

1



2

### Tài liệu tham khảo

- An Introduction to Digital Speech Processing  
Lawrence R. Rabiner, Ronald W. Schafer, Now.2007
- Digital Processing of Speech Signals  
Lawrence R. Rabiner, Ronald W. Schafer, Prentice-Hall , 1978
- Discrete-Time Processing of Speech Signals  
John R. Deller, John G. Proakis, Hansen John H. L.. IEEE Press, 2000
- Fundamentals of Speech Recognition  
Lawrence Rabiner, Biing-Hwang Juang, Pearson College Div, 1993
- Automatic Speech Recognition: A Deep Learning Approach (Signals and Communication Technology)  
Dong Yu, Li Deng, Springer, 2015
- Text-to-Speech Synthesis  
Paul Taylor, Cambridge University Press, 2009
- Improvements of Vietnamese Hidden Markov Model based speech synthesis  
Duy Khanh Ninh, LAP LAMBERT Academic Publishing, 2020
- Tiếng Việt hiện đại (Ngữ âm, ngữ pháp, phong cách)  
Nguyễn Hữu Quỳnh, Hà Nội, 1994
- Dẫn luận Ngôn ngữ học  
Nguyễn Thị Hiền Giáp, Đoàn Thị Thuật, Nguyễn Minh Thuyết, Hà Nội, 1994

 TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

3

### 1. Các khái niệm cơ bản

- Xử lý tiếng nói là gì ?
- Xử lý tiếng nói bao hàm các lĩnh vực:
  - Nhận dạng tiếng nói
  - Nhận dạng người nói
  - Mã hóa và giải mã tiếng nói
  - Tổng hợp tiếng nói
  - Tăng cường chất lượng tín hiệu tiếng nói

 TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

4

## 1. Các khái niệm cơ bản

- Các ứng dụng của xử lý tiếng nói
  - Tương tác người – máy
  - Viễn thông
  - Các công nghệ trợ giúp (khiếm thính, khiếm thị, học ngôn ngữ)
  - Khai thác dữ liệu âm thanh
  - An ninh, bảo mật
- Các lĩnh vực khoa học liên quan
  - Xử lý tín hiệu số
  - Xử lý ngôn ngữ tự nhiên
  - Học máy
  - Ngữ âm học
  - Tương tác người máy
  - Tâm lý học cảm thụ

5

## 1.2 Các khái niệm cơ bản về tiếng nói

- Phân biệt tiếng nói và âm thanh

Tiếng nói được phân biệt với các âm thanh khác bởi các đặc tính âm học có nguồn gốc từ cơ chế tạo tiếng nói.

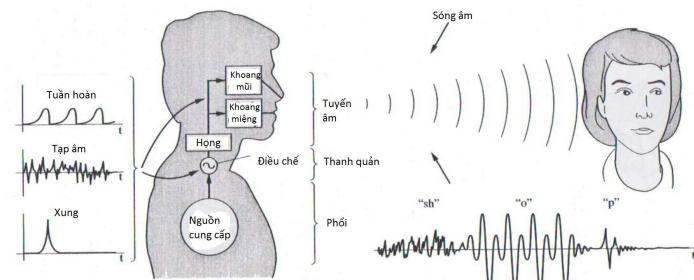
- Có các nguồn âm

- Tuần hoàn (dây thanh rung)
- Tập âm (dây thanh không rung)
- Xung

6

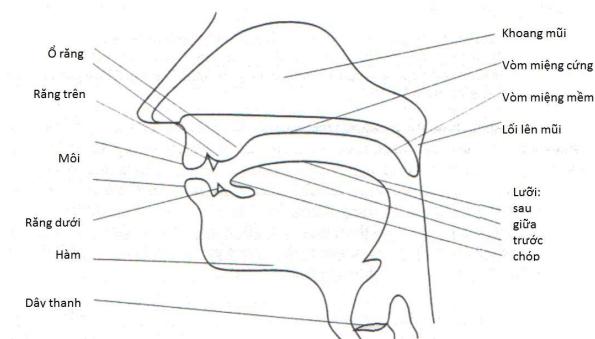
6

## 1.2 Các khái niệm cơ bản về tiếng nói

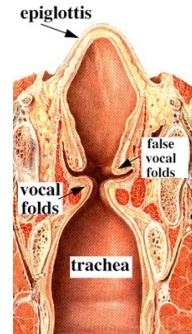


7

## Bộ máy phát âm



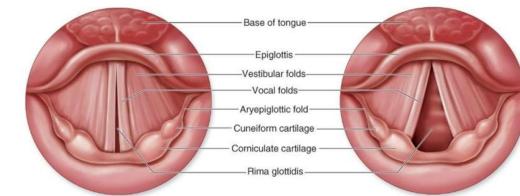
## Bộ máy phát âm



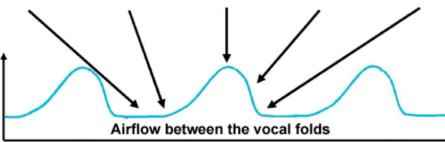
NASAL CAVITY: Khoang mũi  
 SOFT PALATE: Vòm miệng mềm  
 EPIGLOTTIS: Nắp thanh quản  
 VOCAL FOLDS (CORDS): Dây thanh  
 OESOPHAGUS: Thực quản  
 TRACHEA: Khí quản  
 PHARYNX: Họng

9

## Dây thanh và Thanh mòn



Vibration pattern of vocal folds (coronal section)



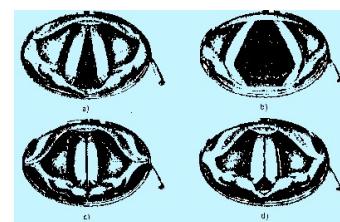
Airflow between the vocal folds

time

10

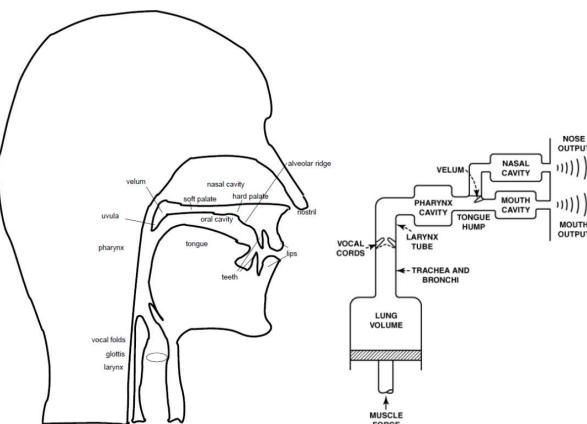
## Thanh mòn

- Ở các vị trí hít, thở, phát âm, nói thì thào



11

## Sơ đồ khái bộ máy phát âm



12

## Hệ thống thính giác

- Hệ thống thính giác có 2 thành phần quan trọng:
  - Cơ quan thính giác ngoại vi (tai)
    - Biến đổi áp suất âm thanh thành dao động cơ học kích thích tế bào thần kinh
  - Hệ thống thần kinh thính giác (não)
    - Trích xuất các thông tin cảm nhận được ở mức độ khác nhau

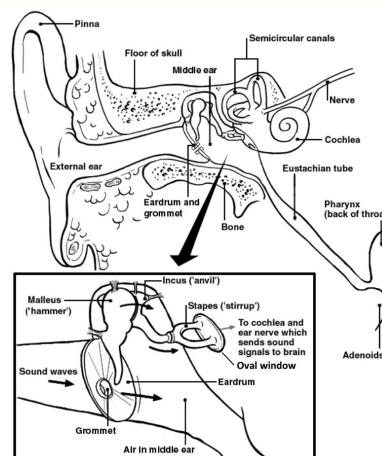
13

## Hệ thống thính giác

- Tai có thể được phân chia
  - Tai ngoài:
    - Bao gồm loa tai, ống tai ngoài và màng nhĩ
    - Biến đổi áp suất âm thanh thành rung động
  - Tai giữa
    - Gồm các xương: xương búa, xương đe và xương bàn đạp
    - Vận chuyển rung động màng nhĩ vào tai trong
  - Tai trong:
    - Gồm ốc tai
    - Biến đổi các rung động thành các xung kích thích màng đáy
    - Màng đáy có thể được mô hình hóa như băng bộ lọc

14

## Hệ thống thính giác

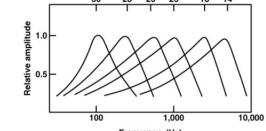
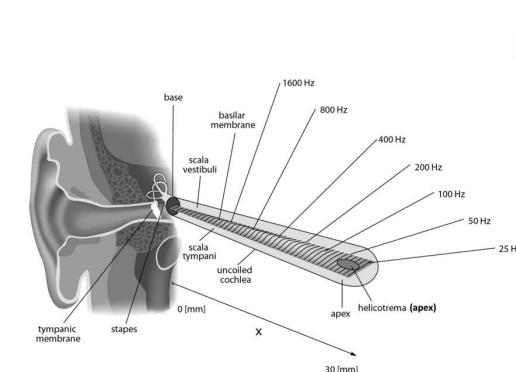


Pinna: Loa tai  
 Middle ear: Tai giữa  
 External ear: Tai ngoài  
 Eardrum: Màng nhĩ  
 Cochlea: Ốc tai  
 Nerve: Dây thần kinh

Malleus: Xương búa  
 Incus: Xương đe  
 Stapes: Xương bàn đạp  
 To cochlea and ear nerve which sends sound signals  
 to brain: Dẫn tới ốc tai và thần kinh tai gửi tín hiệu âm thanh tới não  
 Oval window: Cửa sổ bầu dục

15

## Hệ thống thính giác



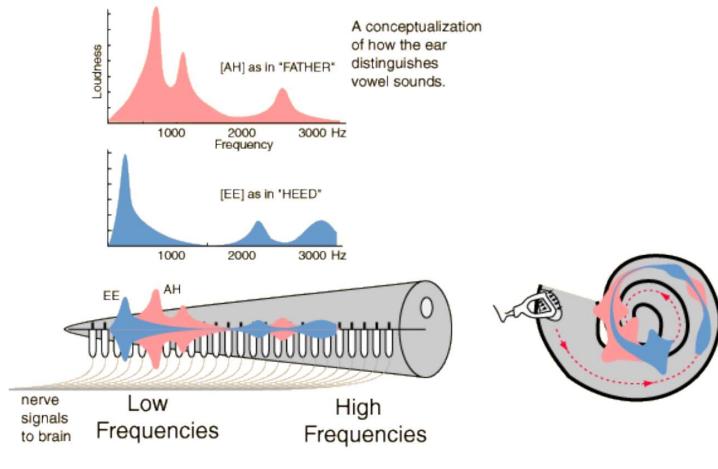
Scala vestibuli: Ông tiền đình  
 Basilar membrane: Màng đáy  
 Scala tympani: Màng định âm  
 Uncoiled cochlea: Ốc tai duỗi thẳng  
 Apex: Đỉnh  
 Stapes: Xương bàn đạp

16

15

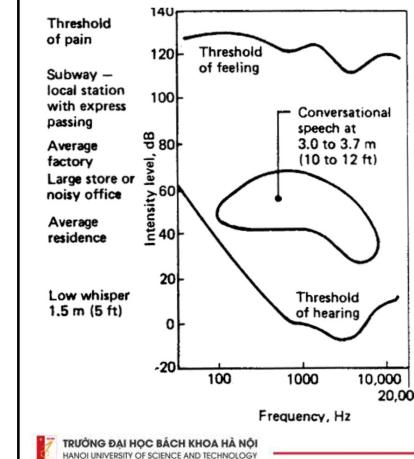
16

## Hệ thống thính giác



17

## Hệ thống thính giác



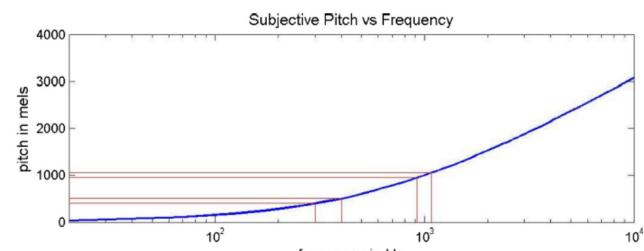
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

18

18

## Cảm nhận cao độ (pitch)

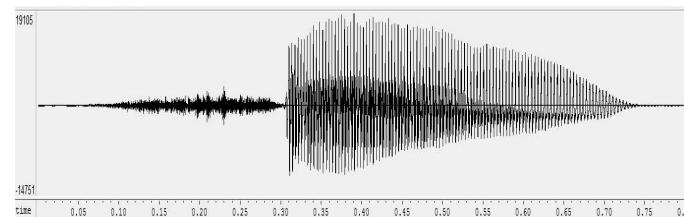
- Cao độ là F0 (tần số cơ bản) được con người cảm nhận, mang tính chủ quan



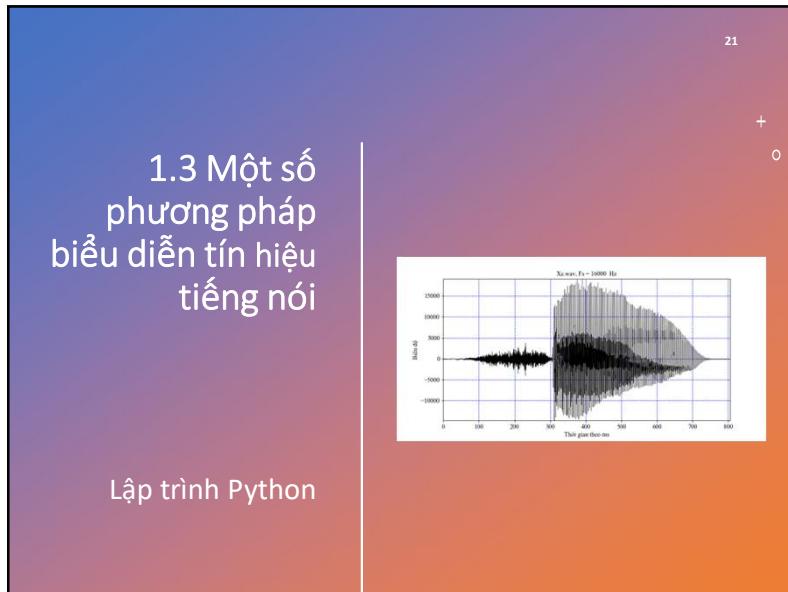
19

## 1.3 Một số phương pháp biểu diễn tín hiệu tiếng nói

- Dạng sóng theo thời gian



20



21

```

21
+ 0
import numpy as np
import scipy.io.wavfile as wf
import matplotlib.pyplot as plt
filename = "Xe.wav"
rate, data = wf.read(filename)
data = data.astype(np.int32)

# Timing axis
Time = np.linspace(0,1000 * len(data) / rate, num=len(data))
fig = plt.figure(figsize=(10, 5))
plt.plot(Time, data, color='k', linewidth=0.5)
plt.ylim(min(data), max(data))
plt.xlabel('Thời gian (ms)')
plt.title('Xe.wav, Fs = 16000 Hz')
plt.show()

# Frequency axis
f = np.fft.rfft(data)
frequencies = np.fft.fftfreq(len(f), 1/rate)
frequencies = frequencies[1:]
amplitude = np.abs(f)
plt.plot(frequencies, amplitude, color='r')
plt.title('Spectrum of Xe.wav')
plt.xlabel('Frequency in Hz')
plt.ylabel('Magnitude')
plt.show()

```

Lập trình Python

```

import numpy as np
import scipy.io.wavfile as wf
import matplotlib.pyplot as plt
filename = "Xe.wav"
rate, data = wf.read(filename)
data = data.astype(np.int32)

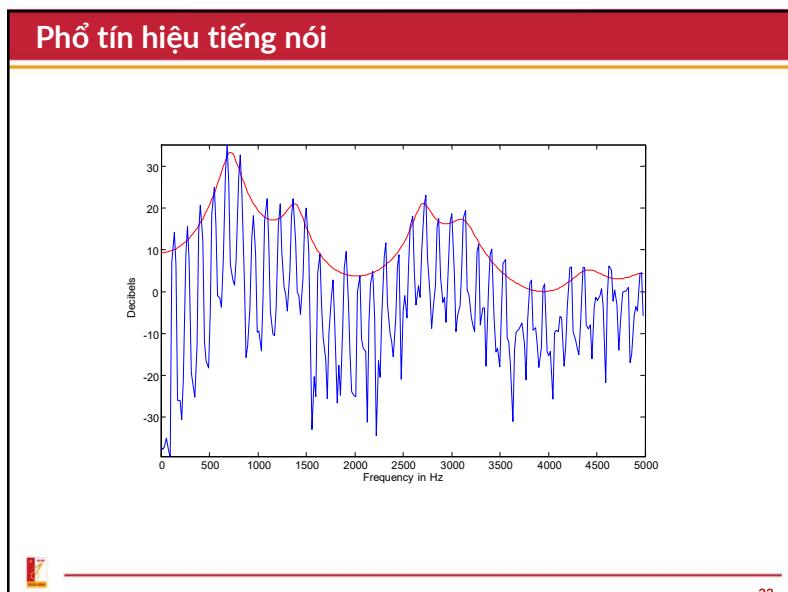
# Timing axis
Time = np.linspace(0,1000 * len(data) / rate, num=len(data))
fig = plt.figure(figsize=(10, 5))
plt.plot(Time, data, color='k', linewidth=0.5)
plt.ylim(min(data), max(data))
plt.xlabel('Thời gian (ms)')
plt.title('Xe.wav, Fs = 16000 Hz')
plt.show()

# Frequency axis
f = np.fft.rfft(data)
frequencies = np.fft.fftfreq(len(f), 1/rate)
frequencies = frequencies[1:]
amplitude = np.abs(f)
plt.plot(frequencies, amplitude, color='r')
plt.title('Spectrum of Xe.wav')
plt.xlabel('Frequency in Hz')
plt.ylabel('Magnitude')
plt.show()

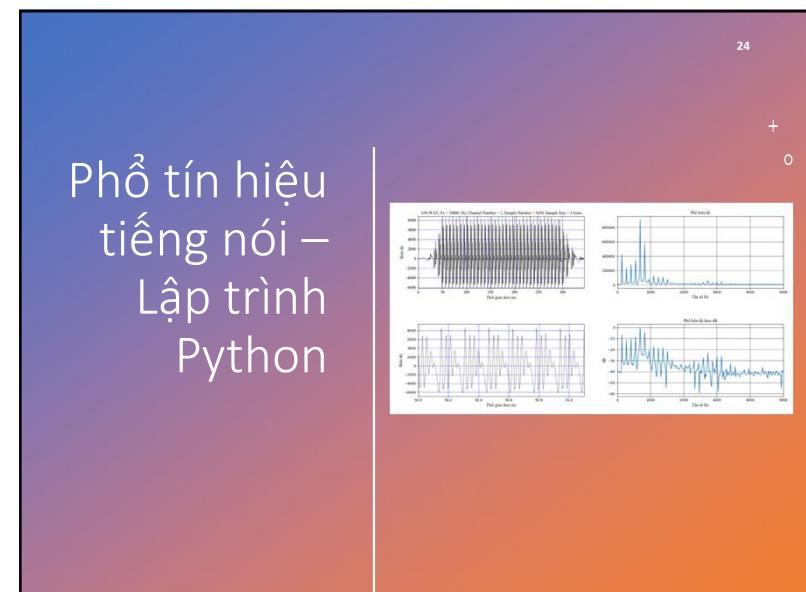
# Mirror operation
def mirror_object_to_mirror(mirror_mod, mirror_object):
    if mirror_mod == "MIRROR_X":
        mirror_mod.use_x = True
        mirror_mod.use_y = False
        mirror_mod.use_z = False
    elif mirror_mod == "MIRROR_Y":
        mirror_mod.use_x = False
        mirror_mod.use_y = True
        mirror_mod.use_z = False
    elif mirror_mod == "MIRROR_Z":
        mirror_mod.use_x = False
        mirror_mod.use_y = False
        mirror_mod.use_z = True
    else:
        print("Selection at the end - add")
        ob.select = 1
        ob.select = 1
        context.scene.objects.active = "Selected" + str(modifier)
        mirror_mod.select = 0
        bpy.context.selected_objects.append(data.objects[one.name].select)
        print("int(\"please select exactly one object\")")
        print("OPERATOR CLASSES -----")
        print("types.Operator):")
        print("    X mirror to the selected object.mirror_mirror_x")
        print("    mirror_X")
        print("context):")
        print("    context.active_object is not None")

```

22

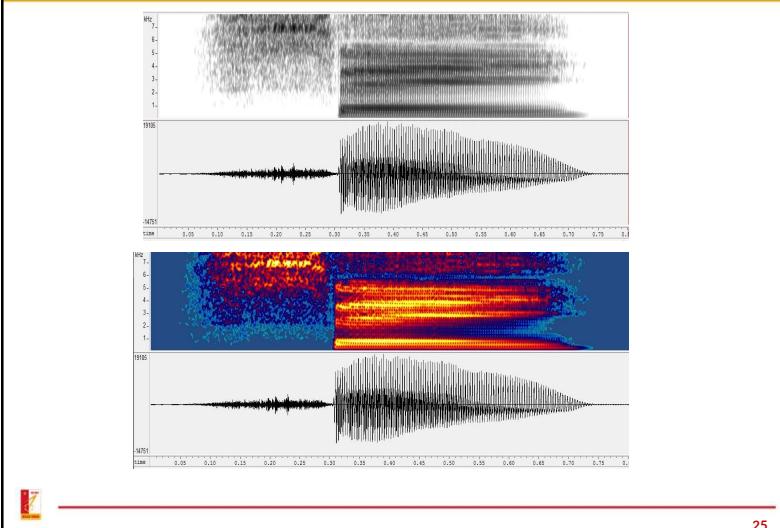


23



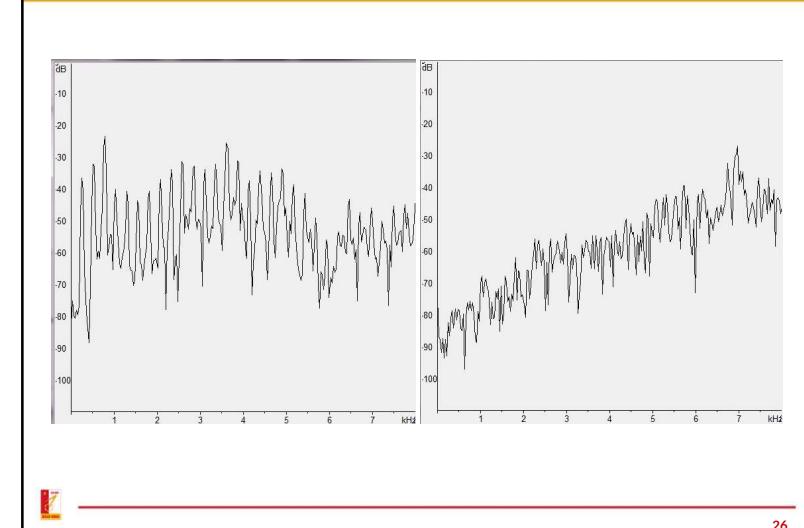
24

### Spectrogram (Sonagram)



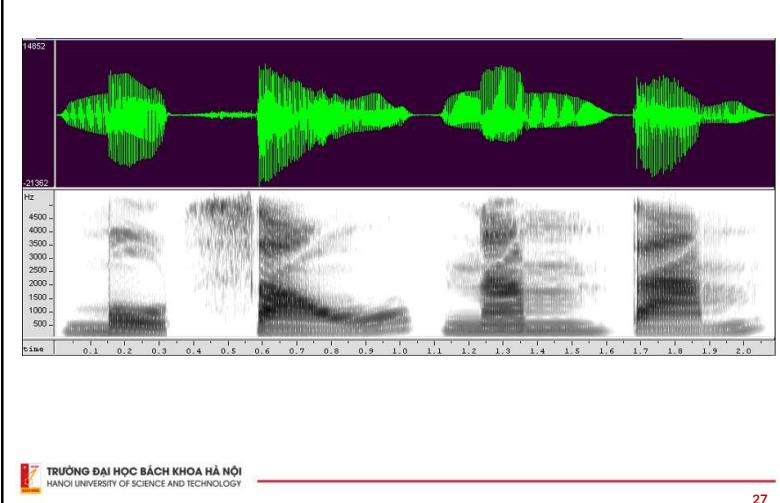
25

### Spectrogram và Phổ



26

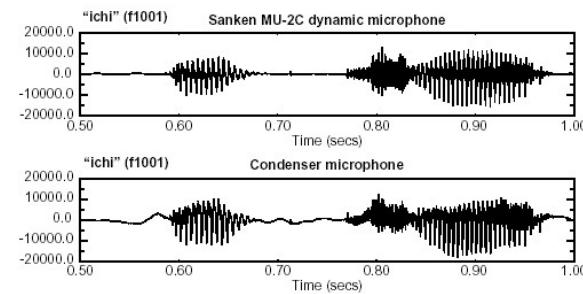
### Tiếng nói rời rạc, tiếng nói liên tục



27

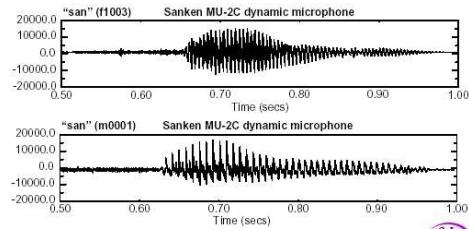
### Tính không nhất quán của tín hiệu tiếng nói

- Tín hiệu tiếng nói thu bằng micro khác loại



## Tính không nhất quán của tín hiệu tiếng nói

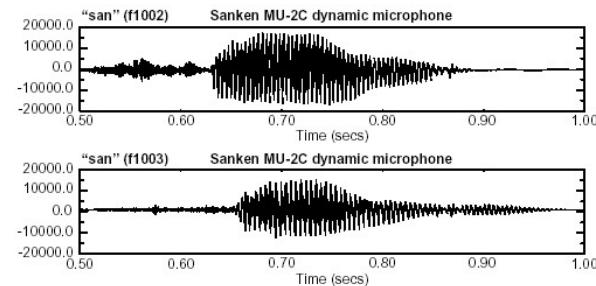
- Hai giọng khác nhau cho cùng một âm



29

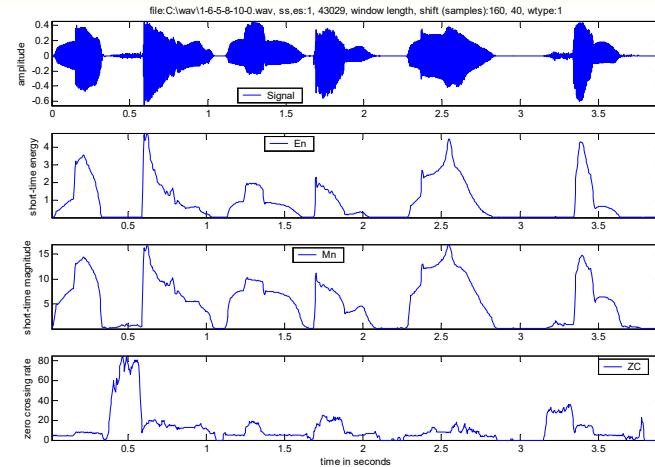
## Tính không nhất quán của tín hiệu tiếng nói

- Cùng người nói, cùng một âm



30

## Năng lượng, tỷ lệ biến thiên qua giá trị không

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

31

## Lập trình Python

```
for_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
```

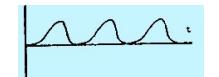
```
selection at the end -add
ob.select= 1
ler_ob.select=1
ntext.scene.objects.acti
("Selected" + str(modifi
mirror_ob.select = 0
bpy.context.selected ob
ata.objects[one.name].sel
int("please select exactly
----- OPERATOR CLASSES -----
```

```
types.Operator):
X mirror to the selected
object.mirror_mirror_x"
mirror_X"
```

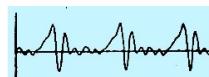
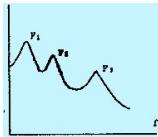
```
context):
next.active_object is not
```

32

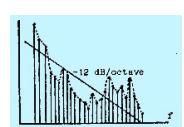
## Tạo âm hữu thanh. Formant và antiformant



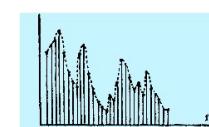
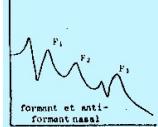
Tín hiệu nguồn hữu thanh



Tín hiệu âm hữu thanh

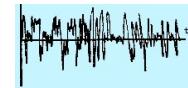


Phổ của nguồn hữu thanh

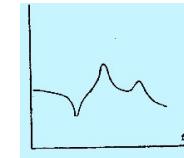


Phổ của âm hữu thanh

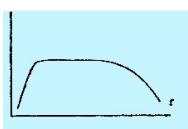
## Tạo âm vô thanh



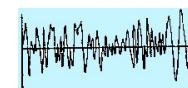
Tín hiệu nguồn vô thanh



Tín hiệu âm vô thanh



Phổ của nguồn vô thanh



## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Đơn âm tiết
- Có thanh điệu (6), biến đổi thanh điệu kèm theo biến đổi nghĩa
- Không biến đổi hình thái

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Hệ thống âm vị: 14 nguyên âm (11 nguyên âm đơn, 3 nguyên âm đôi, 22 phụ âm)

|    |     |        |
|----|-----|--------|
| 1  | i,y | ý chí  |
| 2  | ê   | ê chè  |
| 3  | e   | e dè   |
| 4  | a   | a ha   |
| 5  | ă   | mắt    |
| 6  | ơ   | bơ phờ |
| 7  | â   | ân cần |
| 8  | ư   | từ tú  |
| 9  | ô   | ô ô    |
| 10 | o   | co ro  |
| 11 | u   | lù mù  |

|   |                             |   |
|---|-----------------------------|---|
| 1 | ia,yê,ya,iê<br>(đọc ia, yê) | kia kia, yêu<br>kiều, khuya, tiên<br>tiên |
| 2 | ua,uô<br>(đọc ua)           | tua tua, luôn                             |
| 3 | ưa,uo'<br>(đọc ưa)          | lúa thưa,<br>luợt                         |

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Hệ thống âm vị: 22 phụ âm

|    |      |            |
|----|------|------------|
| 1  | b    | bồng bènh  |
| 2  | p    | óp ép      |
| 3  | v    | vần vơ     |
| 4  | ph   | phôi pha   |
| 5  | m    | mơ màng    |
| 6  | đ    | đất đai    |
| 7  | t    | tin tưởng  |
| 8  | th   | thơ thẩn   |
| 9  | d,gi | duyên, giữ |
| 10 | n    | nóng       |
| 11 | l    | long lanh  |

|    |        |             |
|----|--------|-------------|
| 12 | tr     | trồng       |
| 13 | s      | sinh viên   |
| 14 | r      | rừng        |
| 15 | ch     | chông       |
| 16 | nh     | nhọc        |
| 17 | ng,ngh | ngô nghê    |
| 18 | c,k,q  | con,kết,qua |
| 19 | kh     | khúc        |
| 20 | g,gh   | gồ ghề      |
| 21 | h      | hả hê       |
| 22 | x      | xa xôi      |

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Phân loại nguyên âm theo độ mở của miệng và chuyển động của lưỡi

| Độ mở \ Hàng | hang trước    | hang sau không tròn môi | hang sau tròn môi |
|--------------|---------------|-------------------------|-------------------|
| hở           | i ia,yê,ya,iê | ư ưa                    | u ua              |
| hở rộng      | ê             | ơ â                     | ô                 |
| rộng         | e             |                         | o                 |
|              | a ă           |                         |                   |

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Phân loại nguyên âm theo độ nâng của lưỡi và chuyển động của lưỡi

| Độ nâng \ Hàng | cao | trung bình | thấp |
|----------------|-----|------------|------|
| trước          | i e | e          |      |
| giữa           | ư   | ơ â        | a ă  |
| sau            | u ô | o          |      |

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

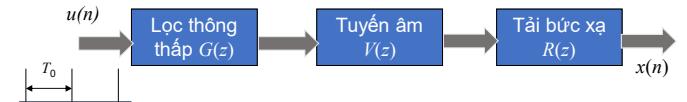
- Âm tắc: tiếng nổ, phát sinh do luồng khí từ phổi đi ra bị cản trở hoàn toàn, phải phá vỡ sự cản trở đó để thoát ra.
- Âm xát: tiếng cọ xát, phát sinh do luồng không khí đi ra bị cản trở không hoàn toàn (chỉ bị khó khăn), phải lách qua một khe hở nhỏ và trong khi thoát ra như vậy phải cọ xát vào thành của bộ máy phát âm.
- Phụ âm bên: đầu lưỡi tiếp xúc với lợi chặn lối thoát của không khí, buộc nó phải lách qua khe hở ở hai bên cạnh lưỡi tiếp giáp với má mà ra ngoài tạo nên tiếng xát nhẹ (l).
- Luồng không khí thoát ra ngoài bị cản trở, tạo nên tiếng xát hay tiếng nổ, dạng tín hiệu không tuân hoà gọi là tiếng động (ôn).
- Trong khi phát âm một số phụ âm, dây thanh cũng hoạt động đồng thời tạo nên tiếng thanh.
- Phụ âm có tỉ lệ tiếng động lớn hơn gọi là phụ âm ồn.
- Phụ âm có tỉ lệ tiếng thanh lớn hơn gọi là phụ âm vang.

## 1.4 Các đặc điểm cơ bản ngữ âm tiếng Việt

- Phân loại phụ âm theo tắc hay xát, hữu thanh hay vô thanh, mũi hóa

| Vị trí câu âm |    |               | Môi | Đầu lưỡi |           | Mặt lưỡi | Cuối lưỡi | Họng   |
|---------------|----|---------------|-----|----------|-----------|----------|-----------|--------|
| Tắc           | Ôn | Bật hơi       |     | Răng     | Vòm miệng |          |           |        |
|               |    | Không bật hơi |     | p        | t         | tr       | ch        | c,k,qu |
|               |    | Hữu thanh     | b   | đ        |           |          |           |        |
| Xát           | Ôn | Vang mũi      | m   | n        |           | nh       | ng,ngh    |        |
|               |    | Vô thanh      | ph  | x        | s         |          | kh        | h      |
|               |    | Hữu thanh     | v   | d,gi     | r         |          | g         |        |
|               |    | Vang bên      | l   |          |           |          |           |        |

## 1.5 Mô hình tạo tiếng nói



$$G(z) = \frac{A}{(1+\alpha z^{-1})(1+\beta z^{-1})}$$

$$R(z) = C(1-z^{-1})$$

$$V(z) = \frac{B}{\prod_{k=1}^K (1+b_{1k}z^{-1} + b_{2k}z^{-2})}$$

## Mô hình toàn điểm cực (AR)

$$T(z) = G(z)V(z)R(z) = \frac{\sigma}{A(z)}$$

- $A(z)$ : Hàm truyền đạt của bộ lọc đảo

$$T(z) = \frac{\sigma}{A(z)}$$

$$A(z) = 1 + \sum_{i=1}^{2K+1} a_i z^{-i} \quad A(z) = \sum_{i=0}^p a_i z^{-i} \quad a_0 = 1$$

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma u(n)$$

$$P = 2K+1$$

## Mô hình ARMA (Autoregressive Moving Average)

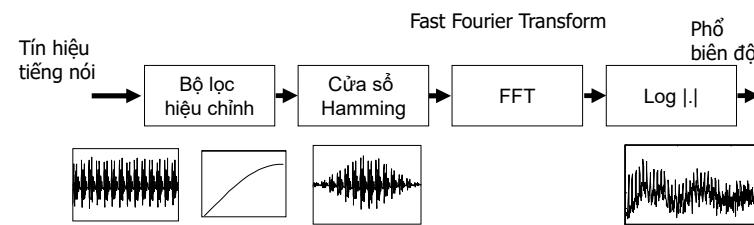
$$T(z) = \frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \sigma \frac{C(z)}{A(z)}$$

$$C(z) = \sum_{i=0}^q c_i z^{-i} \quad c_0 = 1$$

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma \sum_{i=0}^q c_i u(n-i)$$

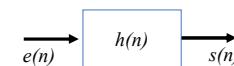
## 1.6 Các kỹ thuật cơ bản xử lý tín hiệu tiếng nói

- Phân tích phổ



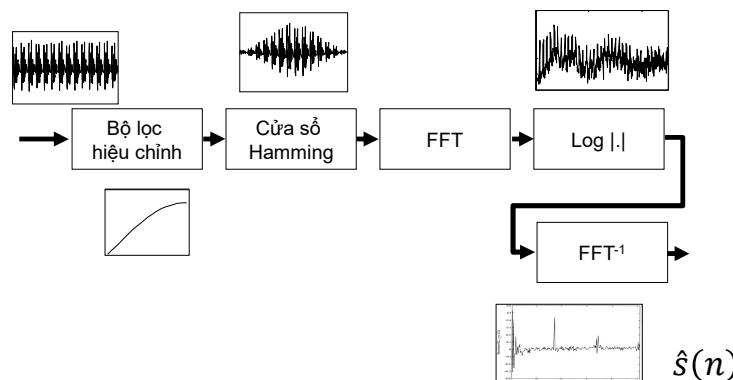
- Bộ lọc hiệu chỉnh  $H(z) = 1 - az^{-1}$ ,  $a = 0,95..0,98$

## Xử lý đồng hình (homomorphic)

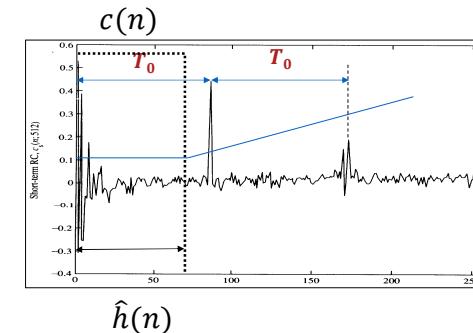


- $s(n) = h(n) * e(n) \rightarrow S(\omega) = H(\omega)E(\omega)$
- $\log S(\omega) = \log H(\omega) + \log E(\omega)$
- $\mathbb{F}^{-1}\{\log S(\omega)\} = \mathbb{F}^{-1}\{\log H(\omega)\} + \mathbb{F}^{-1}\{\log E(\omega)\}$
- $\hat{s}(n) = \hat{h}(n) + \hat{e}(n)$

## Sơ đồ khối xử lý đồng hình



## Ví dụ



### Tiên đoán tuyến tính (Linear Prediction Coding)

- Mô hình AR

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma u(n)$$

Tiên đoán

$$\hat{x}(n) = - \sum_{i=1}^p \hat{a}_i x(n-i)$$

Sai số tiên đoán

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

Sai số bình phương toàn phần

$$E = \sum_n e^2(n)$$

Tối thiểu hóa sai số

$$\frac{\partial E}{\partial \hat{a}_i} = 0, i = 1, 2, \dots, p$$

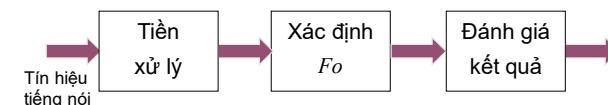
### Một số phương pháp xác định $F_0$

- Dựa vào hàm tự tương quan
- Dựa vào hàm vi sai biên độ trung bình
- Xử lý đồng hình

### Xác định tần số cơ bản

- Giá trị  $F_0$  phụ thuộc vào giới tính và lứa tuổi

- Giọng nam: 80..250 Hz
- Giọng nữ: 150..500 Hz



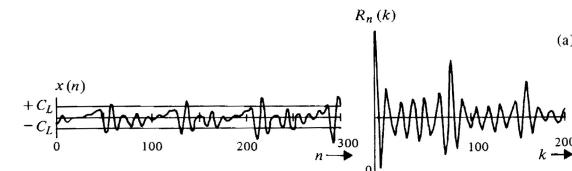
### Dựa vào hàm tự tương quan

- Tính hàm tự tương quan  $R(k)$  của tín hiệu tiếng nói  $x(n)$

$$R(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) \quad k = 0, 1, \dots, K$$

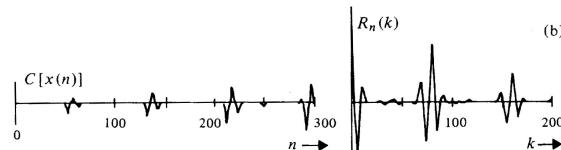
$$Fs = 10 \text{ kHz}, N = 300, K = 150.$$

Tìm cực đại trong khoảng  $(0, K)$

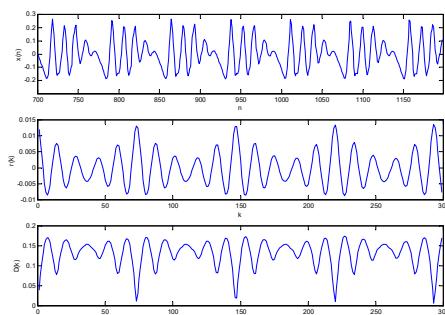


### Phương pháp tự tương quan có cải tiến

- Hạn chế, loại bỏ  $|x| < C_L$



### Ví dụ



### Dựa vào hàm vi sai biên độ trung bình

- AMDF- Average Magnitude Difference Function)

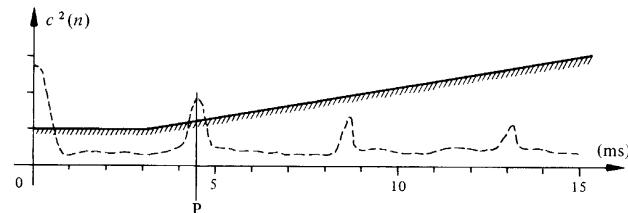
$$D(k) = \sum_{m=0}^{N-1} |x(n+m) - x(n+m-k)| \quad k = 0, 1, \dots, K$$

$$D(iP) = 0, \quad i = 0, 1, \dots \quad \frac{1}{N} \sum_{n=0}^{N-1} |u(n)| \leq \left[ \frac{1}{N} \sum_{n=0}^{N-1} u^2(n) \right]^{1/2}$$

$$\begin{aligned} D(k) &= \lambda \left\{ \frac{1}{N} \sum_{m=0}^{N-1} [x(n+m) - x(n+m-k)]^2 \right\}^{1/2} \\ &= \lambda \left\{ \frac{1}{N} [2r(0) - 2r(k)] \right\}^{1/2} \quad k = 0, 1, \dots, K \end{aligned}$$

với  $\lambda < 1$

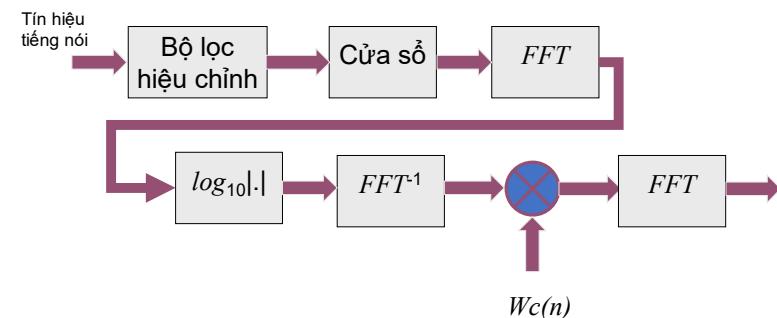
### Xử lý đồng hình



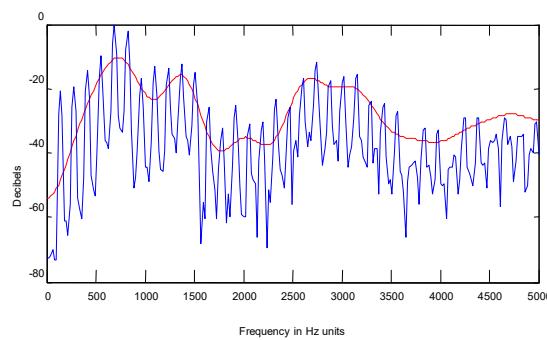
## Xác định formant

- Tham số cần xác định
  - Formant  $F_k$
  - Dải thông  $B_k$
- Phương pháp
  - Xử lý đồng hình
  - LPC

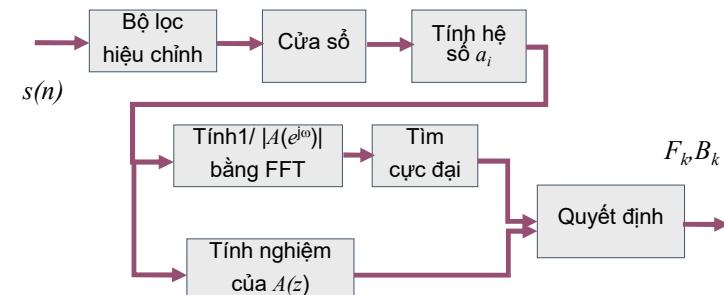
## Xử lý đồng hình



## Xử lý đồng hình



## Phương pháp LPC





## Lập trình Python

ONE LOVE. ONE FUTURE.