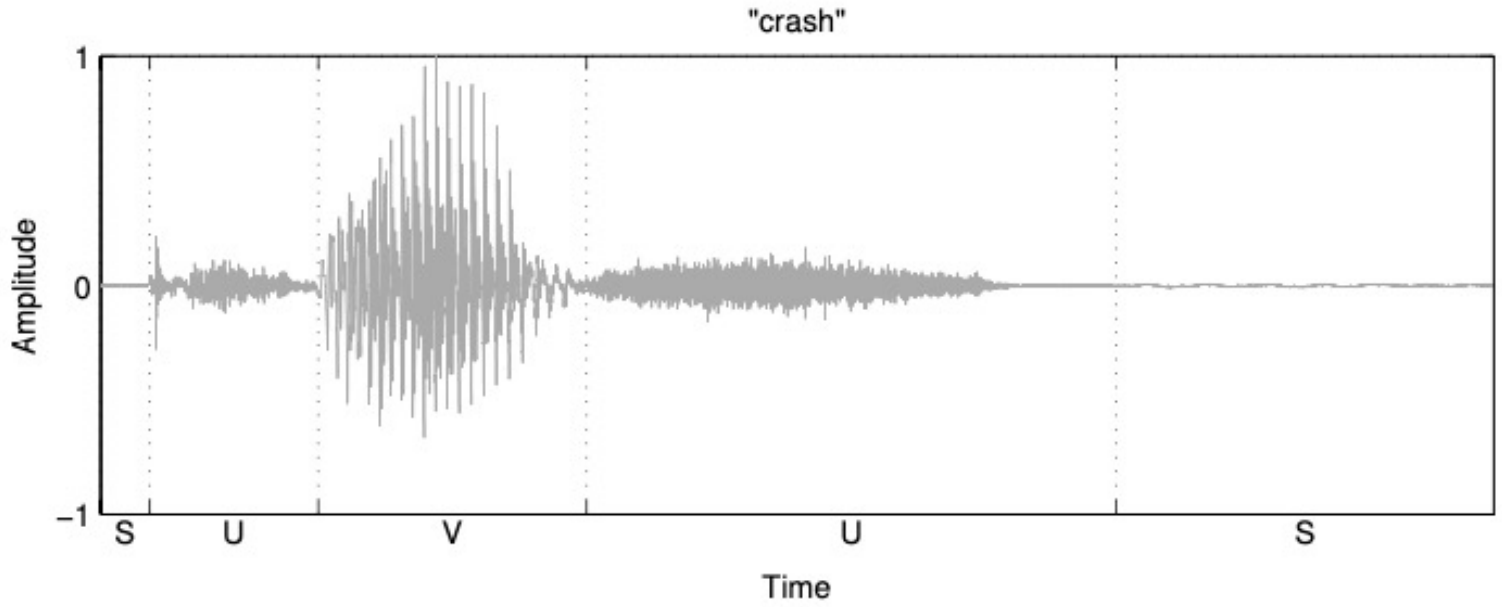
# Introduction to Voiced/Unvoiced/Silence classification

- Recorded signal include speech & silence regions
  - Speech: regions exhibit voice activities (producing phones)
  - Silence: regions exhibit no phone except environmental noise



# Introduction to Voiced/Unvoiced/Silence classification

- A speech region is divided into voiced & unvoiced segments
  - Voiced: exhibit strong periodicity, resulted by vibration of vocal folds
  - Unvoiced: exhibit weak/no periodicity, resulted by closed vocal folds



# Speech/Silence discrimination

- **Problem statement**
  - Input: a signal
  - Output: the signal with vertical boundaries between speech and silence regions
- Constraint
  - The minimum length of silence region is 300ms to exclude very short pauses when speaking

# Speech/Silence discrimination

- Observation

Level of silence is mostly lesser than that of speech segments, except when

- Environmental noise may has level higer than that of unvoiced fricatives (e.g., /s/, /z/)
- Recording environment has a high noise level (or low Signal-to-Noise Ratio (SNR))

→ Use signal level as the discrimination criterion

# Speech/Silence discrimination

- Candidate attribute functions
  - Short-Time Energy (STE): sum of square of the waveform values over a finite number of samples belonging to a frame (20-25 ms)

$$\mathrm{STE}[n] = \sum_{m=0}^{N-1} x^2[n-m]$$

n: frame index
m: sample index
N: frame length (samples)

# Speech/Silence discrimination

- Candidate attribute functions
  - Magnitude Average (MA): sum of absolute of the waveform values over a finite number of samples belonging to a frame

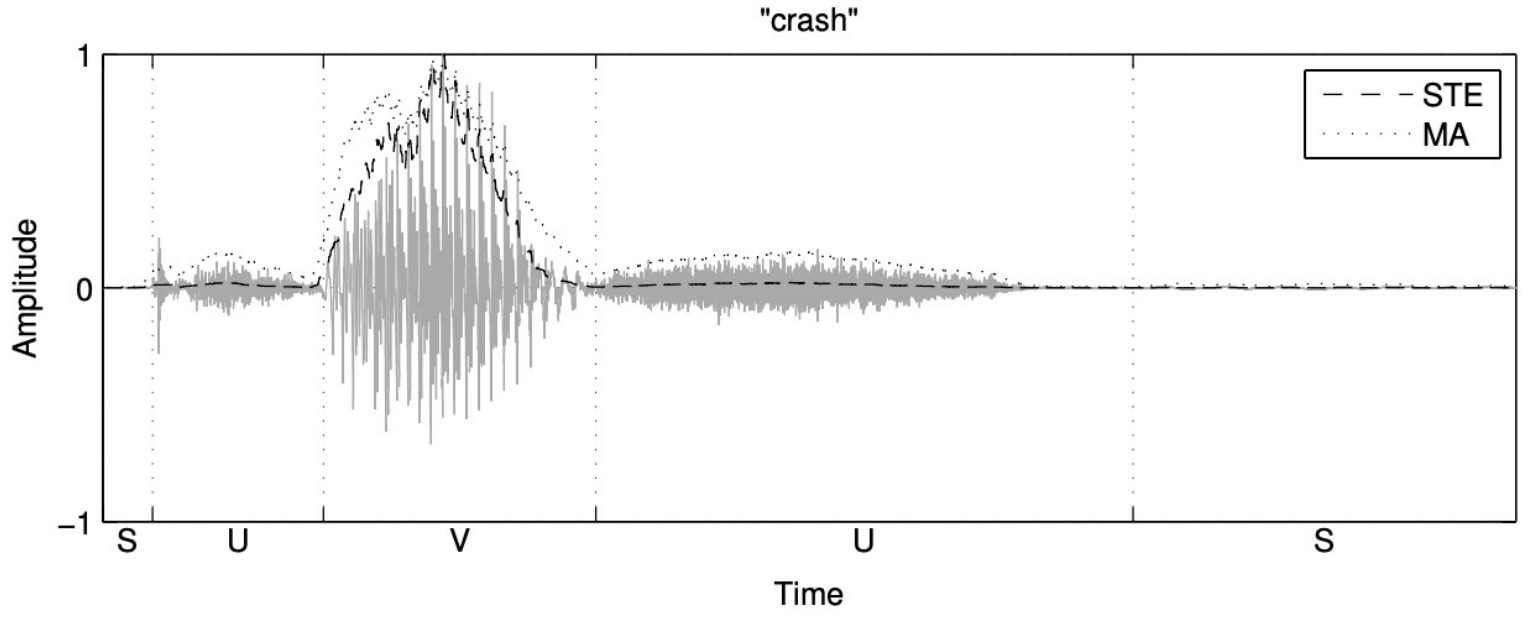
$$\mathrm{MA}[n] = \sum_{m=0}^{N-1} |x[n-m]|$$

n: frame index
m: sample index
N: frame length (samples)

  - For practical uses, we rather use the N values centered around n, from n−N/2 to n+N/2−1

# Speech/Silence discrimination

- Candidate attribute functions
  - Short-Time Energy (STE) vs. Magnitude Average (MA)




Both functions reflect the waveform envelope, but STE emphasizes large values

# Speech/Silence discrimination

- Algorithm in general
  - Based on some threshold of the attribute function to discriminate a frame as speech or silence
  - This threshold is to be found based on given training signals with different environmental noise levels

# Speech/Silence discrimination

- Algorithm to find the threshold
  - Can be set manually or automatically
  - Should be based on the distribution (histogram) of feature data (STE/MA) of frames belong to speech or silence (no label needed), or based on a binary search (label needed)
  - Or should be based on simple statistics (mean & standard deviation) (label needed) (assuming normal distribution)

# Voiced/Unvoiced discrimination

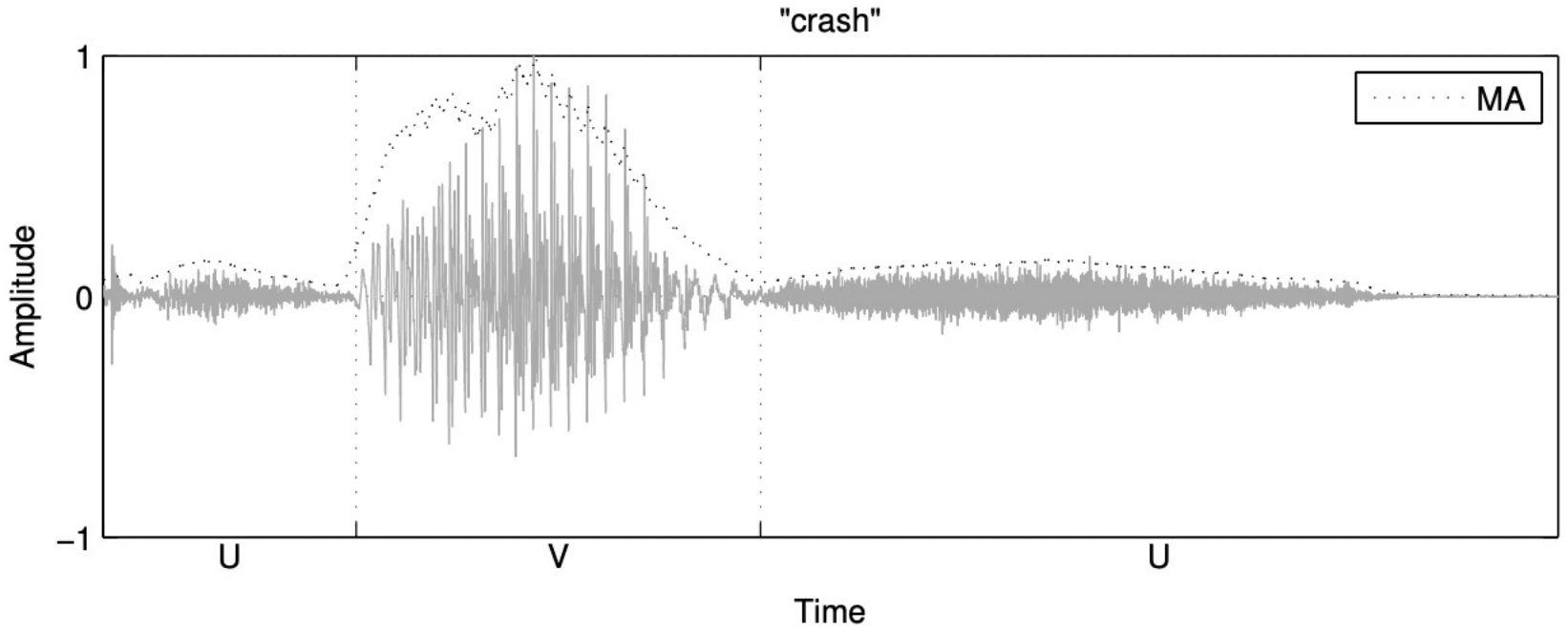
- **Problem statement**
  - Input: a signal including only speech region (assuming no silence)
  - Output: the signal with vertical boundaries between voiced and unvoiced segments
- If input signal includes some silence → no problem because silence is non-periodic & could be considered as unvoiced

# Voiced/Unvoiced discrimination

- Same idea as previous task
  - Look for attributes that characterise contrastingly the states to discriminate
  - Setting for each state a threshold based on training signals
- Different point
  - Combine several features to discriminate voiced vs. unvoiced

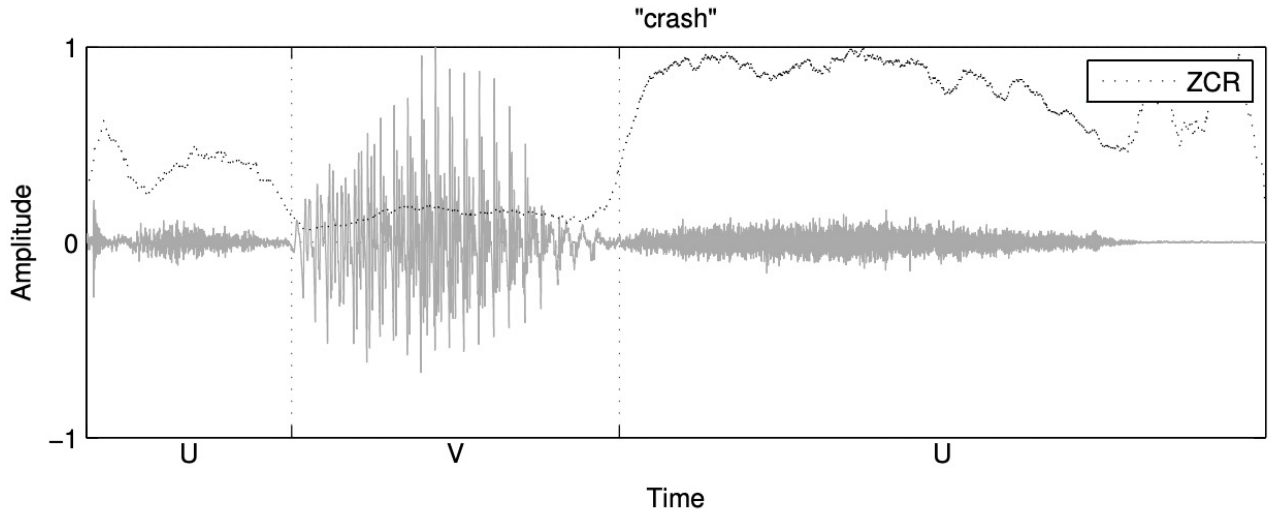
# Voiced/Unvoiced discrimination

- Discriminatory attributes and functions
  - STE or MA: unvoiced segments has level generally lesser than voiced segments



# Voiced/Unvoiced discrimination

- Discriminatory attributes and functions
  - Zero-Crossing Rate (ZCR): the rate at which the waveform crosses the zero-axis
  - Unvoiced segments exhibit a denser waveform, more turbulent than voiced segments → UV has significantly higher ZCR than V



# Voiced/Unvoiced discrimination

- Discriminatory attributes and functions
  - Zero-Crossing Rate (ZCR): the rate at which the waveform crosses the zero-axis

$$\mathrm{ZCR}[n] = \sum_{m=0}^{N-1} |\mathrm{sgn}(x[n-m]) - \mathrm{sgn}(x[n-m-1])|$$

n: frame index
m: sample index
N: frame length

where sgn(.) is the *signum function*,

$$\mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0, \\ -1 & x < 0. \end{cases}$$

# Voiced/Unvoiced discrimination

- Normalisation of attribute functions
  - Useful when combine (e.g., adding) multiple attribute functions into one
  - Then a voicing threshold can be set for the composite function
  - Otherwise, must set various thresholds for dif. attribute functions

# Lecture 6.2

# Time-domain features and applications

- **Outline:**

  1. Voiced/Unvoiced/Silence discrimination

  **2. Time-domain pitch estimation**

# Pitch or Fundamental frequency (F0)

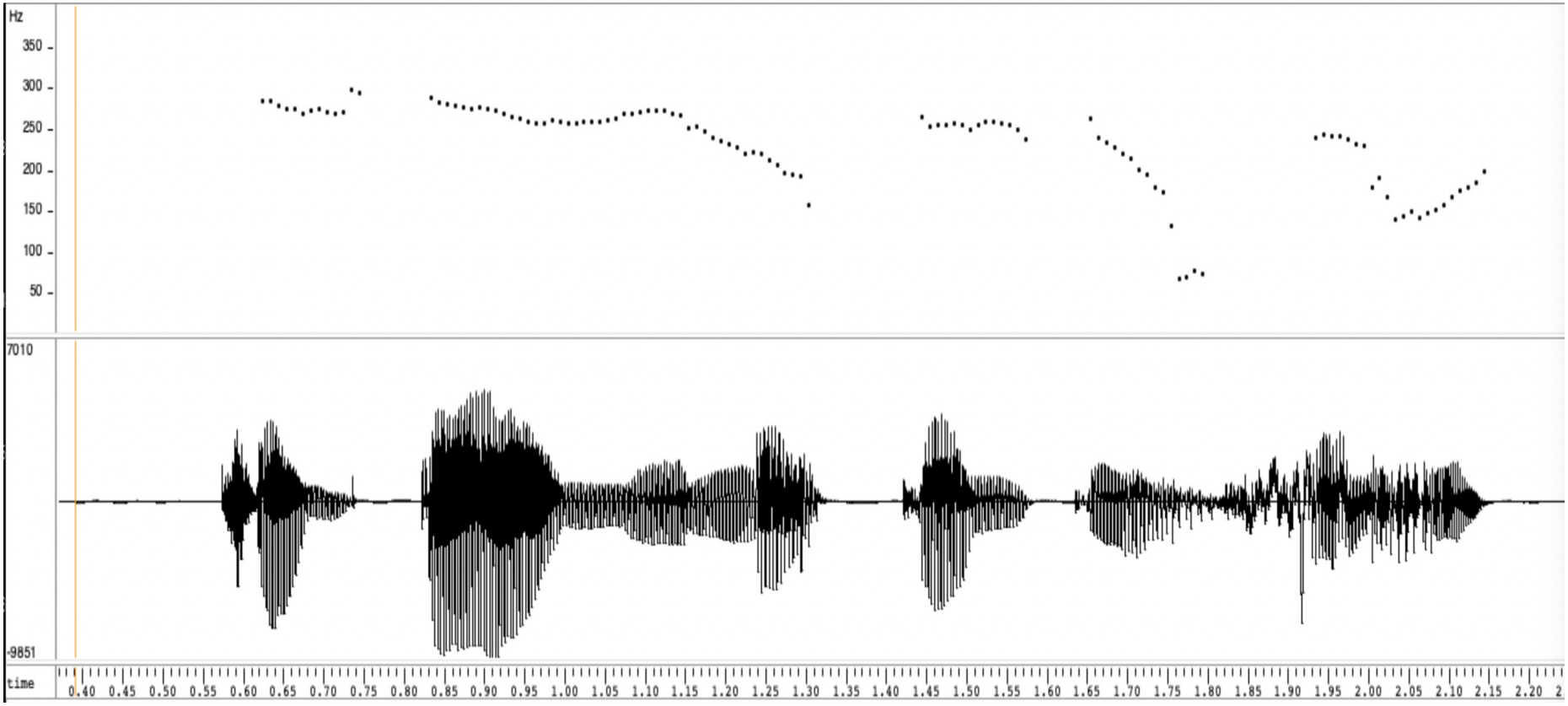
- A feature dedicated only for periodic signals (e.g., voiced segments)
- Definition
  - Fundamental frequency (F0), inverse of the fundamental period, is the number of signal cycles per seconds
    - For speech: F0 is actually the vibration frequency of vocal cords
  - Pitch is the perceptual counterpart of F0 (e.g, high/low-pitched voice)
- Importance
  - Pitch contour conveys the intonation of an utterance (rising/falling)
  - For Vietnamese: 06 tones (ngang, huyền, ngã, hỏi, sắc, nặng)

# Pitch/F0 estimation

- **Problem statement**
  - Input: a signal (may including silence/voiced/unvoiced segments)
  - Output: F0 contour of the signal (a F0 value for each frame)
- Constraint
  - Valid F0 values for adult voices is from 70Hz to 400 Hz

# Pitch/F0 estimation

- An example F0 contour extracted from signal



# Pitch/F0 estimation

- Two time-domain methods
  - Short-Time Autocorrelation function (ACF)
  - Short-Time Average Magnitude Difference Function (AMDF)
- Both based on the following property of periodic signal

$$x[n] = x[n + kN_T], \quad \forall k \in \mathbb{Z}.$$

$N_T$: pitch period/fundamental period (in samples)

- Voiced segments of speech are quasi-periodic

→ "=" never occurs

# Autocorrelation function (ACF)

- The ACF of a signal gives an indication of how alike itself a signal is when shifted
- Definition $$xx[n] = \sum_{m=-\infty}^{\infty} x[m]x[m+n].$$

n: lag/shift
m: sample index

- Application: for a periodic signal x, the ACF is globally maximal at every lag that is an integer multiple of the period
  - For quasi-periodic signal→ local maximal (peak)

# Autocorrelation function (ACF)

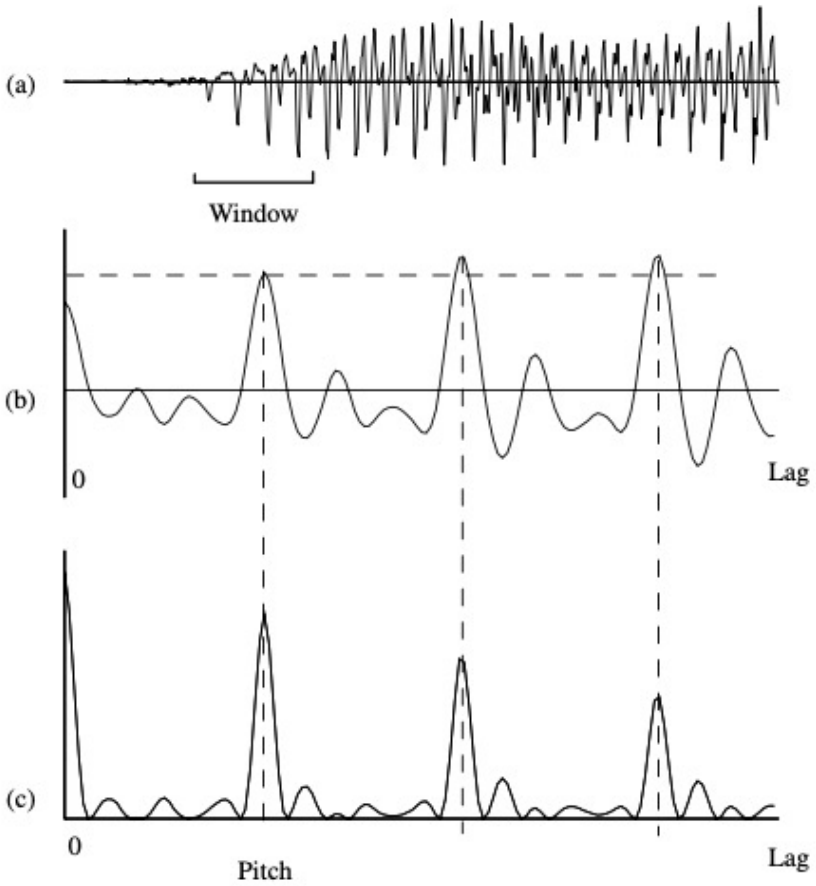
- Short-time ACF of a frame:

$$xx[n] = \sum_{m=0}^{N-1-n} x[m]x[m+n], \quad n \in [0, N[.$$

n: lag (samples)
m: sample index
N: frame length (samples)

- The ACF should be normalized to obtain maximum value of 1 by dividing by ~~largest~~ autocorrelation value at lag zero xx[0]
- Complexity per frame: $O(N^2)$

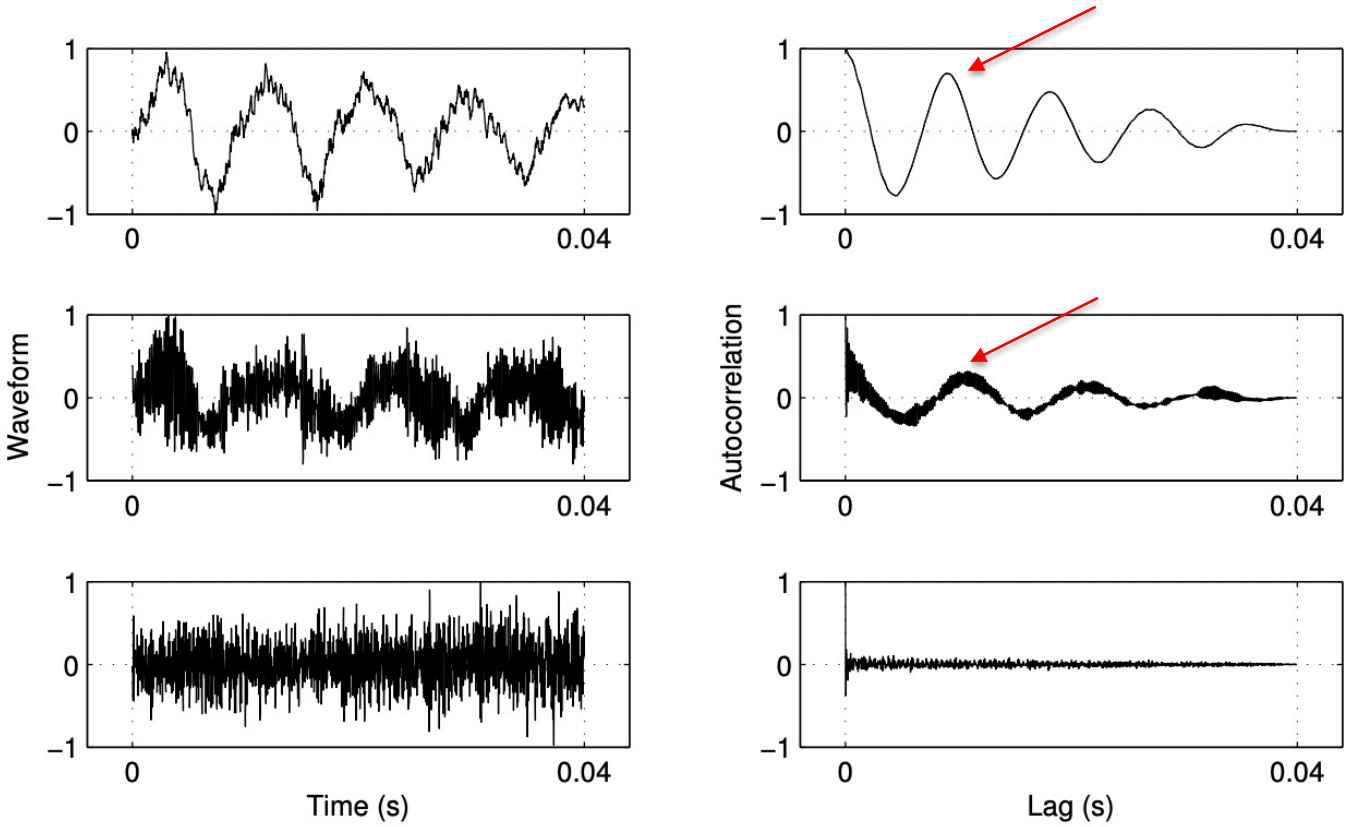
# Short-Time Autocorrelation function




(Kondoz, 2004)

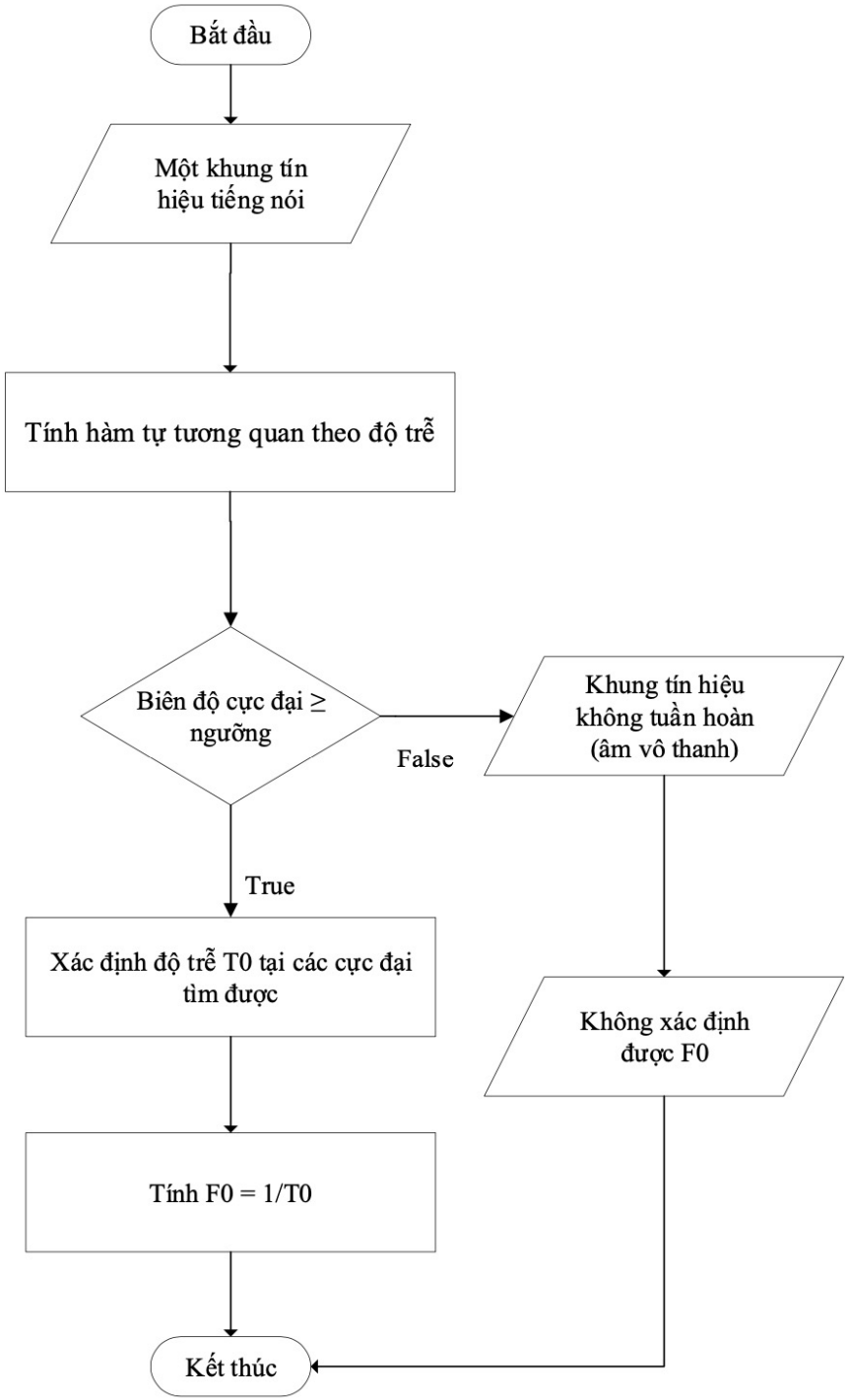
**Figure 6.3** Autocorrelation of speech: (a) original speech, (b) direct autocorrelation function and (c) normalized autocorrelation function

# Short-Time Autocorrelation function




The normalized height of highest local peak is propotional to degree of voicing → can be used for V/U decision

# Algorithm (for a frame)



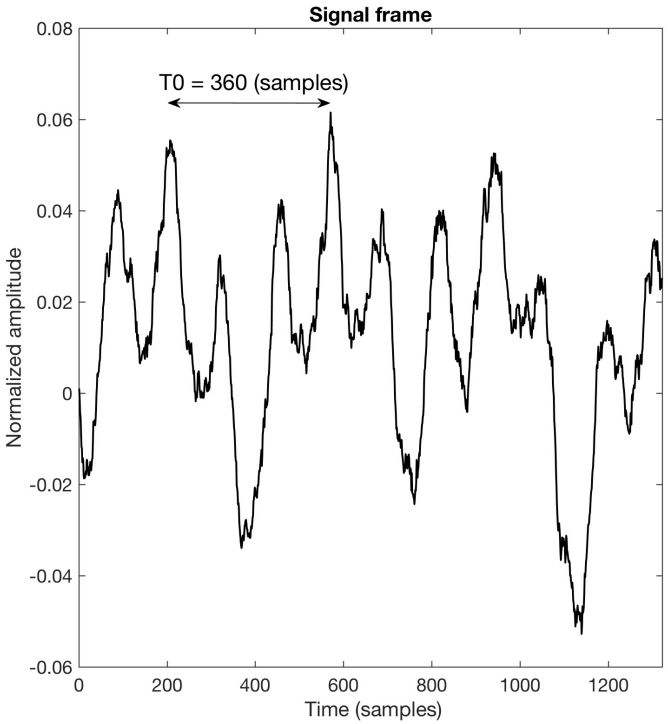

(Trần Văn Tâm, 2019)

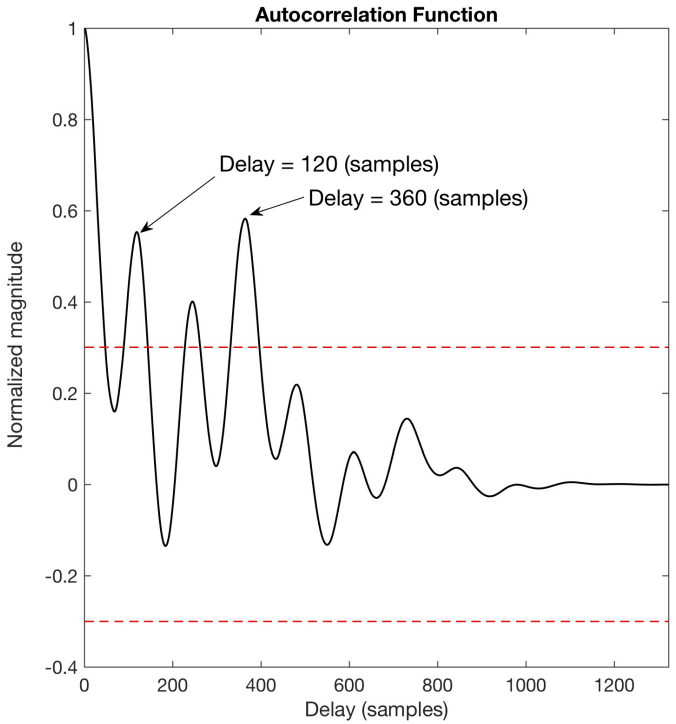
# Short-Time Autocorrelation function

- Autocorrelation peak detection
- Determine a suitable threshod for V/U decision
- Reducing the scope of the search
  - F0 is from 70Hz to 400 Hz → searching range of maximum lag

# Short-Time Autocorrelation function

- Be careful with virtual pitch values

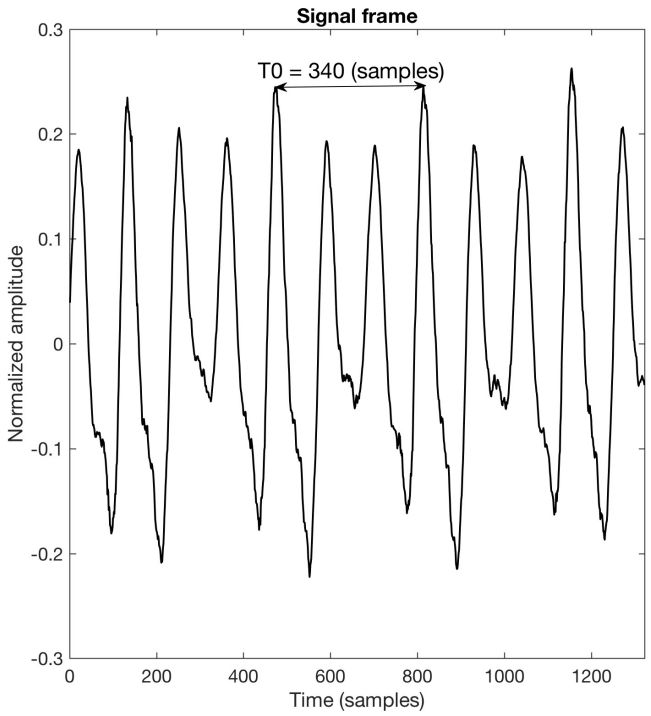


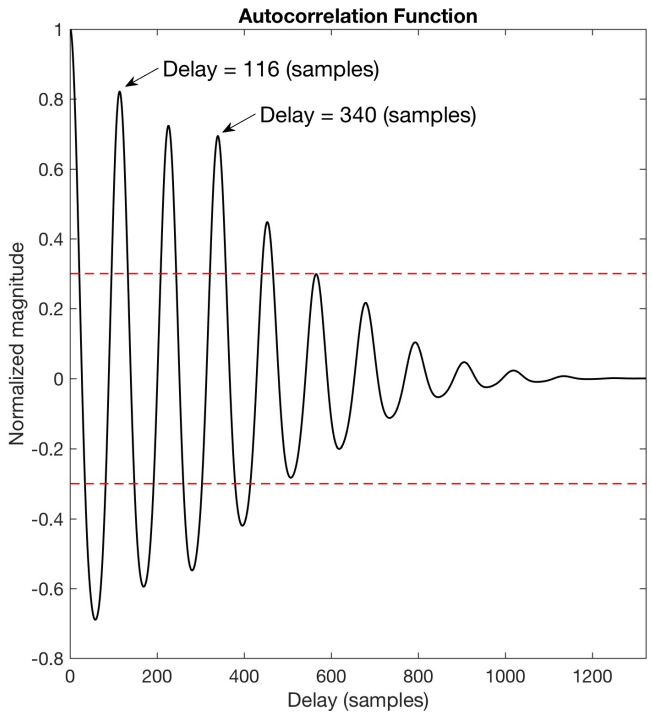




Lucky frame → correct F0

# Short-Time Autocorrelation function

- Be careful with virtual pitch values




Unlucky frame → incorrect F0

# Average Magnitude Difference Function

- The AMDF of a signal gives an indication of how different a signal itself is compared to its shifted version
- Definition

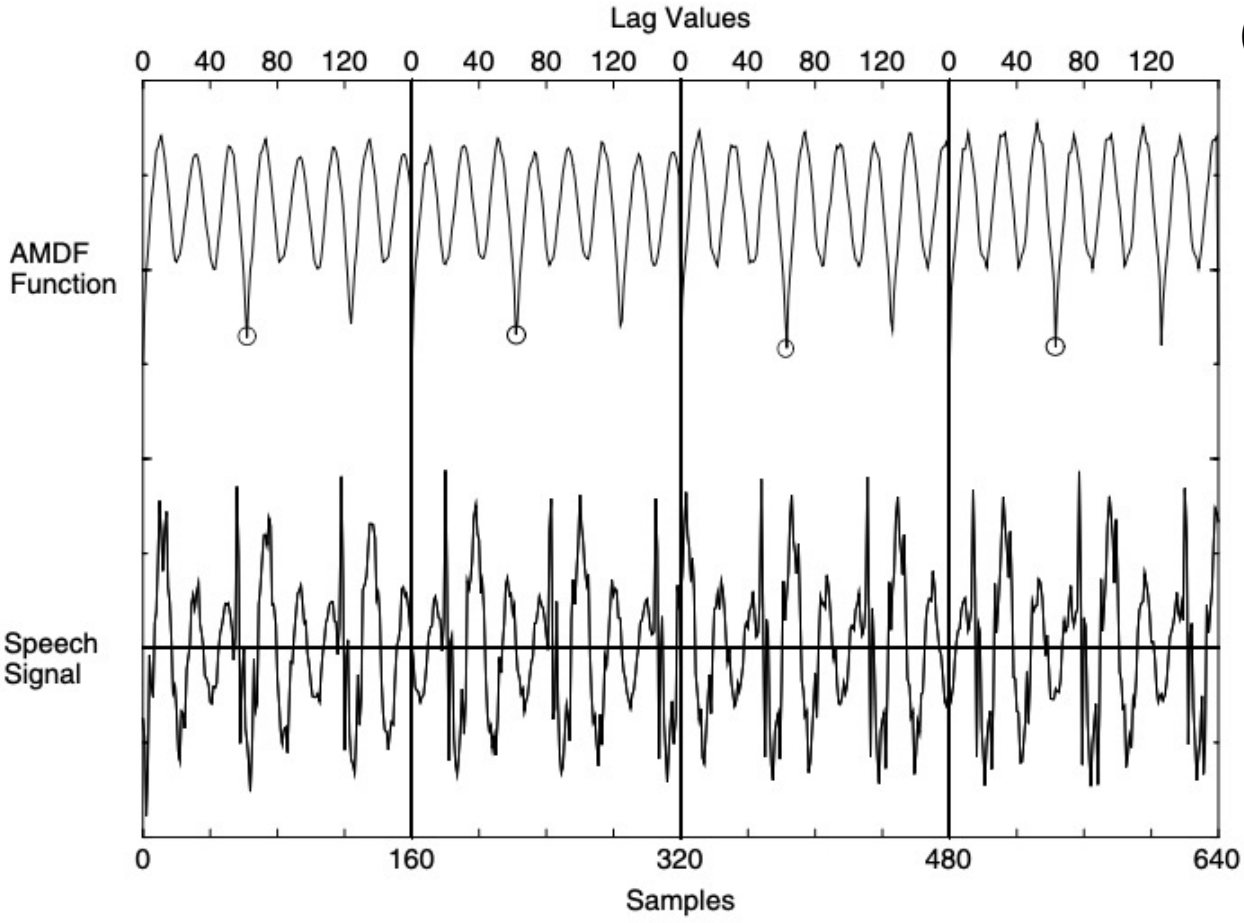
$$d[n] = \sum_{m=-\infty}^{\infty} |x[m] - x[m+n]|.$$

(n: lag, m: sample index)

- Application: for a periodic signal x, the AMDF is zero at every lag that is an integer multiple of the period of the waveform
  - For quasi-periodic signal→ local minimal (dip)

# Average Magnitude Difference Function

(Ex. w/ 4 frames)




(Kondoz, 2004)

**Figure 6.1** AMDF and speech signal: the minima of the AMDF corresponding to the pitch values are indicated by circles

# Average Magnitude Difference Function

- Short-time AMDF of a frame

$$d[n] = \sum_{m=0}^{N-1-n} |x[m] - x[m+n]|, \quad n \in [0, N[.$$

n: lag (samples)
N: frame length (samples)

- Computationally much cheaper than the ACF
- Have similar algorithm & problems to the ACF

# Homework

Các thành viên mỗi nhóm thảo luận và phân công nhiệm vụ, ghi rõ SV nào làm task nào (ko được trùng nhau):
- 1a (phân đoạn speech vs. silence)
- 1b (phân đoạn voiced vs. unvoiced)
- 2a (tính F0 dùng hàm tự tương quan)
- 2b (tính F0 dùng hàm AMDF).

Nhập task (1a/1b/2a/2b) vào link danh sách nhóm.

**Hạn cuối: trước buổi học tuần sau.**

Sau hạn này SV nào ko nhập coi như ko tham gia làm BT nhóm và nhận 0 điểm thi GK.

# CHAPTER 6:
# SPEECH SIGNAL PROCESSING

**Lecture 6.1:** Introduction to speech signals

**Lecture 6.2:** Time-domain features and applications

**Lecture 6.3:** Frequency-domain features and applications

**Duration:** 6 periods

# Lecture 6.3

# Frequency-domain features & applications

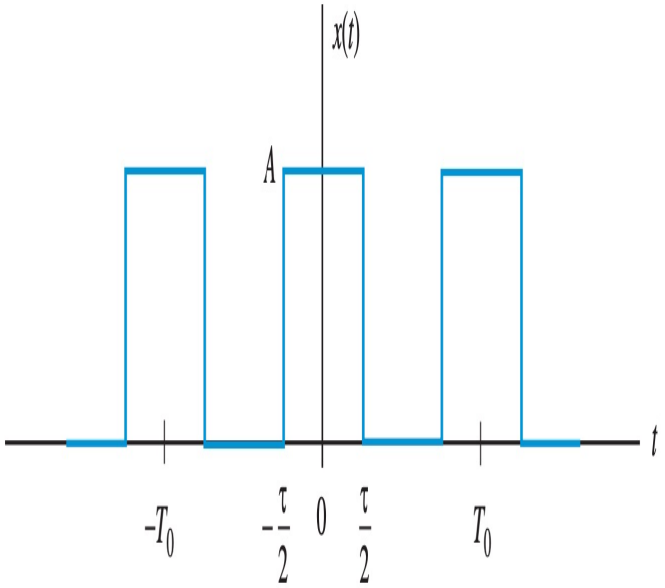
- **Outline:**

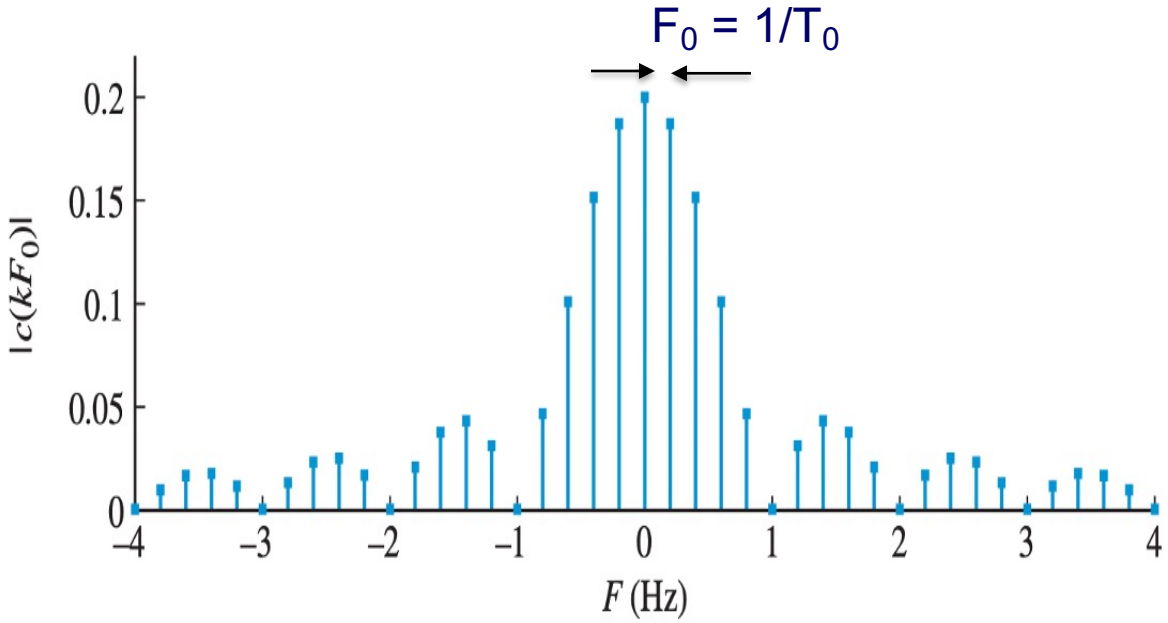
  **1. Frequency-domain pitch (F0) estimation**

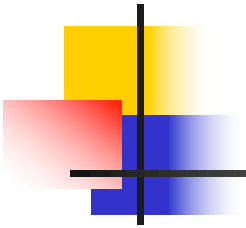
# Theory of CTFS

A periodic signal x(t) has a line spectrum with uniform spacing

$F_0 = 1/T_0$ ($F_0$: fundamental frequency of x(t))





# Main idea

Spectrum of a periodic signal has a harmonic structure with the distance between harmonics being the $F_0$

→ The frame-based solution includes 2 steps:

- Estimate the spectrum using FFT (fast computation of DFT)
- Detect the spacing of adjacent harmonics (i.e., spectral lines)

# Spectrum estimation using FFT

- Important parameters when using function fft(x,N)
  - Window function to reduce spectral leakage (Hamm/Hann)
  - # of FFT points (# of frequency-domain sampling points)
    - Spectral resolution = Sampling frequency / N
    - larger N to have better resolution → more accurate F0 estimates
    - But too large → over-detailed spectrum → harder to detect harmonics
    - ***Should be chosen with high care***
- Log magnitude spectrum should be used for low dynamic range between spectral peaks

# Harmonics spacing detection

- Detect all of harmonic peaks based on estimated spectrum
- Measure the F0 as either the common divisor of these harmonics or the spacing of adjacent harmonics
- Note:
  - Harmonic peaks appear clearer in low-frequency range (<2 kHz)
- Algorithm:
  - Self-proposed (searching for spectral peaks in low-frequency range)
  - Harmonic product spectrum (HPS)