

**TRƯỜNG ĐẠI HỌC HÀNG HẢI VIỆT NAM**  
**KHOA CÔNG NGHỆ THÔNG TIN**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**

-----\*\*\*-----



# **BÀI GIẢNG**

## **XỬ LÝ TIẾNG NÓI**

<b>TÊN HỌC PHẦN</b>	<b>: XỬ LÝ TIẾNG NÓI</b>
<b>MÃ HỌC PHẦN</b>	<b>: 17412</b>
<b>TRÌNH ĐỘ ĐÀO TẠO</b>	<b>: ĐẠI HỌC CHÍNH QUY</b>
<b>DÙNG CHO SV NGÀNH</b>	<b>: CÔNG NGHỆ THÔNG TIN</b>

## MỤC LỤC

<b>CHƯƠNG I: CÁC KIẾN THỨC CƠ BẢN .....</b>	<b>6</b>
1.1 Tổng quan về xử lý tiếng nói.....	6
1.1.1 Nhận dạng tiếng nói tự động .....	6
1.1.2 Chuyển đổi văn bản thành tiếng nói.....	7
1.1.3 Hệ thống hiểu ngôn ngữ nói .....	7
1.2 Cấu trúc ngôn ngữ nói .....	9
1.2.1 Hệ thống tiếng nói con người.....	9
1.2.2 Ngữ âm học và âm vị học.....	10
1.2.3 Âm tiết và từ ngữ.....	11
<b>CHƯƠNG II: XỬ LÝ TÍN HIỆU SỐ TRONG XỬ LÝ TIẾNG NÓI .....</b>	<b>13</b>
2.1 Xử lý tín hiệu số .....	13
2.1.1 Phép biến đổi Fourier .....	14
2.1.2 Phép biến đổi Fourier rời rạc.....	14
2.1.3 Các bộ lọc số và cửa sổ .....	15
2.2 Biểu diễn tín hiệu tiếng nói .....	15
2.2.1 Phân tích Fourier thời gian ngắn .....	15
2.2.2 Mô hình âm học của việc tạo tiếng nói .....	15
2.3 Mã hóa tiếng nói.....	19
2.3.1 Các tính chất của bộ mã hóa tiếng nói .....	19
2.3.2 Các bộ mã hóa dạng sóng tiếng nói vô hướng .....	20
<b>CHƯƠNG III: NHÂN DẠNG TIẾNG NÓI.....</b>	<b>22</b>
3.1 Các hệ thống nhận dạng tiếng nói .....	22
3.1.1 Nhận dạng từ riêng lẻ .....	22
3.1.2 Nhận dạng từ liên tục .....	24
3.2 Các mô hình Markov ẩn .....	27
3.2.1 Chuỗi Markov .....	27
3.2.2 Mô hình Markov.....	28
<b>CHƯƠNG IV: CÁC HỆ THỐNG CHUYỂN VĂN BẢN THÀNH GIỌNG NÓI .....</b>	<b>30</b>
4.1 Phân tích ngữ âm và văn bản .....	30
4.1.1 Từ vựng .....	30
4.1.2 Xác định cấu trúc tài liệu.....	30
4.1.3 Chuẩn hóa văn bản .....	31
4.1.4 Phân tích ngôn ngữ.....	32

4.1.5 Chuyển đổi ký tự sang âm thanh .....	32
4.2 Tổng hợp tiếng nói .....	33
4.2.1 Các tính chất của tổng hợp tiếng nói .....	33
4.2.2 Tổng hợp tiếng nói bằng các Formant.....	34
4.2.3 Tổng hợp tiếng nói bằng ghép nối.....	34
4.2.4 Đánh giá các hệ thống tổng hợp tiếng nói.....	36

**Tên học phần:** Xử lý tiếng nói  
**Bộ môn phụ trách giảng dạy:** Hệ thống Thông tin  
**Mã học phần:** 17412

**Loại học phần:** 2  
**Khoa phụ trách:** CNTT.  
**Tổng số TC:** 4

Tổng số tiết	Lý thuyết	Thực hành/Xemina	Tự học	Bài tập lớn	Đồ án môn học
75	45	30	0	không	không

**Điều kiện tiên quyết:**

Không yêu cầu.

**Mục tiêu của học phần:**

Cung cấp các kiến thức cơ bản về lĩnh vực xử lý tiếng nói, hiểu các hệ thống chuyển văn bản thành tiếng nói, các hệ thống nhận dạng tiếng nói.

**Nội dung chủ yếu:**

Các vấn đề liên quan đến tiếng nói và ngữ âm học; Các hệ thống chuyển văn bản thành tiếng nói; Cơ sở xử lý tín hiệu số trong xử lý tiếng nói; Nhận dạng tiếng nói.

**Nội dung chi tiết:**

TÊN CHƯƠNG MỤC	PHÂN PHỐI SỐ TIẾT				
	TS	LT	TH	BT	KT
<b>CHƯƠNG I: CÁC KIẾN THỨC CƠ BẢN</b>	<b>15</b>	<b>9</b>	<b>6</b>		
<b>1.1 Tổng quan về xử lý tiếng nói</b>		<b>3</b>			
1.1.1 Nhận dạng tiếng nói tự động					
1.1.2 Chuyển đổi văn bản thành tiếng nói					
1.1.3 Hệ thống hiểu ngôn ngữ nói					
<b>1.2 Cấu trúc ngôn ngữ nói</b>		<b>6</b>			
1.2.1 Hệ thống tiếng nói con người					
1.2.2 Ngữ âm học và âm vị học					
1.2.3 Âm tiết và từ ngữ					
<b>CHƯƠNG II: XỬ LÝ TÍN HIỆU SỐ TRONG XỬ LÝ TIẾNG NÓI</b>	<b>15</b>	<b>9</b>	<b>6</b>		
<b>2.1 Xử lý tín hiệu số</b>		<b>3</b>			
2.1.1 Phép biến đổi Fourier					
2.1.2 Phép biến đổi Fourier rời rạc					
2.1.3 Các bộ lọc số và cửa sổ					
<b>2.2 Biểu diễn tín hiệu tiếng nói</b>		<b>3</b>			
2.2.1 Mô hình âm học của việc tạo tiếng nói					
<b>2.3 Mã hóa tiếng nói</b>		<b>3</b>			
2.3.1 Các tính chất của bộ mã hóa tiếng nói					
2.3.2 Các bộ mã hóa dạng sóng tiếng nói vô hướng					
<b>CHƯƠNG III: NHẬN DẠNG TIẾNG NÓI</b>	<b>21</b>	<b>12</b>	<b>9</b>		
<b>3.1 Các hệ thống nhận dạng tiếng nói</b>		<b>3</b>			
3.1.1 Nhận dạng từ riêng lẻ					
3.1.2 Nhận dạng từ liên tục					
<b>3.2 Các mô hình Markov ẩn</b>		<b>9</b>			
3.2.1 Chuỗi Markov					
3.2.2 Mô hình Markov					
<b>CHƯƠNG IV: CÁC HỆ THỐNG CHUYỂN VĂN BẢN THÀNH GIỌNG NÓI</b>	<b>24</b>	<b>15</b>	<b>9</b>		
<b>4.1 Phân tích ngữ âm và văn bản</b>		<b>6</b>			
4.1.1 Từ vựng					
4.1.2 Xác định cấu trúc tài liệu					
4.1.3 Chuẩn hóa văn bản					
4.1.4 Phân tích ngôn ngữ					

4.1.5 Chuyển đổi ký tự sang âm thanh					
<b>4.2 Tổng hợp tiếng nói</b>		<b>9</b>			
4.2.1 Các tính chất của tổng hợp tiếng nói					
4.2.2 Tổng hợp tiếng nói bằng các Formant					
4.2.3 Tổng hợp tiếng nói bằng ghép nối					
4.2.4 Đánh giá các hệ thống tổng hợp tiếng nói					

### Nhiệm vụ của sinh viên:

Tham dự các buổi học lý thuyết và thực hành, làm các bài tập được giao, làm các bài thi giữa học phần và bài thi kết thúc học phần theo đúng quy định.

### Tài liệu học tập:

1. Xuedong Huang, Alex Acero, Hsiao Wuen Hon, *Spoken Language Processing- A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.
2. Lawrence R. Rabiner, Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1978.

### Hình thức và tiêu chuẩn đánh giá sinh viên:

- Hình thức thi: thi viết.
- Tiêu chuẩn đánh giá sinh viên: căn cứ vào sự tham gia học tập của sinh viên trong các buổi học lý thuyết và thực hành, kết quả làm các bài tập được giao, kết quả của các bài thi giữa học phần và bài thi kết thúc học phần.

**Thang điểm:** Thang điểm chữ A, B, C, D, F.

**Điểm đánh giá học phần:**  $Z = 0,3X + 0,7Y$ .

Bài giảng này là tài liệu **chính thức và thống nhất** của Bộ môn Hệ thống Thông tin, Khoa Công nghệ Thông tin và được dùng để giảng dạy cho sinh viên.

**Ngày phê duyệt:**     /     /

**Trưởng Bộ môn**

## CHƯƠNG I : CÁC KIẾN THỨC CƠ BẢN

### 1.1 Tổng quan về xử lý tiếng nói

Kể từ khi xuất hiện, máy tính càng ngày càng chứng tỏ rằng đó là một công cụ vô cùng hữu ích trợ giúp con người xử lý thông tin. Cùng với sự phát triển của xã hội, khối lượng thông tin mà máy tính cần xử lý tăng rất nhanh trong khi thời gian dành cho những công việc này lại giảm đi. Vì vậy, việc tăng tốc độ xử lý thông tin, trong đó có tốc độ trao đổi thông tin giữa con người và máy tính, trở thành một yêu cầu cấp thiết. Hiện tại, giao tiếp người-máy được thực hiện bằng các thiết bị như bàn phím, chuột, màn hình,... với tốc độ tương đối chậm nên cần có các phương pháp trao đổi thông tin mới giúp con người làm việc hiệu quả hơn với máy tính. Một trong những hướng nghiên cứu này là sử dụng tiếng nói trong trao đổi thông tin người-máy. Những nghiên cứu này liên quan trực tiếp tới các kết quả của chuyên ngành xử lý tiếng nói, trong đó có tổng hợp tiếng nói.

#### 1.1.1 Nhận dạng tiếng nói tự động

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với mục đích là phân lớp (classify) thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ trong bộ nhớ. Các mẫu là các đơn vị nhận dạng, chúng có thể là các từ, hoặc các âm vị. Nếu các mẫu này là bất biến và không thay đổi thì công việc nhận dạng tiếng nói trở nên đơn giản bằng cách so sánh dữ liệu tiếng nói cần nhận dạng với các mẫu đã được học và lưu trữ trong bộ nhớ. Khó khăn cơ bản của nhận dạng tiếng nói đó là tiếng nói luôn biến thiên theo thời gian và có sự khác biệt lớn giữa tiếng nói của những người nói khác nhau, tốc độ nói, ngữ cảnh và môi trường âm học khác nhau.

Xác định những thông tin biến thiên nào của tiếng nói là có ích và những thông tin nào là không có ích đối với nhận dạng tiếng nói là rất quan trọng. Đây là một nhiệm vụ rất khó khăn mà ngay cả với các kỹ thuật xác suất thống kê mạnh cũng khó khăn trong việc tổng quát hoá từ các mẫu tiếng nói những biến thiên quan trọng cần thiết trong nhận dạng tiếng nói.

Các nghiên cứu về nhận dạng tiếng nói dựa trên ba nguyên tắc cơ bản:

- Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn (short-term amplitude spectrum). Nhờ vậy ta có thể trích ra các đặc điểm tiếng nói từ những khoảng thời gian ngắn và dùng các đặc điểm này làm dữ liệu để nhận dạng tiếng nói.
- Nội dung của tiếng nói được biểu diễn dưới dạng chữ viết, là một dãy các ký hiệu ngữ âm. Do đó ý nghĩa của một phát âm được bảo toàn khi chúng ta phiên âm phát âm thành dãy các ký hiệu ngữ âm.

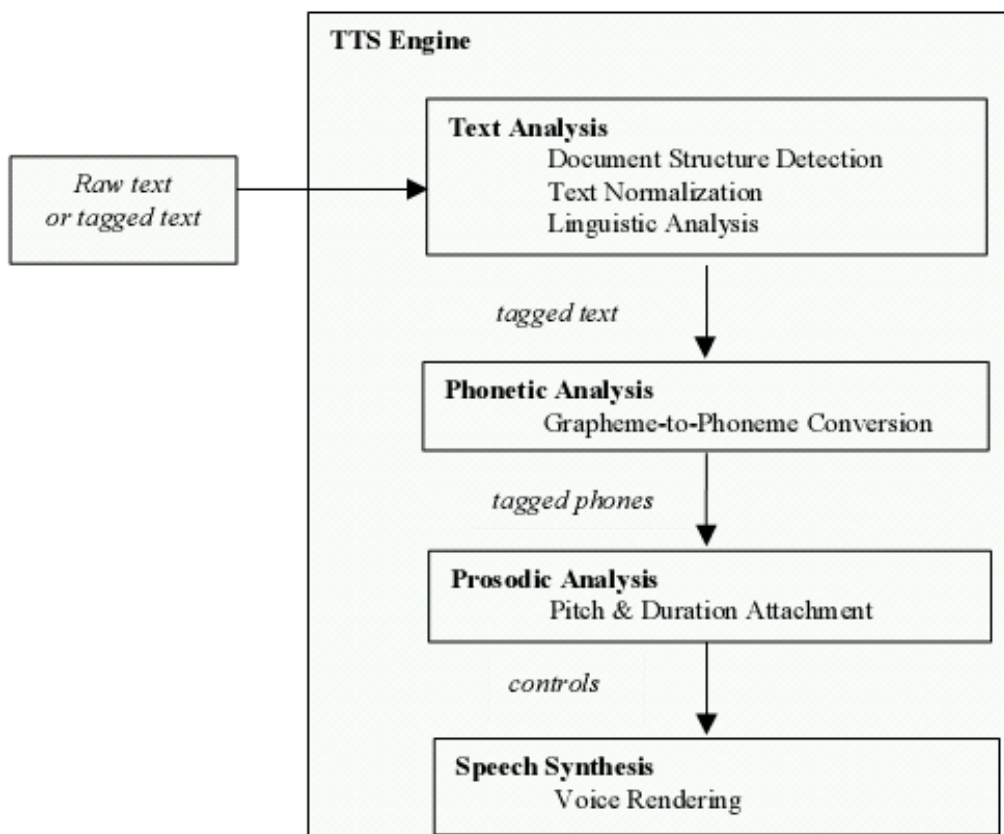
- Nhận dạng tiếng nói là một quá trình nhận thức. Thông tin về ngữ nghĩa (semantics) và suy đoán (pragmatics) có giá trị trong quá trình nhận dạng tiếng nói, nhất là khi thông tin về âm học là không rõ ràng.

### 1.1.2 Chuyển đổi văn bản thành tiếng nói

Các hệ thống chuyển đổi văn bản thành giọng nói có thể được xem như là hệ thống mã hóa tiếng nói cho phép lựa chọn kiểu cách nói, tốc độ, cường độ và các hiệu ứng. Hệ thống chuyển văn bản thành tiếng nói (Text-to-Speech) là một hệ thống có thể sinh ra tiếng nói gần giống với con người từ các văn bản được đưa vào (còn được gọi là hệ thống tổng hợp tiếng nói) Sự chuyển đổi các từ dưới dạng viết sang tiếng nói là một công việc khó khăn vì hệ thống TTS cần dữ liệu từ vựng rất lớn và nhiều ngữ điệu của âm thanh.

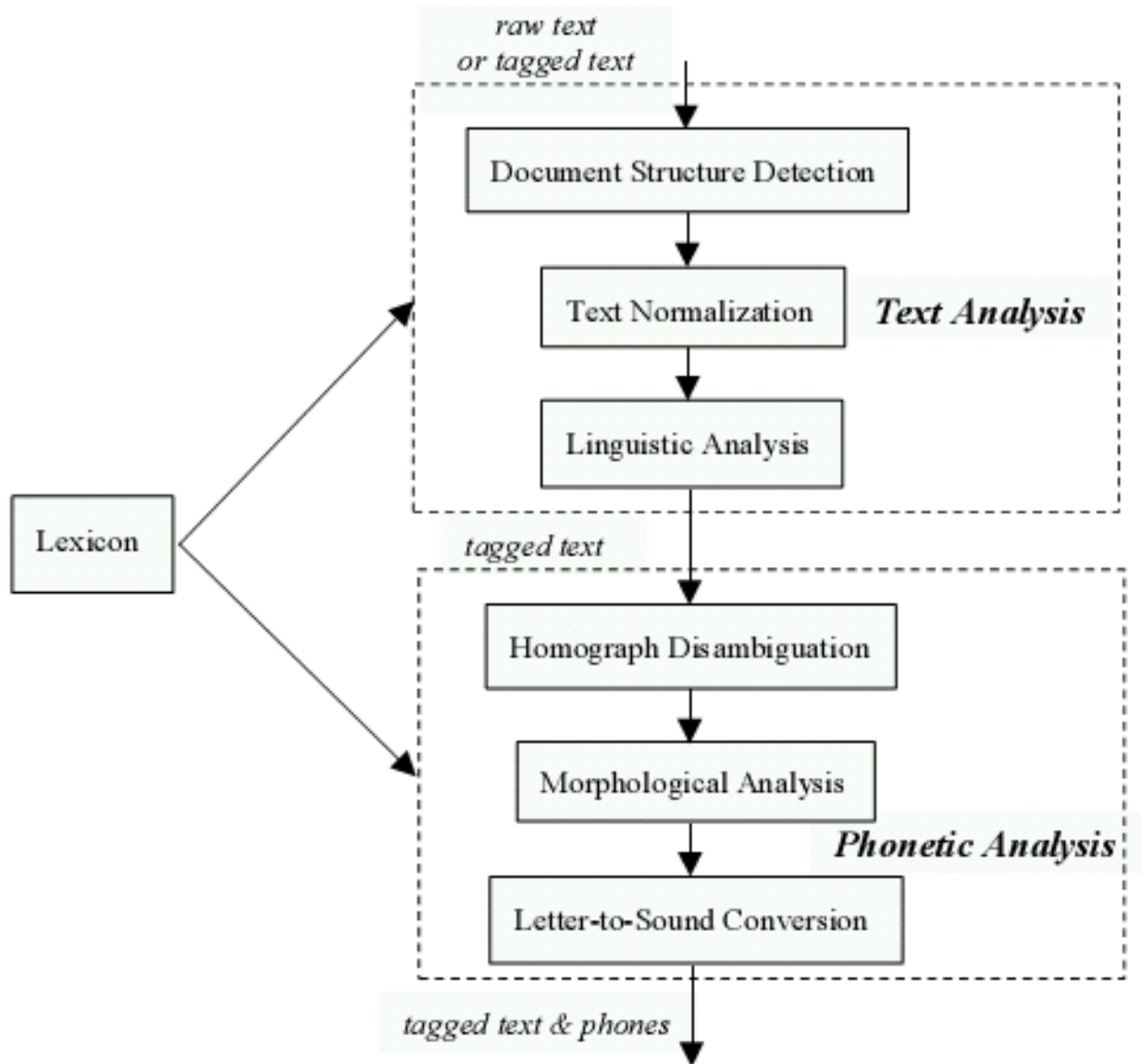
#### Các thành phần cơ bản của một hệ thống chuyển đổi văn bản thành tiếng nói

- Bộ phân tích văn bản: chuẩn hóa văn bản sang dạng thích hợp cho hệ thống TTS
- Bộ phân tích ngữ âm chuyển đổi văn bản đã được xử lý thành dãy các âm tương ứng sau đó được phân tích ngữ điệu để xác định trọng âm, ngắt nhịp, thời gian, ..
- Cuối cùng, bộ tổng hợp tiếng nói nhận các tham số đầu vào từ dãy âm vị đã xử lý đầy đủ



- Thành phần phân tích văn bản:
  - + Xác định cấu trúc tài liệu, chuyển đổi ký hiệu, phân tích cấu trúc ngôn ngữ
  - + Chuyển đổi các ký hiệu sang dạng chuẩn.
  - + Chuyển đổi các số sang dạng chữ tương ứng

- + Phân tích khoảng trống, dấu chấm câu để xác định cấu trúc ngôn ngữ
- Thành phần phân tích ngữ âm:
  - + Chuyển đổi các từ đã chuẩn hóa sang các âm vị tương ứng (với thông tin như trọng âm, thời gian phát âm)



### 1.1.3 Hệ thống hiểu ngôn ngữ nói

Tổng hợp tiếng nói là lĩnh vực đang được nghiên cứu khá rộng rãi trên thế giới và đã cho những kết quả khá tốt. Có ba phương pháp cơ bản dùng để tổng hợp tiếng nói là mô phỏng bộ máy phát âm, tổng hợp bằng formant và tổng hợp bằng cách ghép nối. Phương pháp mô phỏng bộ máy phát âm cho chất lượng tốt nhưng đòi hỏi nhiều tính toán vì việc mô phỏng chính xác bộ máy phát âm rất phức tạp. Phương pháp tổng hợp formant không đòi hỏi chi phí cao trong tính toán nhưng cho kết quả chưa tốt. Phương pháp tổng hợp ghép nối cho chất lượng tốt, chi phí tính toán không cao nhưng số lượng từ vựng phải rất lớn.

Ở các nước phát triển, những nghiên cứu xử lý tiếng nói, đã cho các kết quả khả quan, làm tiền đề cho việc giao tiếp người-máy bằng tiếng nói. Ở Việt Nam, các nghiên cứu trong lĩnh



vực này tuy mới được phát triển trong những năm gần đây nhưng cũng đã có một số kết quả khả quan

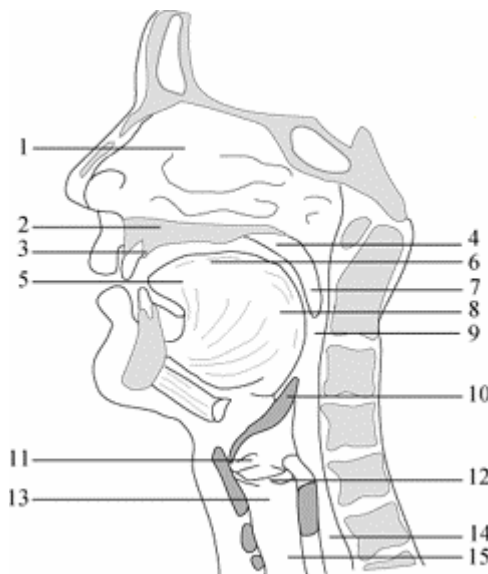
## 1.2 Cấu trúc ngôn ngữ nói

### 1.2.1 Hệ thống tiếng nói con người

#### a) Bộ máy phát âm

Bộ máy phát âm bao gồm các thành phần riêng rẽ như phổi, khí quản, thanh quản, và các đường dẫn miệng, mũi. Trong đó:

- Thanh quản chứa hai dây thanh có thể dao động tạo ra sự cộng hưởng cần thiết để tạo ra âm thanh.
- Tuyến âm là ống không đều bắt đầu từ môi, kết thúc bởi dây thanh hoặc thanh quản.
- Khoang mũi là ống không đều bắt đầu từ môi, kết thúc bởi vòm miệng, có độ dài cố định khoảng 12cm đối với người lớn.
- Vòm miệng là các nếp cơ chuyển động.



1. Hốc mũi
2. Vòm miệng trên
3. Ổ răng
4. Vòm miệng mềm
5. Đầu lưỡi
6. Thân lưỡi
7. Lưỡi gà
8. Cơ miệng
9. Yết hầu
10. Nắp đóng của thanh quản
11. Dây thanh giả
12. Dây thanh
13. Thanh quản
14. Thực quản
15. Khí quản

#### b) Cơ chế phát âm

Trong quá trình tạo âm thanh không phải là âm mũi, vòm miệng mở, khoang mũi đóng lại, dòng khí sẽ chỉ đi qua khoang mũi. Khi phát âm mũi, vòm miệng hạ thấp và dòng khí sẽ chỉ đi qua khoang mũi.



Tuyến âm sẽ được kích thích bởi nguồn năng lượng chính tại thanh môn. Tiếng nói được tạo ra

do tín hiệu nguồn từ thanh môn phát ra, đẩy không khí có trong phổi lên tạo thành dòng khí, va chạm vào hai dây thanh trong tuyến âm. Hai dây thanh dao động sẽ tạo ra cộng hưởng, dao động âm sẽ được lan truyền theo tuyến âm (tính từ tuyến âm đến khoang miệng) và sau khi đi qua khoang mũi và môi, sẽ tạo ra tiếng nói.

### 1.2.2 Ngữ âm học và âm vị học

Ngành nghiên cứu âm thanh cho một ngôn ngữ được gọi là âm vị học. Ngữ âm học là một ngành khoa học nghiên cứu các đặc điểm âm thanh của tiếng nói con người. Ngữ âm học nghiên cứu các phổ quát âm thanh. Ví dụ: Nhờ vào bộ máy cấu âm, con người có thể phát ra các chuỗi âm thanh khác nhau. Ngữ âm học chia các loại âm thanh này thành các phạm trù ngữ âm khác nhau: nguyên âm, phụ âm, tắc, xát... Còn âm vị học thì không nghiên cứu rộng như vậy. Âm vị học nghiên cứu xem trong một ngôn ngữ có bao nhiêu đơn vị âm thanh là có chức năng khu biệt nghĩa. Hoặc, trong ngôn ngữ, những nét ngữ âm nào trở thành những nét khu biệt và có ý nghĩa. Chính vì vậy, ngữ âm học có số đơn vị là vô hạn, quen gọi là các âm tố (sounds). Còn âm vị học, có số đơn vị hữu hạn, đếm được. Đơn vị của âm vị học là âm vị (phonemes).

Ví dụ:

<b>[p]</b> =	<b>[b]</b> =	<b>[m]</b> =
+ PAT	+ PAT	+ PAT
+ Môi	+ Môi	+ Môi
+ Tắc	+ Tắc	+ Tắc
+ Vô thanh	- Vô thanh	- Vô thanh
- Mũi	- Mũi	+ Mũi
+ Mạnh (cường độ)	- Mạnh (cường độ)	- Mạnh (cường độ)
		+ Dài
		
nét ngữ âm	nét ngữ âm	nét ngữ âm

PAT: Phụ âm tính (Consonantal)  
(+): Present  
(-): Absent

Về mặt ngữ âm học, 3 nguyên âm này đều có nội dung ngữ âm là như nhau ở tất cả các ngôn ngữ trên thế giới. Ví dụ như [m] phân biệt với [p] và [b] ở đặc tính [mũi/không mũi]. [p] phân biệt với [m] và [b] ở đặc tính [+ vô thanh]: +vô thanh/ +hữu thanh. Những đối lập kiểu như vậy thì ở bất cứ ngôn ngữ nào cũng giống nhau. Vì vậy, đó chỉ là các thuộc tính ngữ âm học thuần túy. Tuy nhiên, dưới con mắt âm vị học, tài nguyên ngữ âm của các âm vị phải được lựa chọn dưới con mắt của người bản ngữ (native), được tận dụng và chọn lựa, được khai thác sao cho có lợi và hợp với hệ thống (cái tạng của ngôn ngữ) của mình nhất. Nói tóm lại, các nét ngữ âm đã biến thành các nét âm vị học; từ cái chung, cái phổ quát trở thành cái riêng, cái đặc thù. Cả một tiến trình lịch sử phát triển của một hệ thống ngữ âm, từ lúc xa xưa cho đến ngày nay, suy cho cùng, là sự chọn lựa và khai thác tài nguyên nhân loại ấy cho tộc người mình, cho cộng đồng nói năng cụ thể. Quá trình chọn lựa đó cũng chặt vật, và có thể nói là “đầy máu và nước mắt”. Chính vì vậy, các nhà âm vị học hiện đại không quay lưng lại với lịch sử của một ngôn ngữ mà tìm ở đó ra những hệ thống cứ liệu chắc chắn cho việc chứng minh những chức năng của hệ âm thanh một ngôn ngữ. Phương pháp luận này khác hoàn toàn với âm vị học cấu trúc luận xưa kia. Vì vậy, có thể nói, âm vị học hiện đại là hình ảnh thu

nhỏ một cách logic và có tính hình thức hoá cao con đường phát triển của một hệ thống âm thanh một ngôn ngữ.

### 1.2.3 Âm tiết và từ ngữ

#### a) Âm tiết

Chuỗi lời nói mà con người phát ra gồm nhiều khúc đoạn dài ngắn khác nhau. Đơn vị phát âm ngắn nhất là âm tiết (syllable).

Về phương diện phát âm, âm tiết có tính chất toàn vẹn, không thể phân chia được là bởi nó được phát âm bằng một đợt căng của cơ thịt của bộ máy phát âm.

Khi phát âm một âm tiết, các cơ thịt của bộ máy phát âm đều phải trải qua ba giai đoạn: tăng cường độ căng, đỉnh điểm căng thẳng và giảm độ căng.

Dựa vào cách kết thúc, các âm tiết được chia thành hai loại lớn: mở và khép. Trong mỗi loại lại có hai loại nhỏ hơn. Như vậy có 4 loại âm tiết như sau:

- Những âm tiết được kết thúc bằng một phụ âm vang (/m, n, ɲ/...) được gọi là những âm tiết nửa khép.
- Những âm tiết được kết thúc bằng một phụ âm không vang (/p, t, k/) được gọi là những âm tiết khép.
- Những âm tiết được kết thúc bằng một bán nguyên âm (/w, j/) được gọi là những âm tiết nửa mở.
- Những âm tiết được kết thúc bằng cách giữ nguyên âm sắc của nguyên âm ở đỉnh âm tiết thì được gọi là âm tiết mở.

#### b) Đặc điểm của âm tiết tiếng Việt

- Có tính độc lập cao:
  - + Trong dòng lời nói, âm tiết tiếng Việt bao giờ cũng thể hiện khá đầy đủ, rõ ràng, được tách và ngắt ra thành từng khúc đoạn riêng biệt.
  - + Khác với âm tiết các ngôn ngữ châu Âu, âm tiết nào của tiếng Việt cũng mang một thanh điệu nhất định.
  - + Do được thể hiện rõ ràng như vậy nên việc vạch ranh giới âm tiết tiếng Việt trở nên rất dễ dàng.
- Có khả năng biểu hiện ý nghĩa
  - + Ở tiếng Việt, tuyệt đại đa số các âm tiết đều có ý nghĩa. Hay, ở tiếng Việt, gần như toàn bộ các âm tiết đều hoạt động như từ...

+ Có thể nói, trong tiếng Việt, âm tiết không chỉ là một đơn vị ngữ âm đơn thuần mà còn là một đơn vị từ vựng và ngữ pháp chủ yếu. Ở đây, mối quan hệ giữa âm và nghĩa trong âm tiết cũng chặt chẽ và thường xuyên như trong từ của các ngôn ngữ Âu châu, và đó chính là một nét đặc trưng loại hình chủ đạo của tiếng Việt.

- Có một cấu trúc chặt chẽ

Mô hình âm tiết tiếng Việt không phải là một khối không thể chia cắt mà là một cấu trúc. Cấu trúc âm tiết tiếng Việt là một cấu trúc hai bậc, ở dạng đầy đủ nhất gồm 5 thành tố, mỗi thành tố có một chức năng riêng.

### **CÂU HỎI ÔN TẬP**

1. Trình bày khái niệm về xử lý tiếng nói? Ý nghĩa trong thực tiễn? Cho ví dụ minh họa?
2. Trình bày các nguyên tắc cơ bản trong quá trình nhận dạng tiếng nói?
3. Trình bày hệ thống chuyển đổi văn bản thành giọng nói?
4. Trình bày cấu trúc của ngôn ngữ nói?

## CHƯƠNG II : XỬ LÝ TÍN HIỆU SỐ TRONG XỬ LÝ TIẾNG NÓI

### 2.1 Xử lý tín hiệu số

Phân tích và thiết kế các hệ thống tuyến tính được thực hiện dễ dàng nhờ *các biểu diễn miền tần số* frequency-domain representation) của cả các tín hiệu và hệ thống. Do vậy, cần xét các biểu diễn của *biến đổi Fourier* (Fourier Transform, FT) và của *biến đổi Z* (Z - Transform, ZT) của các tín hiệu và hệ thống rời rạc.

*Biến đổi Z* (ZT) : Biểu diễn ZT của dãy được xác định bởi 2 phương trình:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n).z^{-n} \quad (2a)$$

$$x(n) = \frac{1}{2\pi j} \oint_C X(\tau) z^{-n-1} d\tau \quad (2b)$$

"*Biến đổi Z*" (ZT) hay "biến đổi trực tiếp" của  $x(n)$  được xác định bởi (2a). Tổng quan, có thể thấy  $X(z)$  là chuỗi lũy thừa vô hạn theo biến  $z^{-1}$ , trong đó dãy các giá trị,  $x(n)$ , đóng vai trò các hệ số trong chuỗi lũy thừa. Nói chung, các chuỗi lũy thừa này sẽ hội tụ đến giá trị hữu hạn chỉ với các giá trị xác định của  $z$ . *Điều kiện đủ* của hội tụ là:

$$\sum_{n=-\infty}^{\infty} |x(n)| |z^{-n}| < \infty \quad (3)$$

Tập hợp các giá trị mà chuỗi hội tụ xác định một miền trên mặt phẳng phức  $Z$  gọi là *miền hội tụ*. Nói chung, miền này có dạng:

$$R_1 < |z| < R_2 \quad (4)$$

Có nhiều định lý và tính chất của biểu diễn ZT tiện dụng cho việc nghiên cứu các hệ thống thời gian rời rạc. Danh sách các định lý quan trọng cho trong bảng 1. Về hình thức, các định lý này giống với các định lý tương ứng của biến đổi Laplace cho các hàm thời gian liên tục. Tuy nhiên, điều này không có nghĩa là ZT là một dạng xấp xỉ nào đó của biến đổi Laplace. biến đổi Laplace là biểu diễn chính xác của các hàm thời gian liên tục, còn ZT là biểu diễn chính xác của dãy các số

	Dãy	ZT
1. Tuyến tính (Linear)	$ax_1(n) + bx_2(n)$	$aX_1(Z) + bX_2(Z)$
2. Dịch chuyển (Shift)	$x(n + n_0)$	$Z^{n_0} X(Z)$
3. Trọng số lũy thừa	$a^n x(n)$	$X(a^{-1}Z)$
4. Trọng số tuyến tính	$nx(n)$	$-Z \frac{dX(Z)}{dZ}$
5. Đảo ngược thời gian	$x(-n)$	$X(Z^{-1})$
6. Tích chập	$x(n)*h(n)$	$X(Z)H(Z)$
7. Nhân dãy	$x(n)w(n)$	$\frac{1}{2\pi j} \oint_C X(\nu) W\left(\frac{z}{\nu}\right) \nu^{-1} d\nu$

### 2.1.1 Phép biến đổi Fourier

Biểu diễn biến đổi Fourier (FT) của tín hiệu thời gian rời rạc cho bởi các phương trình

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}, \quad (5a)$$

$$x(n) = \frac{1}{2\pi j} \oint_C X(e^{j\omega}) e^{j\omega n} d\omega. \quad (5b)$$

### 2.1.2 Phép biến đổi Fourier rời rạc

Cũng như trong trường hợp các tín hiệu tương tự, nếu dãy tuần hoàn với chu kỳ N,

$$\tilde{x}(n) = \tilde{x}(n + N) \quad -\infty < n < \infty \quad (7)$$

thì  $x(n)$  có thể biểu diễn bởi tổng rời rạc của các đường hình sin hơn là bởi dạng tích phân như ở (5b). Các biểu diễn dạng chuỗi Fourier cho dãy tuần hoàn là:

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) e^{-j\frac{2\pi}{N}kn} \quad (8a)$$

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) e^{j\frac{2\pi}{N}kn} \quad (8b)$$

Đó là biểu diễn chính xác của dãy tuần hoàn. Tuy nhiên, người ta hay dùng biểu diễn khác của (8). Xét dãy độ dài hữu hạn,  $x(n)$ , bằng 0 ngoài đoạn  $0 \leq n \leq N-1$ . Biến đổi ZT của  $x(n)$  là

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \quad (9)$$

Nếu ta đánh giá  $X(z)$  tại N điểm cách đều nhau trên đường tròn đơn vị,  $z_k = e^{j2\pi k/N}$ ,  $k = 0..(N-1)$ , thì có

$$X(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 0..(N-1) \quad (10)$$

	Dãy	DFT N-điểm
1. Tuyến tính (Linear)	$ax_1(n) + bx_2(n)$	$aX_1(k) + bX_2(k)$
2. Dịch chuyển (Shift)	$x((n + n_0))_N$	$e^{j\frac{2\pi}{N}kn_0} X(k)$
3. Đảo ngược thời gian (Time Reversal)	$x((-n))_N$	$X^*(k)$
4. Chập (Convolution)	$\sum_{m=0}^{N-1} x(m)h((n-m))_N$	$X(k)H(k)$
5. Nhân dãy (Multiplication of Sequence)	$x(n)w(n)$	$\frac{1}{N} \sum_{r=0}^{N-1} X(r)W((k-r))_N$

Bảng 2. Các dãy và DFT tương ứng của chúng.

Biểu diễn DFT với tất cả các nét riêng của nó là quan trọng do một số lý do:

- Biến đổi DFT,  $X(k)$ , có thể coi là bản mẫu của biến đổi ZT (hoặc biến đổi FT) của dãy có độ dài hữu hạn.
- Biến đổi DFT có các tính chất rất giống (có các sửa đổi do sự tuần hoàn nội tại) với nhiều tính chất hữu ích của biến đổi ZT và FT.
- $N$  giá trị của  $X(k)$  có thể tính toán rất hiệu quả (với thời gian tỷ lệ với  $N \log N$ ) bằng tập hợp các thuật toán tính toán được biết chung là *biến đổi Fourier nhanh* (Fast Fourier Transform, FFT).
- DFT được dùng rộng rãi để tính các *ước lượng phổ* (Spectrum estimate), *hàm tương quan* (Correlation function) và để thực hiện các lọc số.

### 2.1.3 Các bộ lọc số và cửa sổ

Lọc số là hệ thống bất biến dịch chuyển tuyến tính thời gian rời rạc (Discrete-Time Linear Shift-Invariant System). Nhớ rằng với hệ thống như vậy, cái vào và cái ra có quan hệ theo biểu thức tích chập (1). Quan hệ tương ứng giữa biến đổi ZT của cái vào và cái ra cho ở bảng 1

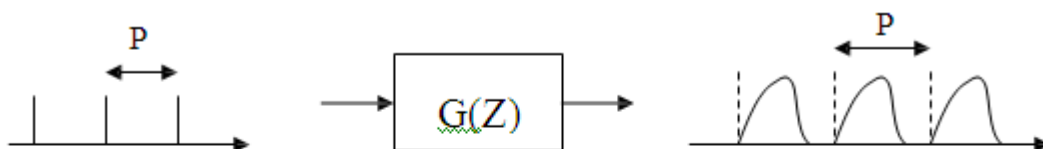
$Y(z) = H(z)X(z)$  Biến đổi ZT của đáp ứng mẫu đơn vị,  $H(z)$ , được gọi là hàm hệ thống (system function) của hệ, biến đổi FT của đáp ứng xung đơn vị,

## 2.2 Biểu diễn tín hiệu tiếng nói

### 2.2.1 Mô hình âm học của việc tạo tiếng nói

Nhằm đơn giản hoá việc phân tích và nghiên cứu bộ máy phát âm, người ta chia bộ máy phát âm ra làm hai phần cơ bản: nguồn âm và hệ thống đáp ứng.

- ⌚ Hệ thống đáp ứng bao gồm thanh môn, tuyến âm, môi và mũi. Việc mô hình hoá này sử dụng hàm truyền đạt trong biến đổi Z.
- ⌚ Đối với các âm hữu thanh, nguồn âm là một dạng sóng tuần hoàn đặc biệt. Dạng sóng này được mô phỏng bởi đáp ứng của bộ lọc thông thấp có hai điểm cực thực và tần số cắt vào khoảng 100 Hz.



Hình 1.4. Mô Hình hoá nguồn âm đôi với âm hữu thanh

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

Trong đó  $\alpha, \beta$  là các hằng số đặc trưng cho nguồn âm với  $\alpha < 1, \beta < 1$ .

Đối với âm vô thanh nguồn âm là một nhiễu trắng với biên độ biến đổi gần như ngẫu nhiên.

Để tạo tiếng nói, người ta dùng các mô hình khác nhau để mô phỏng bộ máy phát âm. Theo quan điểm giải phẫu học, ta có thể giả thiết rằng tuyến âm được biểu diễn bằng một chuỗi  $M$  đoạn ống âm học lý tưởng, là những đoạn ống có độ dài bằng nhau, và từng đoạn riêng biệt có thiết diện mặt cắt là  $A_m$  (gọi tắt là thiết diện) khác nhau theo chiều dài đoạn ống. Tổ hợp thiết diện

$\{A_m\}$  của các đoạn ống được chọn sao cho chúng xấp xỉ với hàm thiết diện  $A(x)$  của tuyến âm.

Các đoạn ống được coi là lý tưởng khi:

- ⌚ Độ dài mỗi đoạn đủ nhỏ so với bước sóng âm truyền qua nó được coi là sóng phẳng.
- ⌚ Các đoạn đủ cứng sao cho sự hao tổn bên trong do dao động thành ống, tính dính và dẫn nhiệt không đáng kể.

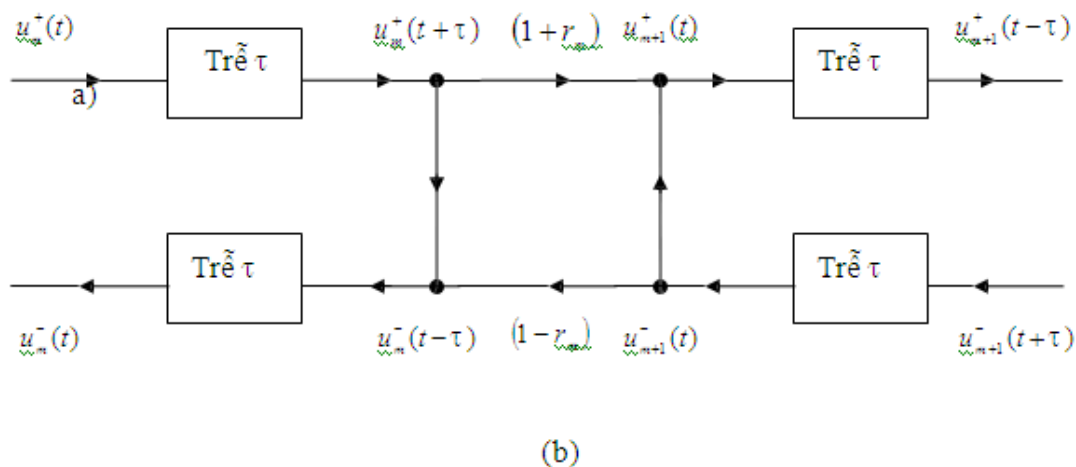
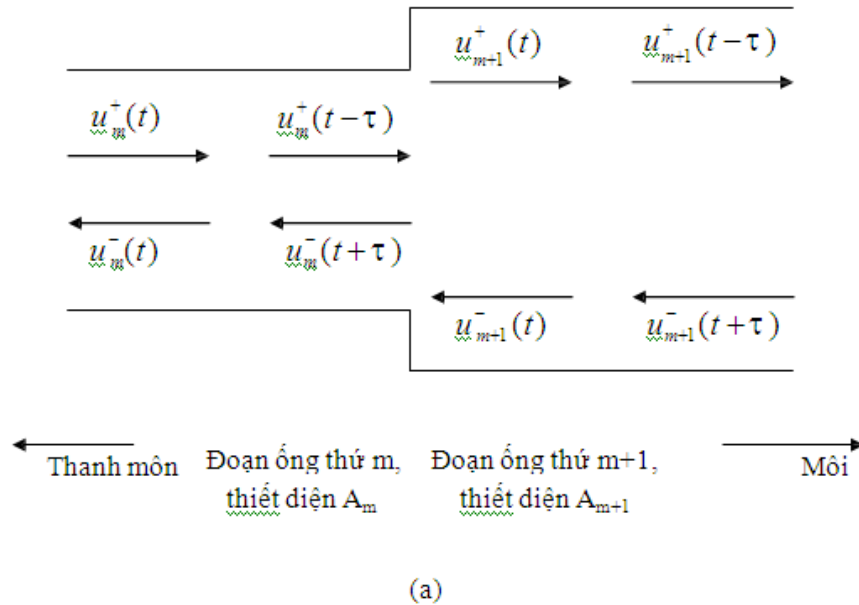
Ngoài ra ta giả thiết thêm mô hình tuyến âm lúc này là tuyến tính và không nối với thanh môn, hiệu ứng của tuyến mũi được bỏ qua, ta sẽ có mô hình tạo tiếng nói lý tưởng và việc phân tích mô hình ống âm học trở nên phức tạp hơn. Tiếp theo chúng ta có thể thấy rằng mô hình này có nhiều tính chất chung với mạch lọc số nên nó có thể được biểu diễn bằng cấu trúc mạch lọc số với các tham số thay đổi phù hợp với sự thay đổi tham số của ống âm học.

Sự chuyển động của không khí trong một đoạn ống âm học có thể được mô tả bằng áp suất âm thanh và thông lượng, đó là những hàm phụ thuộc độ dài ống ( $x$ ) và thời gian ( $t$ ). Trong những đoạn riêng biệt đó, các giá trị của hai hàm này được coi là tổ hợp tuyến tính các giá trị của chúng đối với sóng thuận và sóng ngược (được ký hiệu lần lượt bằng dấu cộng '+' và dấu trừ '-'). Sóng



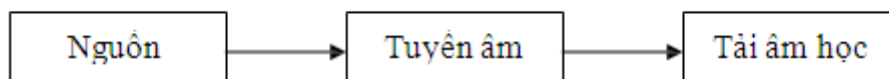
thuận là sóng truyền từ thanh môn đến môi, trong khi sóng ngược lại truyền từ môi đến thanh môn.

Mối quan hệ giữa sóng thuận và sóng ngược trong những đoạn kế tiếp phải đảm bảo áp suất và thông lượng liên tục cả về thời gian và không gian tại mọi điểm trong hệ thống. Trong hình 1.6.a ta thấy khi sóng thuận trong một đoạn gặp phần thay đổi về thiết diện (mối nối giữa hai đoạn kế tiếp), một phần của nó truyền sang đoạn kế tiếp, một phần kia lại phản xạ dưới dạng sóng ngược. Hoàn toàn tương tự, khi sóng ngược gặp mối nối, một phần được chuyển tiếp sang đoạn trước đó, còn phần kia lại phản xạ lại dưới dạng sóng thuận.



**Hình 1.6 Cách biểu diễn lý học và toán học**

- Mô hình lý học giữa đoạn ống m và m+1
- Mô hình toán học của đoạn ống thứ m



Tuyến âm được coi như một chuỗi liên tiếp các ống âm học và được mô hình hoá

bởi một chuỗi gồm  $K$  bộ cộng hưởng. Khi đó hàm truyền đạt của tuyến âm có dạng:

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

Mỗi bộ cộng hưởng sẽ tạo ra một formant được đặc trưng bởi tần số trung tâm, tính theo công thức:

$$f_i = \frac{1}{2\pi} \cos^{-1} \frac{-b_{1i}}{2\sqrt{b_{2i}}}$$

Với  $f_e$  là tần số lấy mẫu của tín hiệu lấy mẫu

Cuối cùng âm thanh được phát ra ở môi, nơi được coi như một tải âm học.

Sự tán xạ của môi được biểu diễn bởi hàm truyền đạt:

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

Hàm truyền đạt của hệ thống có dạng:

$$T(z) = G(z)V(z)R(z)$$

Nếu giả thiết một trong hai điểm cực của thanh môn gần bằng 1 ( $\beta = -1$ ) ta có:

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

$$\text{Với } A(z) = (1 + \alpha z^{-1}) \prod_{i=1}^K (1 + b_{1i} z^{-1} + b_{2i} z^{-2})$$

$$\text{Hay } A(z) = 1 + \sum_{i=1}^{2K+1} \alpha_i z^{-i}$$

là hàm truyền đạt của bộ lọc đảo.  $T(z)$  là hàm truyền đạt của mô hình toàn điểm cực. Các hệ số  $a_i$  của bộ lọc đảo sẽ là các tham số quan trọng trong phương pháp dự đoán tuyến tính để xác định các formant của tuyến âm.

Hạn chế của mô hình này là không thể tạo ra các âm xát hữu thanh và các âm mũi. Đối với các âm mũi mô hình trên được cải tiến bằng cách thêm vào phần đặc trưng cho mũi đặt song song với mô hình. Lúc đó hàm truyền đạt của hệ thống mới là:

$$\frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \frac{\sigma_1 A_2(z) + \sigma_2 A_1(z)}{A_1(z)A_2(z)}$$

Hệ thống trên không còn là hệ thống toàn điểm cực mà nó còn xuất hiện các điểm không trong mặt phẳng Z. Việc xuất hiện các điểm không này sẽ gây khó khăn cho phương pháp tiên đoán tuyến tính là phương pháp áp dụng cho các hệ thống toàn điểm cực. Song người ta đã khắc phục được khó khăn trên bằng cách thay một điểm không bằng hai điểm cực theo phương pháp giảm bậc gần đúng, công thức giảm bậc như sau:

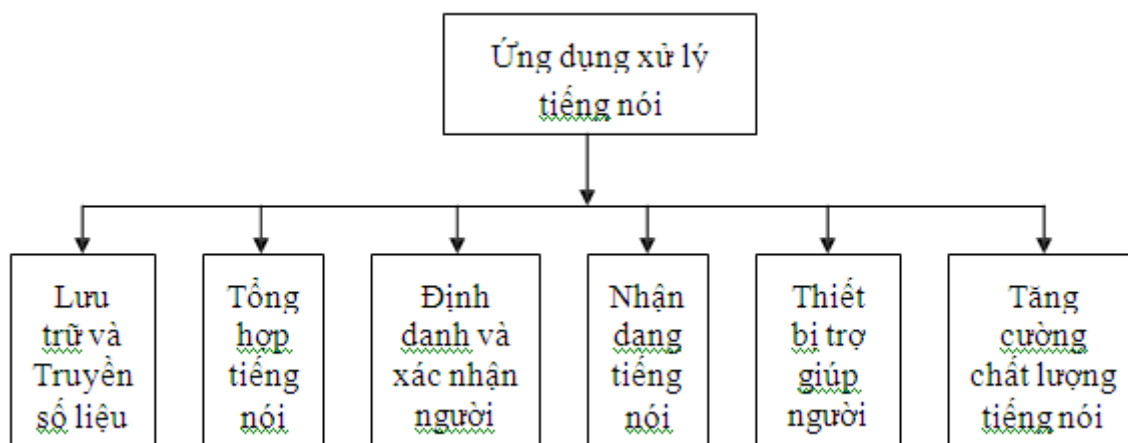
$$1 - \alpha z^{-1} \approx \frac{1}{1 + \alpha z^{-1} + \alpha^2 z^{-2} + \dots}$$

Tín hiệu âm thanh không phải là tín hiệu dừng, do đó mô hình phải được xây dựng một cách liên tục, nghĩa là các tham số của mô hình phải biến thiên theo thời gian. Sự biến thiên này rất chậm nên các tham số có thể coi như không đổi trong khoảng thời gian mà tín hiệu được coi là dừng: 20 ms.

## 2.3 Mã hóa tiếng nói

### 2.3.1 Các tính chất của bộ mã hóa tiếng nói

Dựa trên cơ sở lựa chọn các cách biểu diễn tín hiệu và phương pháp xử lý, đã có rất nhiều các ứng dụng quan trọng đã được triển khai. Hình vẽ dưới đây sẽ chỉ ra một số ứng dụng trong lĩnh vực xử lý tiếng nói.



Tổng hợp tiếng nói là quá trình tạo ra tín hiệu âm thanh bằng cách điều khiển một mô hình mẫu với một tập các tham số. Nếu mô hình mẫu này và các tham số được xây dựng một cách hoàn hảo thì tiếng nói tổng hợp có thể giống với tiếng nói tự nhiên. Hiện có hai phương pháp tổng hợp tiếng nói:

#### a. Tổng hợp tiếng nói theo cách phát âm

Đây là cách tiếp cận trực tiếp để mô hình hoá hệ thống một cách chi tiết. Trong phương pháp này hệ thống tổng hợp được mô phỏng giống như quá trình tạo ra âm thanh và lan truyền âm thanh trong hệ thống phát âm của con người. Hướng nghiên cứu này vẫn đang tiếp tục và

cho một số kết quả nhất định.

Phương pháp này có thể tạo ra hầu hết các tiếng nói tự nhiên.

### **b. Tổng hợp đầu cuối tự nhiên**

Theo hướng mô hình hoá này, người ta dựa trên các đặc tính đáp ứng tần số của dây thanh và tuyến âm để mô phỏng lại cơ chế tạo tiếng nói. Mô hình này gọi là mô hình nguồn-lọc. Bộ tổng hợp tiếng nói theo hướng này được thực hiện bằng cách sử dụng hệ thống tương tự với cơ chế tạo tiếng nói tại những điểm quan sát.

Cơ quan phát âm được mô hình hoá thành một hệ thống bao gồm một nguồn âm biểu diễn cho thanh môn và một bộ lọc biểu diễn cho tuyến âm. Quá trình tổng hợp sẽ bao gồm hai phần cơ bản:

- Tổng hợp tín hiệu nguồn dựa vào tần số cơ bản và tính chất tuần hoàn của nguồn.
- Xây dựng lại hàm truyền đạt của tuyến âm (bao gồm cả mũi và miệng) dựa vào các tham số đặc trưng cho tuyến âm.

Hiện nay người ta thường sử dụng hai bộ tham số đặc trưng cho tuyến âm:

- Bộ tham số formant
- Bộ tham số của bộ lọc đảo

Các bộ tham số này có thể được tổng kết từ các quá trình phân tích tiếng nói.

### **2.3.2 Các bộ mã hóa dạng sóng tiếng nói vô hướng**

Nhận dạng tiếng nói là lĩnh vực nghiên cứu với mục đích tạo ra được một thiết bị, máy móc hoặc phần mềm có khả năng nhận biết một cách chính xác tiếng nói của con người từ bất kỳ một nguồn phát âm nào. Nhận dạng tiếng nói có hai ứng dụng chính là nhận dạng tiếng nói và nhận dạng người nói.

#### **a. Nhận dạng ngữ nghĩa**

Thông thường để điều khiển các thiết bị máy móc người ta thường sử dụng cách giao tiếp thông qua sự vào ra cơ khí. Khi áp dụng tiếng nói vào giao tiếp, lợi ích của nó có thể dễ dàng nhận thấy: đó là tính tiện lợi, dễ sử dụng, tốc độ giao tiếp cao... Để có thể sử dụng tiếng nói như một công cụ giao tiếp thì hệ thống cần có khả năng tiếng nói về ngữ nghĩa. Nhận dạng ngữ nghĩa bao gồm nhận dạng từ và nhận dạng câu.

#### **b. Nhận dạng người nói**

Trong thế giới ngày nay tồn tại nhiều hệ thống yêu cầu độ an toàn bảo mật cao. Từ đó nảy sinh ra yêu cầu phải nhận dạng được người nói bằng những đặc điểm riêng biệt mà không ai có thể sao chép được. Bên cạnh các cách thức nhận dạng qua chữ ký, ảnh chân dung, chữ viết..., ngày nay người ta còn dùng tiếng nói để nhận dạng bởi vì tiếng nói có những đặc tính riêng biệt với từng người. Tại một số công ty đã xuất hiện những hệ thống kiểm tra người qua cửa bằng nhận dạng tiếng nói hoặc nhận dạng mỗi người qua thẻ nhận dạng mà những thông tin lưu trữ trên thẻ

chính là đặc điểm về tiếng nói của người đó.

Nguyên tắc của nhận dạng người nói là sử dụng những từ khoá đã được xác định từ trước mà những từ khoá này đặc trưng cho từng người một. Có hai yếu tố để khẳng định sự khác nhau trong tiếng nói của mỗi người:

- Các đặc tính cơ quan phát âm khác nhau như: độ dài của tuyến âm, tần số cộng hưởng của dây thanh, các tần số formant, dải thông, sự biến đổi của đường bao phổ... Đó là tập hợp những đặc tính có liên quan đến tính độc lập của nội dung âm vị của từ ngữ.
- Sự khác nhau trong cách phát âm của từng người: tốc độ và chiều dài từ luôn luôn khác nhau.

Trong tất cả các đặc tính trên đường bao phổ và tần số cơ bản là hai đặc tính quan trọng nhất. Đường bao phổ được miêu tả bằng những giá trị trung bình của các bộ lọc thông dải, của các tần số formant, của các hệ số tiên đoán tuyến tính, của hệ số cepstre và các tham số khác.

## **CÂU HỎI ÔN TẬP**

1. Trình bày ứng dụng của xử lý tín hiệu số trong xử lý tiếng nói ?
2. Trình bày mô hình âm học của việc tạo tiếng nói ?
3. Trình bày các tính chất của bộ mã hóa tiếng nói ?

## CHƯƠNG III : NHẬN DẠNG TIẾNG NÓI

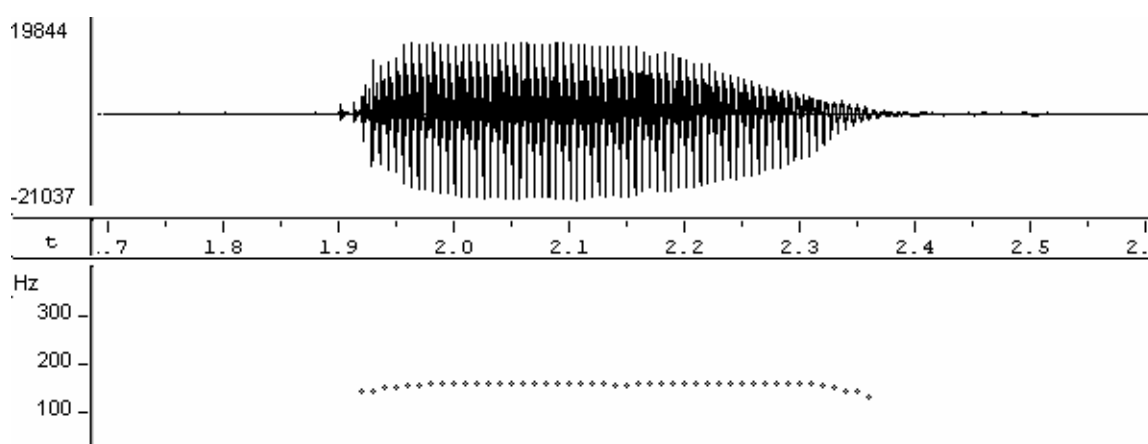
### 3.1 Các hệ thống nhận dạng tiếng nói

#### 3.1.1 Nhận dạng từ riêng lẻ

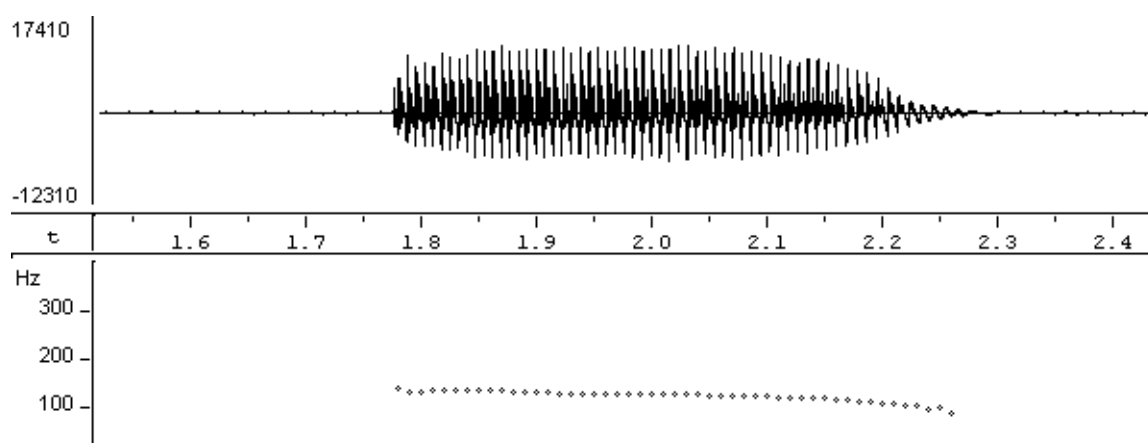
Trong tiếng Việt, ngữ nghĩa của một từ phụ thuộc vào thanh điệu. Khi thanh điệu thay đổi, nghĩa của từ cũng thay đổi theo. Có 6 thanh điệu trong tiếng Việt: không dấu, huyền, sắc, nặng, hỏi, ngã. Tương ứng với mỗi thanh điệu, tần số cơ bản thay đổi theo một quy luật riêng.

##### a. Không dấu

Với thanh điệu không dấu, tần số cơ bản không thay đổi.



##### b. Dấu huyền

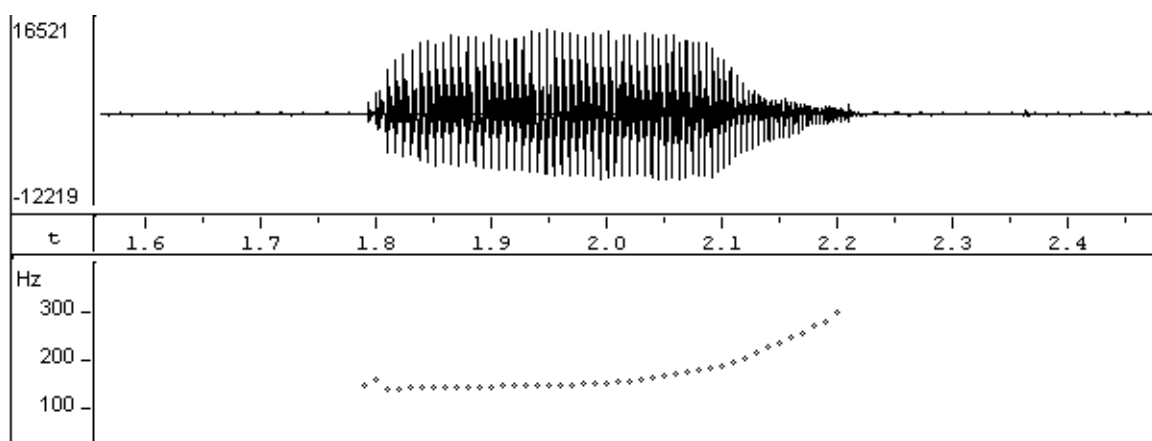


Với dấu huyền, tần số cơ bản giảm dần.

Nếu gọi  $F_0$  là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu huyền có thể được mô tả như sau:

$$F_0, F_0-10, F_0-20, F_0-30, F_0-50, F_0-60$$

##### c. Dấu sắc

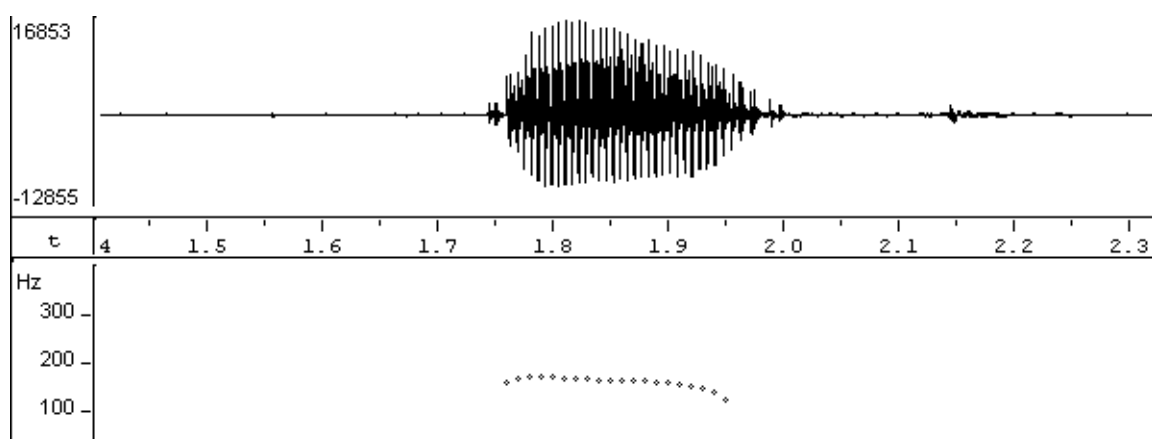


Với dấu sắc, tần số cơ bản tăng dần.

Nếu gọi  $F_0$  là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu sắc có thể được mô tả như sau:

$$F_0-20, F_0-20, F_0-15, F_0-10, F_0-5, F_0+5, F_0+30, F_0+70, F_0+80$$

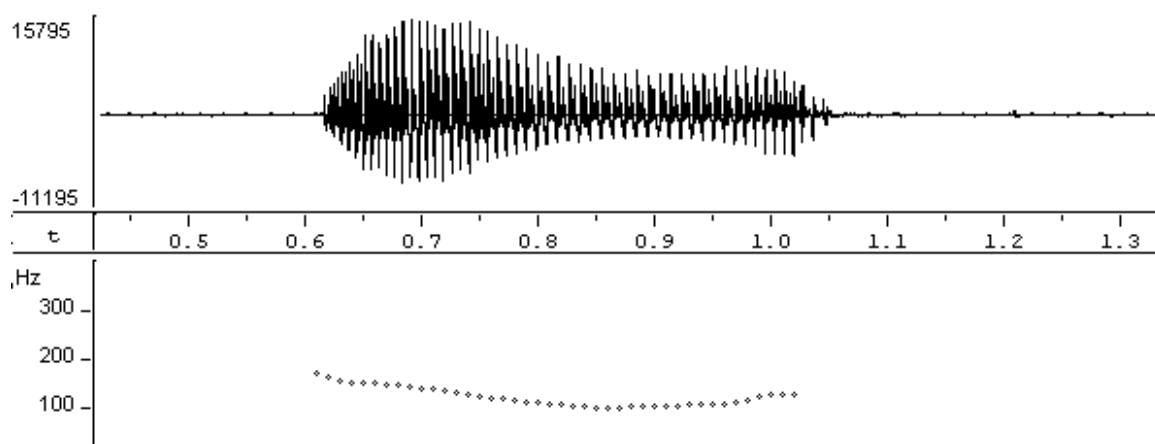
#### d. Dấu nặng



Nếu gọi  $F_0$  là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu nặng có thể được mô tả như sau:

$$F_0, F_0, F_0-35, F_0-50, F_0-90, F_0-120, F_0-140$$

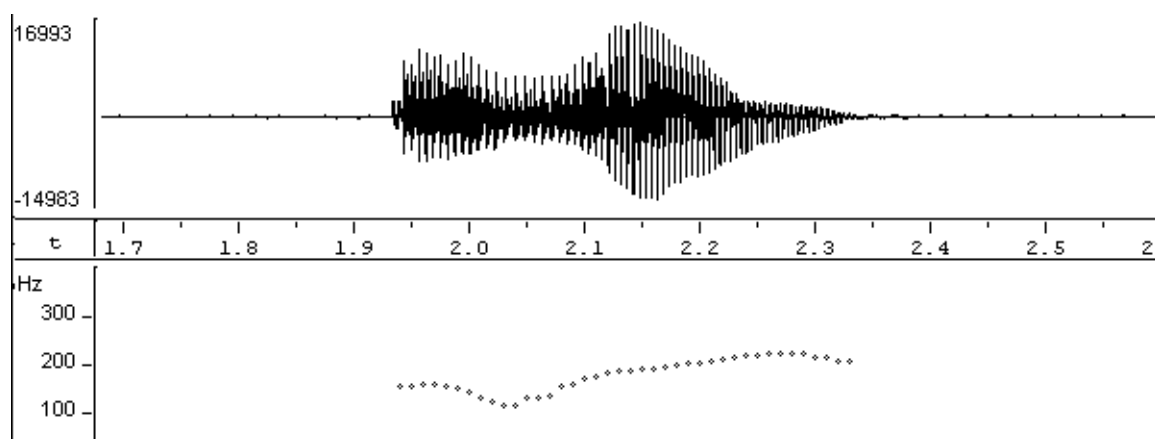
#### e. Dấu hỏi



Nếu gọi  $F_0$  là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu hỏi có thể được mô tả như sau:

$F_0-30, F_0-15, F_0-20, F_0-35, F_0-55, F_0-70, F_0-75, F_0-85, F_0-90, F_0-95, F_0-90,$   
 $F_0-80, F_0-90, F_0-30$

#### f. Dấu ngã



Nếu gọi  $F_0$  là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu ngã có thể được mô tả như sau:

$F_0, F_0-40, F_0+20, F_0+50, F_0+60$

### 3.1.2 Nhận dạng từ liên tục

Sự thay đổi các thông số của tín hiệu tiếng nói khi phát âm một câu trong tiếng Việt khá phức tạp, vì việc phát âm này phụ thuộc vào nhiều yếu tố như loại câu (câu hỏi, câu trần thuật, câu cảm thán...), hoàn cảnh phát âm (nói chuyện, đọc,...), địa phương... Để có được những hiểu biết về việc phát âm một câu trong tiếng Việt cần có những nghiên cứu đầy đủ.

Với mục đích thử nghiệm việc ghép từ để tạo thành câu trong tiếng Việt, phần này sẽ đưa ra một số nhận xét về sự biến đổi của tín hiệu tiếng nói khi phát âm hai loại câu điển hình của tiếng Việt: câu trần thuật và câu hỏi. Những nhận xét này được rút ra qua sự so sánh với câu không có ngữ điệu.

#### a. Câu trần thuật

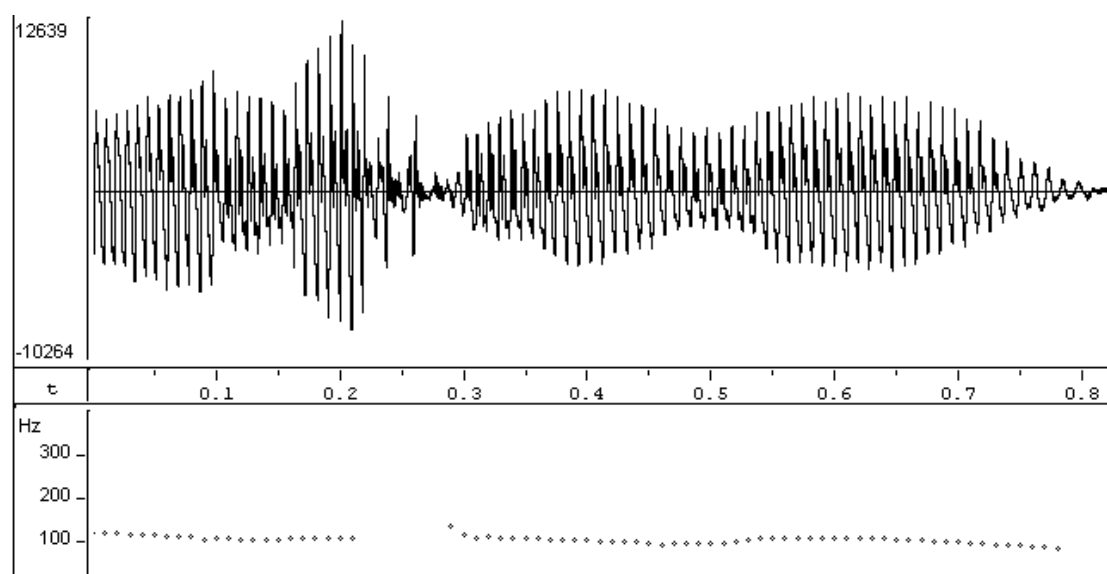
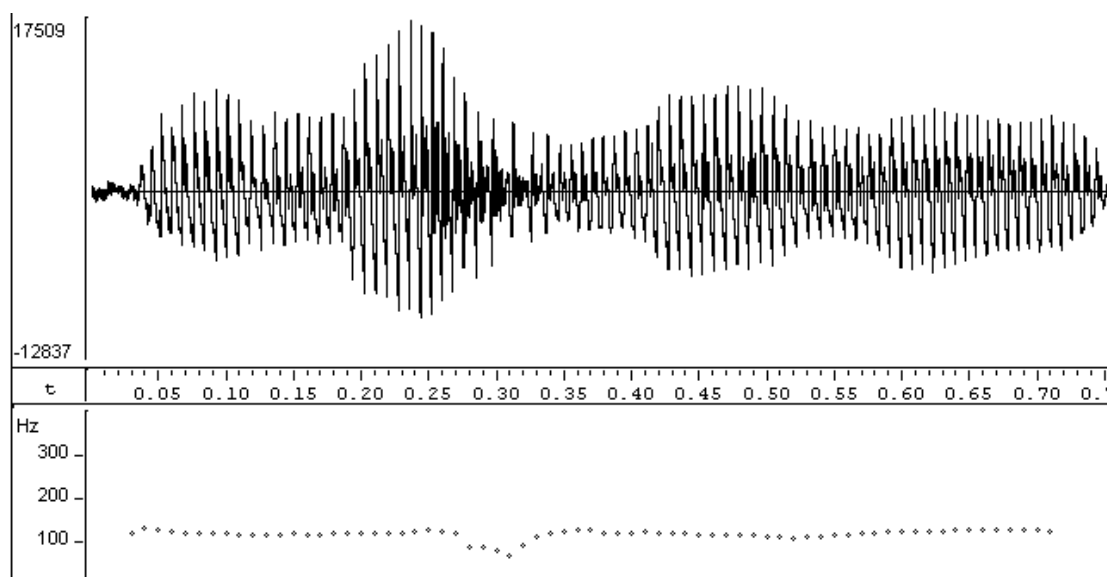


Khi phát âm câu trần thuật, tùy theo hoàn cảnh có thể có một số từ nào đó được nhấn mạnh. Việc xác định từ cần nhấn mạnh trong câu trần thuật liên quan tới phân tích bậc cao và không được đề cập tới ở đây. Để đơn giản, giả sử không có từ nào được nhấn mạnh rõ ràng trong câu.

So sánh hai cách phát âm có thể rút ra các nhận xét sau:

- Về thời gian phát âm: Do không có từ nhấn mạnh nên các từ trong câu không ngữ điệu và câu trần thuật được phát âm trong khoảng thời gian gần như nhau.
- Về biên độ tín hiệu: Các từ trong câu không ngữ điệu được phát âm với biên độ tương đối đều. Biên độ các từ trong câu trần thuật giảm dần ở cuối câu.
- Về tần số cơ bản: Trong câu không ngữ điệu, tần số cơ bản của các từ (không có thanh điệu) đi theo đường nằm ngang. Tần số cơ bản của từ trong câu trần thuật giảm dần.

Như vậy, các từ trong câu trần thuật được phát âm với biên độ và tần số cơ bản giảm dần về phía cuối câu.



**b. Câu hỏi**

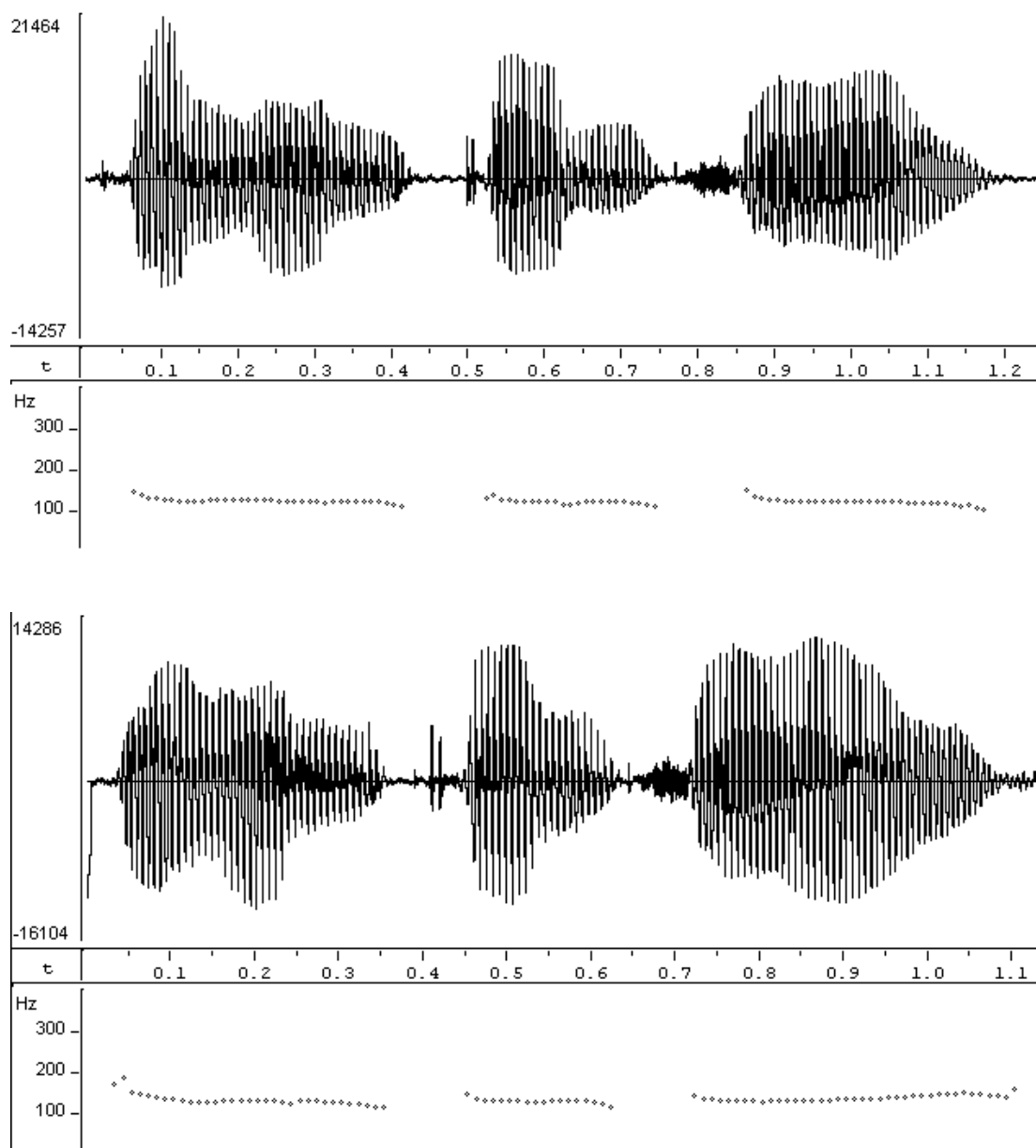
Trong câu hỏi, người nói thường nhấn mạnh vào từ cần hỏi. Những từ cần hỏi này thường không có vị trí cố định trong câu.

Ví dụ: Cùng một câu hỏi *Anh đi?* Nếu muốn hỏi về chủ ngữ (anh hoặc ai đó) thì người hỏi sẽ nhấn mạnh vào từ *anh*, nếu muốn hỏi về hành động (đi hoặc chạy) thì người hỏi sẽ nhấn mạnh vào từ *đi*.

Việc xác định từ để hỏi trong câu liên quan tới việc phân tích bậc cao trong quá trình tổng hợp và không được đề cập ở đây. Để đơn giản, từ để hỏi

trong các câu thử nghiệm được coi là từ cuối câu. Câu hỏi sẽ được so sánh với câu không có ngữ điệu.

Dưới đây là hình ảnh dạng sóng và tần số cơ bản của câu: *Anh ăn chưa* (không có ngữ điệu) và câu *Anh ăn chưa?* (từ để hỏi là *chưa*)



So sánh hai cách phát âm có thể rút ra các nhận xét sau:

- ☐ Về thời gian phát âm: Các từ trong câu không ngữ điệu được phát âm trong khoảng thời gian gần như nhau. Từ để hỏi trong câu hỏi (*chưa*) được phát âm dài hơn (0.45s) các từ *anh* (0.35s) và *ăn* (0.20s) trong câu này.

- Về biên độ tín hiệu: Các từ trong câu không ngữ điệu được phát âm với biên độ tương đối đều. Từ để hỏi *chưa* trong câu hỏi được phát âm với biên độ lớn hơn từ *chưa* trong câu không ngữ điệu.
- Về tần số cơ bản: Trong câu không ngữ điệu, tần số cơ bản của các từ (không có thanh điệu) đi theo đường nằm ngang. Tần số cơ bản của từ *anh* và *ăn* trong câu hỏi không tăng dần. Tần số cơ bản của từ *chưa* trong câu hỏi tăng dần.

Như vậy, các từ để hỏi trong câu hỏi được phát âm dài hơn, với biên độ lớn hơn và tần số cơ bản tăng dần so với câu không ngữ điệu.

## 3.2 Các mô hình Markov ẩn

### 3.2.1 Chuỗi Markov

Trong toán học, một xích Markov hay chuỗi Markov (thời gian rời rạc), đặt theo tên nhà toán học người Nga Andrei Andreyevich Markov, là một quá trình ngẫu nhiên thời gian rời rạc với tính chất Markov. Trong một quá trình như vậy, quá khứ không liên quan đến việc tiên đoán tương lai mà việc đó chỉ phụ thuộc theo kiến thức về hiện tại.

Xích Markov là một dãy  $X_1, X_2, X_3, \dots$  gồm các biến ngẫu nhiên. Tập tất cả các giá trị có thể có của các biến này được gọi là không gian trạng thái  $S$ , giá trị của  $X_n$  là trạng thái của quá trình (hệ) tại thời điểm  $n$ .

Nếu việc xác định (dự đoán) phân bố xác suất có điều kiện của  $X_{n+1}$  khi cho biết các trạng thái quá khứ là một hàm chỉ phụ thuộc  $X_n$  thì:

$$P(X_{n+1} = x | X_0, X_1, X_2, \dots, X_n) = P(X_{n+1} = x | X_n),$$

trong đó  $x$  là một trạng thái nào đó của quá trình ( $x$  thuộc không gian trạng thái  $S$ ). Đó là thuộc tính Markov.

Một cách đơn giản để hình dung một kiểu chuỗi Markov cụ thể là qua một ô tômat hữu hạn (finite state machine). Nếu hệ ở trạng thái  $y$  tại thời điểm  $n$  thì xác suất mà hệ sẽ chuyển tới trạng thái  $x$  tại thời điểm  $n+1$  không phụ thuộc vào giá trị của thời điểm  $n$  mà chỉ phụ thuộc vào trạng thái hiện tại  $y$ . Do đó, tại thời điểm  $n$  bất kỳ, một xích Markov hữu hạn có thể được biểu diễn bằng một ma trận xác suất, trong đó phần tử  $x, y$  có giá trị bằng  $P(X_{n+1} = x | X_n = y)$  và độc lập với chỉ số thời gian  $n$  (nghĩa là để xác định trạng thái kế tiếp, ta không cần biết đang ở thời điểm nào mà chỉ cần biết trạng thái ở thời điểm đó là gì). Các loại xích Markov hữu hạn rời rạc này còn có thể được biểu diễn bằng đồ thị có hướng, trong đó các cung được gắn nhãn bằng xác suất chuyển từ trạng thái tại đỉnh (vertex) đầu sang trạng thái tại đỉnh cuối của cung đó.

Markov đã đưa ra các kết quả đầu tiên (1906) về các quá trình này. Andrey Nikolaevich Kolmogorov (1936) đã đưa ra một suy rộng tới các không gian trạng thái vô hạn đếm được.

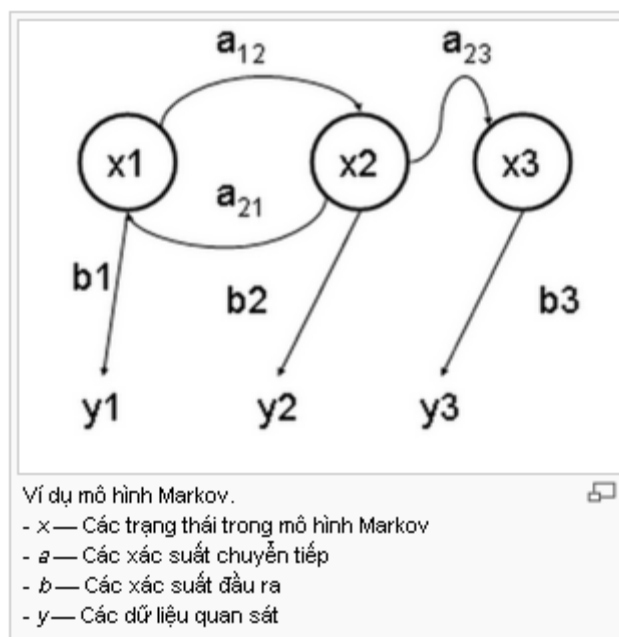
Các xích Markov có liên quan tới chuyển động Brown (Brownian motion) và Tổng hợp ergodic, hai chủ đề quan trọng của vật lý trong những năm đầu của thế kỷ 20, nhưng Markov có vẻ phải tham gia vào quá trình phát triển của toán học, còn gọi là sự mở rộng của luật số lớn cho các sự kiện độc lập.

### 3.2.2 Mô hình Markov

Mô hình Markov ẩn (tiếng Anh là Hidden Markov Model - HMM) là mô hình thống kê trong đó hệ thống được mô hình hóa được cho là một quá trình Markov với các tham số không biết trước và nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được, dựa trên sự thừa nhận này. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp, ví dụ cho các ứng dụng nhận dạng mẫu.

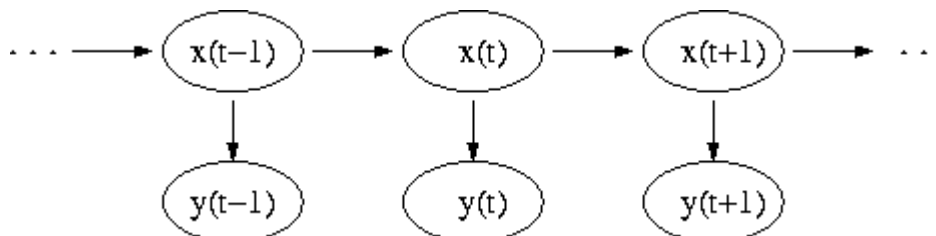
Trong một mô hình Markov điển hình, trạng thái được quan sát trực tiếp bởi người quan sát, và vì vậy các xác suất chuyển tiếp trạng thái là các tham số duy nhất. Mô hình Markov ẩn thêm vào các đầu ra: mỗi trạng thái có xác suất phân bố trên các biểu hiện đầu ra có thể. Vì vậy, nhìn vào dãy của các biểu hiện được sinh ra bởi HMM không trực tiếp chỉ ra dãy các trạng thái.

#### Các chuyển tiếp trạng thái trong mô hình Markov



### Sự tiến hóa của mô hình Markov

Biểu đồ trên đây làm nổi bật các chuyển tiếp trạng thái của mô hình Markov ẩn. Nó cũng có ích để biểu diễn rõ ràng sự tiến hóa của mô hình theo thời gian, với các trạng thái tại các thời điểm khác nhau  $t_1$  và  $t_2$  được biểu diễn bằng các tham biến khác nhau,  $x(t_1)$  và  $x(t_2)$ .



Trong biểu đồ này, nó được hiểu rằng thời gian chia cắt ra  $(x(t), y(t))$  mở rộng tới các thời gian trước và sau đó như một sự cần thiết. Thông thường lát cắt sớm nhất là thời gian  $t=0$  hay  $t=1$ .

### CÂU HỎI ÔN TẬP

1. Trình bày phương pháp nhận dạng từ riêng lẻ ?
2. Trình bày phương pháp nhận dạng từ liên tục ?
3. Trình bày mô hình Markov và ứng dụng của mô hình này trong hệ thống xử lý tiếng nói ?

## CHƯƠNG IV : CÁC HỆ THỐNG CHUYỂN VĂN BẢN THÀNH GIỌNG NÓI

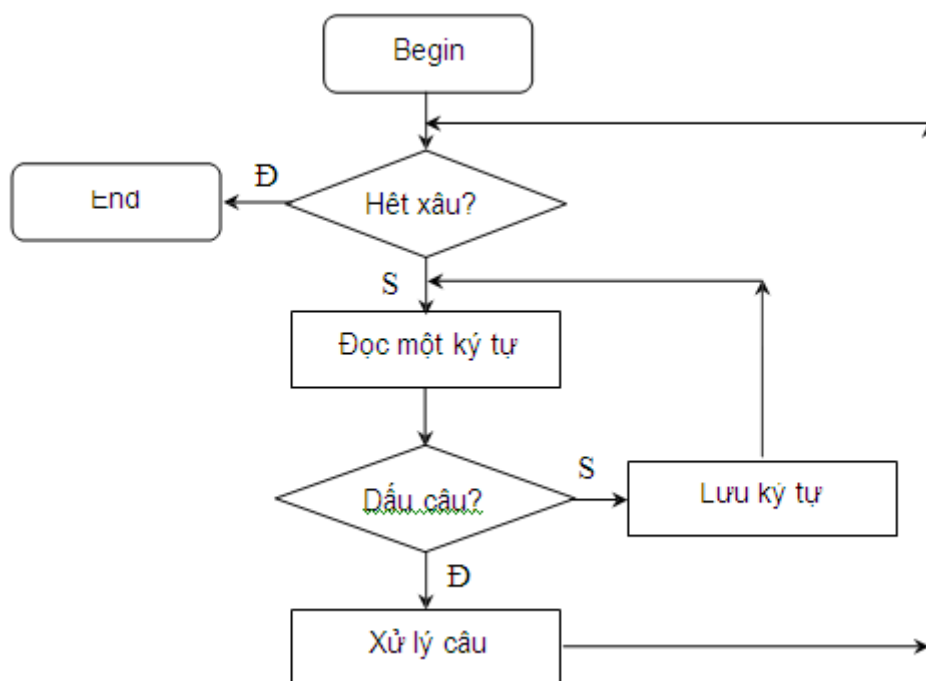
### 4.1 Phân tích ngữ âm và văn bản

#### 4.1.1 Từ vựng

Câu trong văn bản được ngăn cách với nhau bởi các dấu câu. Các dấu câu được cho trong bảng. Cần chú ý rằng khái niệm “*câu*” ở đây nhằm chỉ các loại câu khác nhau (trần thuật, hỏi...) để xác định sự biến đổi của tần số cơ bản và có thể không chặt chẽ về ngữ pháp.

Loại dấu câu	Cách viết
Dấu chấm	.
Dấu phẩy	,
Dấu chấm phẩy	;
Dấu hai chấm	:
Dấu chấm than	!
Dấu chấm hỏi	?
Các dấu ngoặc	( ) [ ] { }

Do chương trình chỉ xét các văn bản dưới dạng text nên toàn bộ văn bản được coi như một xâu ký tự. Các câu được xác định theo lưu đồ thuật toán sau:



#### 4.1.2 Xác định cấu trúc tài liệu

Sau khi được xác định, câu được phân loại để xử lý. Với mục đích thử

nghiệm tổng hợp câu, báo cáo này chỉ chia câu làm ba loại:

- ⌚ Loại 1 (câu trần thuật): tương ứng với các dấu: “.”, “,” “)”, “]”, “}”
- ⌚ Loại 2 (câu hỏi): tương ứng với dấu câu: “?”
- ⌚ Loại 3 (câu hơi lên giọng ở cuối câu): dấu “,”, “!”

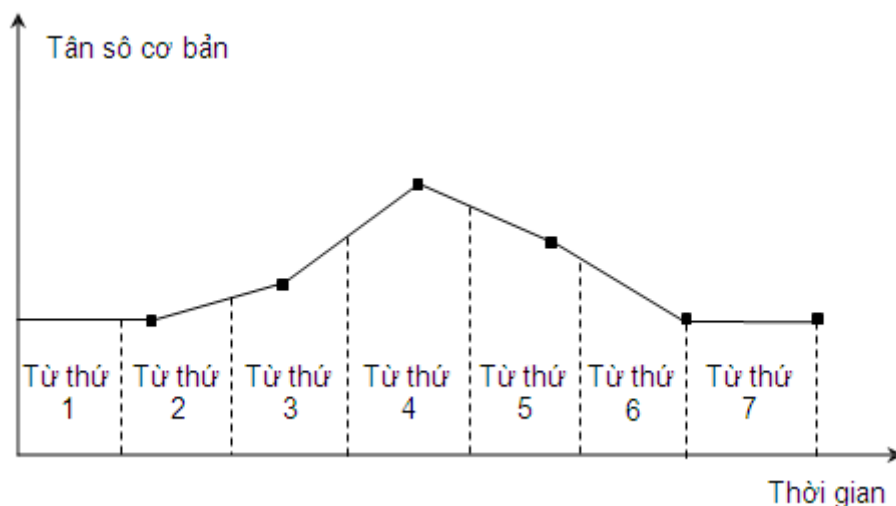
Sự biến đổi các thông số của tín hiệu tiếng nói tổng hợp phụ thuộc vào từng loại câu.

Vấn đề này được trình bày chi tiết trong mục 4.6.2.

Căn cứ vào sự biến đổi các thông số của tín hiệu tiếng nói, câu được phân tích thành các từ đi kèm với các thông số của từ. Các thông số của từ bao gồm:

- ⌚ Sự biến đổi tần số cơ bản
- ⌚ Biên độ
- ⌚ Trường độ

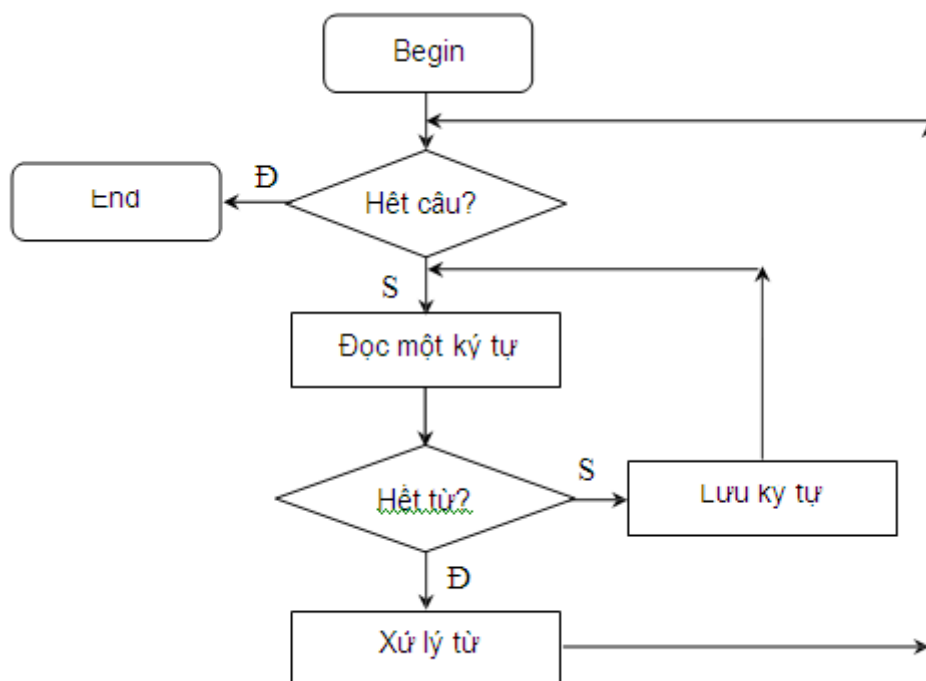
Hình dưới minh họa sự biến đổi tần số cơ bản của các từ theo sự biến đổi tần số cơ bản của câu.



Các từ được nhấn mạnh trong câu (ví dụ từ để hỏi trong câu hỏi) có biên độ và trường độ của từ này lớn hơn các từ khác.

Việc tách từ trong câu được thực hiện theo lưu đồ thuật toán ở trên.





### 4.1.3 Chuẩn hóa văn bản

Để tiện xử lý về sau (sử dụng các bảng mã tiếng Việt khác nhau), trước khi tách thành hai diphone từ được chuyển thành dạng telex. Dấu của từ được viết ở cuối từ.

Ví dụ: từ *trường* được chuyển thành *truwowngf*

Việc chuyển từ dạng tiếng Việt thông thường sang dạng telex tùy thuộc vào loại bảng mã được sử dụng. Chương trình sử dụng bảng mã 8 bit TCVN3- ABC

### 4.1.4 Phân tích ngôn ngữ

#### 4.1.5 Chuyển đổi ký tự sang âm thanh

Từ ở dạng biểu diễn telex được tách thành hai diphone bắt đầu và kết thúc tương ứng. Diphone bắt đầu được phân biệt bằng dấu “\_” phía trước, diphone kết thúc có dấu “\_” phía sau.

Ví dụ: từ *truwowngf* được tách thành hai diphone *\_truw* và *uwowng\_*

Mấu chốt của việc tách một từ thành hai diphone là phát hiện được vị trí bắt đầu và kết thúc của nguyên âm đầu tiên (theo chiều từ trái sang phải).

Ví dụ: nếu tìm được nguyên âm *u* (*uw*) thì dễ dàng tách từ *truwowng* thành *truw* và *uwowng*.

Thuật toán xác định vị trí bắt đầu và kết thúc của nguyên âm đầu tiên được cho trong hình 4.7.

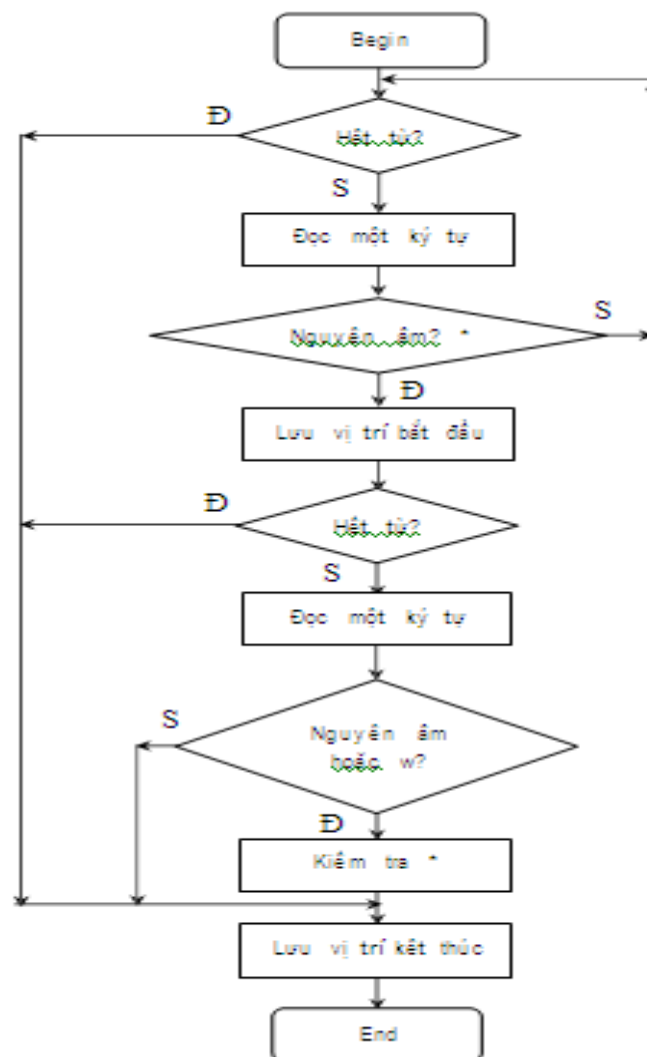
Trong lưu đồ 4.7. \* ứng với quá trình kiểm tra xem hai ký tự liên tiếp có phải là *aa*, *aw*, *ee*, *oo*, *ow*, *uw* hay không.

Việc xác định diphone kết thúc phải đi kèm với việc xác định dấu của từ,

vì có trường hợp diphone kết thúc không thể tạo thành từ diphone không dấu.

Ví dụ: từ *các* và *cạc* đều có diphone kết thúc là *ac\_*, diphone này không thể tạo thành từ diphone không dấu nên phải căn cứ vào dấu của từ để xác định diphone là *acs\_* hay *acj\_*.

Các trường hợp này tương ứng với những diphone in đậm trong bảng 4.1. Đa số các diphone được lưu trong cơ sở dữ liệu với tên là cách biểu diễn diphone, ví dụ diphone *an\_* có tên là *an\_* trong cơ sở dữ liệu, nhưng với diphone có cách biểu diễn dài, ví dụ *uwowng\_*, thì tên lưu trong cơ sở dữ liệu khác với cách biểu diễn *wog\_* (tên của các diphone trong cơ sở dữ liệu với kích thước 4 byte) nên cần chuyển đổi cách biểu diễn diphone phù hợp với tên trong cơ sở dữ liệu.



## 4.2 Tổng hợp tiếng nói

### 4.2.1 Các tính chất của tổng hợp tiếng nói

Tổng hợp tiếng nói là phát sinh tiếng nói từ sóng tiếng nói. Trong vài thập niên gần đây, các bộ tổng hợp tiếng nói có chất lượng ngày càng cao. Tuy nhiên chất lượng của các phương pháp

hiện nay mới chỉ đạt đến mức phù hợp cho một vài ứng dụng, chẳng hạn như đa phương tiện và truyền thông.

Hiện nay có ba phương pháp tổng hợp tiếng nói. Phương pháp đơn giản nhất để phát sinh tiếng nói tổng hợp là phát các mẫu tiếng nói đã thu từ tiếng nói tự nhiên (như các từ hoặc câu). Phương pháp này cho chất lượng tương đối tốt nhưng gặp phải hạn chế là số lượng từ vựng trong cơ sở dữ liệu rất lớn. Bên cạnh đó tiếng nói cũng có thể tạo ra bằng cách mô phỏng hệ thống phát âm. Phương pháp này cho chất lượng rất tốt nhưng thực hiện khá phức tạp. Một phương pháp nữa cũng được dùng để tổng hợp tiếng nói là tổng hợp formant. Các phương pháp tổng hợp tiếng nói cùng với những đặc điểm cơ bản nhất sẽ được giới thiệu trong phần tiếp theo.

#### **4.2.2 Tổng hợp tiếng nói bằng các Formant**

Phương pháp tổng hợp formant (formant synthesis) yêu cầu phải tổng hợp được tối thiểu 3 formant để hiểu được tiếng nói, và để có được tiếng nói chất lượng cao thì cần tới 5 formant. Tiếng nói được tạo ra từ các bộ tổng hợp formant với thành phần chính là các bộ cộng hưởng. Tùy theo cách bố trí các bộ cộng hưởng mà ta có bộ tổng hợp formant là nối tiếp hay song song.

##### **a. Bộ tổng hợp formant nối tiếp**

Bộ tổng hợp formant nối tiếp là một bộ tổng hợp formant có các tầng nối tiếp, đầu ra của bộ cộng hưởng này là đầu vào của bộ cộng hưởng kia.

##### **b. Bộ tổng hợp formant song song**

Bộ tổng hợp formant song song bao gồm các bộ cộng hưởng mắc song song. Đầu ra là kết hợp của tín hiệu nguồn và tất cả các formant. Cấu trúc song song cần nhiều thông tin để điều khiển hơn.

#### **4.2.3 Tổng hợp tiếng nói bằng ghép nối**

Tổng hợp bằng cách ghép nối các âm được tổng hợp từ các lời nói tự nhiên đã được thu từ trước có lẽ là cách dễ nhất để sản sinh lời nói. Phương pháp tổng hợp ghép nối cho chất lượng cao và tương đối tự nhiên. Phương pháp này rất phù hợp với các hệ thống phát thanh và các hệ thống thông tin. Tuy nhiên phương pháp này thường chỉ áp dụng cho một giọng và phải sử dụng nhiều bộ nhớ hơn các phương pháp khác do số lượng từ vựng rất lớn. Để khắc phục nhược điểm này người ta xây dựng các phương pháp tổng hợp ghép nối từ những đơn vị nhỏ như âm vị, âm tiết, diphone (âm vị kép)... Ngoài các diphone, chúng ta còn sử dụng triphone, tetraphone hay syllable, demisyllable, nhưng chủ yếu vẫn là các diphone, được thu từ tiếng nói tự nhiên. Các diphone được cắt ra từ tín hiệu rồi sau đó được tổng hợp lại theo yêu cầu dựa trên một thuật toán ghép nối.

Phương pháp này có một số khác biệt so với các phương pháp khác:

- Xuất hiện sự biến dạng của tiếng nói tổng hợp do tính không liên tục của việc ghép nối các diphone với nhau. Vì vậy phải sử dụng biện pháp làm trơn tín hiệu.
- Bộ nhớ yêu cầu cao, nhất là khi các đơn vị kết nối dài như là các âm vị hay các từ.
- Suu tầm và gấn nhãn dữ liệu tiếng nói cần nhiều thời gian và công sức. Về lý thuyết tất cả các mẫu cần phải được lưu trữ. Số lượng và chất lượng các mẫu lưu trữ là một vấn đề cần giải quyết khi tiến hành lưu trữ.

Hiện nay phương pháp này đang được sử dụng rộng rãi trên thế giới và ngày càng cho chất lượng tốt hơn nhờ sự trợ giúp của máy tính.

Phần tiếp theo sẽ giới thiệu về một phương pháp tổng hợp ghép nối được áp dụng phổ biến cho tín hiệu tiếng nói, phương pháp ghép nối dựa trên giải thuật PSOLA.

#### **a. Phương pháp tổng hợp PSOLA**

PSOLA (Pitch Synchronous Overlap Add) là phương pháp tổng hợp dựa trên sự phân tích một tín hiệu thành một chuỗi các tín hiệu thành phần. Khi cộng xếp chồng (overlap-add) các tín hiệu thành phần ta có thể khôi phục lại tín hiệu ban đầu.

PSOLA thao tác trực tiếp với tín hiệu dạng sóng, không dùng bất cứ loại mô hình nào nên không làm mất thông tin của tín hiệu. PSOLA cho phép điều khiển độc lập tần số cơ bản, chu kỳ cơ bản và các formant của tín hiệu. Ưu điểm chính của phương pháp PSOLA là giữ nguyên đường bao phổ khi thay đổi tần số cơ bản (pitch shifting). Phương pháp này cho phép biến đổi tín hiệu ngay trên miền thời gian nên chi phí tính toán rất thấp. PSOLA đã được dùng rất phổ biến với tín hiệu tiếng nói.

#### **b. Các phiên bản của PSOLA**

Dựa trên PSOLA, người ta đã đưa ra nhiều phiên bản khác nhau, dưới đây là các phiên bản chính:

##### **TD-PSOLA**

Phương pháp TD-PSOLA (Time Domain- Pitch Synchronous Overlap Add) là phiên bản miền thời gian của PSOLA (TD-PSOLA). Phương pháp này thao tác với tín hiệu trên miền thời gian nên được sử dụng nhiều vì hiệu quả trong tính toán của nó. Phương pháp này sẽ được trình bày chi tiết trong chương tiếp theo.

##### **FD-PSOLA**

Phương pháp tổng hợp FD-PSOLA (Frequency Domain- Pitch Synchronous Overlap Add) là phương pháp bao gồm các bước giống như TD-PSOLA nhưng thao tác trên miền tần số. Phương pháp này có chi phí tính toán cao hơn TD-PSOLA. Đối với mỗi trường hợp riêng biệt thì mỗi phương pháp sẽ cho hiệu quả khác nhau, nên phải dựa vào từng hoàn cảnh để chọn phương pháp thích hợp.

##### **⌚ LP-PSOLA**

Ngoài các phương pháp trên miền thời gian, miền tần số, còn có một phương pháp gọi là phương pháp dự đoán tuyến tính (Linear Prediction - Pitch Synchronous Overlap Add). Phương pháp dự đoán tuyến tính được thiết kế để mã hoá tiếng nói nhưng phương pháp này cũng có thể dùng cho tổng hợp.

Cơ sở của phương pháp dự đoán tuyến tính dựa trên các mẫu  $y(n)$  có thể lấy xấp xỉ hoặc dự đoán từ  $p$  mẫu trước đó  $y(n-1)$  đến  $y(n-p)$  với sai số nhỏ nhất. Tín hiệu kích thích được lấy xấp xỉ bằng một dãy các tín hiệu tiếng nói và nhiễu ngẫu nhiên. Tín hiệu nguồn được cho qua bộ lọc số với hệ số  $a(k)$ .

Phương pháp LP-PSOLA cho kết quả chưa tốt. Người ta đã cải biến phương pháp này để thu được chất lượng tốt hơn, mà đại diện là phương pháp WLP (Warped Linear Prediction).

#### 4.2.4 Đánh giá các hệ thống tổng hợp tiếng nói

Sau khi giới thiệu những đặc điểm cơ bản nhất của các phương pháp tổng hợp tiếng nói ta có thể rút ra một số nhận xét về các phương pháp này. Các nhận xét này nhằm mục đích đưa ra đánh giá về ba phương pháp dựa trên chất lượng tiếng nói tổng hợp, chi phí tính toán và kích thước dữ liệu.

- ⌚ *Về chất lượng của tiếng nói tổng hợp:* Trong ba phương pháp nói trên thì phương pháp mô phỏng bộ máy phát âm về nguyên tắc sẽ cho chất lượng tốt nhất. Để đạt được điều này thì vấn đề quan trọng là làm sao mô phỏng chính xác bộ máy phát âm của con người. Công việc này hoàn toàn không đơn giản, mặc dù đã có sự trợ giúp của máy tính nhưng do cấu trúc phức tạp của bộ máy phát âm nên chi phí tính toán sẽ rất lớn. Trong hai phương pháp còn lại thì thực tế cho thấy phương pháp ghép nối thường cho chất lượng tốt hơn.
- ⌚ *Về hiệu quả tính toán:* Rõ ràng là phương pháp mô phỏng bộ máy phát âm đòi hỏi chi phí tính toán lớn nhất vì phải mô phỏng một cách chính xác nhất bộ máy phát âm phức tạp của con người. Hai phương pháp còn lại có chi phí tính toán thấp hơn do đặc điểm các thuật toán được sử dụng.
- ⌚ *Về kích thước dữ liệu:* Phương pháp ghép nối có kích thước dữ liệu lớn nhất do số lượng từ vựng là rất lớn. Hai phương pháp còn lại do không phải lưu trữ các mẫu nên có kích thước dữ liệu nhỏ hơn.

### CÂU HỎI ÔN TẬP

1. Trình bày ý nghĩa của việc chuẩn hóa văn bản?

2. Trình bày quá trình chuyển đổi ký tự sang âm thanh?
3. Trình bày các tính chất của tổng hợp tiếng nói?
4. Trình bày tổng hợp tiếng nói bằng các Formant?
5. Trình bày tổng hợp tiếng nói bằng phương pháp ghép nối?

# MỘT SỐ ĐỀ THI MẪU

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>1</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói dưới dạng sóng theo thời gian?
- Mã hóa tiếng nói dạng sóng vô hướng: xung tuyến tính?

**Câu 3: (4 điểm)**

- Mô hình chung của hệ thống nhận dạng tiếng nói?
- Mô hình markov? Ứng dụng của Markov trong nhận dạng tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi



Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>2</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền không gian 3 chiều: Spectrogram.?
- Mã hóa tiếng nói dạng sóng vô hướng: xung tuyến tính?

**Câu 3: (4 điểm)**

- Trình bày hệ thống chuyên đổi văn bản thành giọng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>3</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền tần số?
- Trình bày về Formant và Antiformant?

**Câu 3: (4 điểm)**

- Mô hình chung của hệ thống nhận dạng tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>4</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền tần số?
- Mã hóa tiếng nói dạng sóng vô hướng: xung tuyến tính?

**Câu 3: (4 điểm)**

- Các phương pháp nhận dạng tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>5</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền không gian 3 chiều: Spectrogram.?
- Phân tích đặc tính ngữ âm, âm học của tiếng nói?

**Câu 3: (4 điểm)**

- Trình bày hệ thống chuyển đổi văn bản thành giọng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>6</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền không gian 3 chiều: Spectrogram.?
- Trình bày về Formant và Antiformant?

**Câu 3: (4 điểm)**

- Trình bày hệ thống chuyển đổi văn bản thành giọng nói?
- Trình bày cấu trúc của mô hình Markov? Các vấn đề trong mô hình Markov?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>7</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

**Câu 2: (3 điểm)**

- Trình bày phương pháp biểu diễn tín hiệu tiếng nói trong miền không gian 3 chiều: Spectrogram.?
- Phân tích đặc tính ngữ âm, âm học của tiếng nói?

**Câu 3: (4 điểm)**

- Trình bày hệ thống chuyển đổi văn bản thành giọng nói?
- Mô hình markov? Ứng dụng của Markov trong nhận dạng tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

Trường Đại Học Hàng Hải Việt Nam  
Khoa Công nghệ Thông tin  
BỘ MÔN HỆ THỐNG THÔNG TIN  
-----\*\*\*-----

## ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>2009- 2010</b>	Đề thi số:  <b>8</b>	Ký duyệt đề:
Thời gian: <b>60 phút</b>		

**Câu 1: (3 điểm)**

- Trình bày khái niệm chung và các ứng dụng của xử lý tiếng nói?
- Phân biệt 2 hệ thống: nhận dạng tiếng nói và tổng hợp tiếng nói?
- Các tính chất có thể thay đổi được trong tín hiệu tiếng nói?

**Câu 2: (3 điểm)**

- Phân tích đặc tính ngữ âm, âm học của tiếng nói?
- Mã hóa tiếng nói dạng sóng vô hướng: xung tuyến tính?

**Câu 3: (4 điểm)**

- Các phương pháp nhận dạng tiếng nói?
- Trình bày cấu trúc của mô hình Markov? Các vấn đề trong mô hình Markov?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

**Trường Đại Học Hàng Hải Việt Nam**  
**Khoa Công nghệ Thông tin**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**  
 -----\*\*\*-----

## THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>x</b>	Đề thi số:  <b>X</b>	Ký duyệt đề:  x
Thời gian: <b>60 phút</b>		

**Câu 1: (2 điểm)**

Âm tiết là gì? Trình bày đặc điểm và cấu trúc của âm tiết tiếng Việt.

**Câu 2: (2 điểm)**

Trình bày kiến trúc chung của hệ thống chuyển văn bản thành tiếng nói và chức năng của từng thành phần.

**Câu 3: (3 điểm)**

- Tìm biến đổi Fourier  $X(e^{j\omega})$  của dãy  $x(n) = n\alpha^n u(n-2)$  với  $|\alpha| < 1$
- Tìm biến đổi Fourier rời rạc N điểm  $X(k)$  của dãy  $x(n) = a^{|n|}$  với  $0 \leq n \leq N-1$ ;  $|a| < 1$

**Câu 4: (3 điểm)**

- Phổ của tín hiệu tiếng nói là gì? Các loại tần số được sử dụng khi vẽ đồ thị phổ?
- Ảnh phổ của tín hiệu tiếng nói là gì? Trình bày các bước thực hiện phân tích phổ tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi



**Trường Đại Học Hàng Hải Việt Nam**  
**Khoa Công nghệ Thông tin**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**  
 -----\*\*\*-----

## THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>x</b>	Đề thi số:  <b>X</b>	Ký duyệt đề:  x
Thời gian: <b>60 phút</b>		

**Câu 1: (2 điểm)**

Trình bày hiểu biết của bạn về đặc điểm âm học của các loại nguyên âm, phụ âm. Lấy ví dụ.

**Câu 2: (2 điểm)**

Trình bày giải pháp tổng hợp tiếng nói tiếng Việt bằng cách ghép các âm vị kép (diphone).

**Câu 3: (3 điểm)**

- Tìm biến đổi Fourier  $X(e^{j\omega})$  của dãy  $x(n) = n\alpha^n u(-n+2)$  với  $|\alpha| > 1$
- Tìm biến đổi Fourier rời rạc N điểm  $X(k)$  của dãy  $x(n)$ ,

**Câu 4: (3 điểm)**

- Phổ của tín hiệu tiếng nói là gì? Các loại tần số được sử dụng khi vẽ đồ thị phổ?
- Ảnh phổ của tín hiệu tiếng nói là gì? Trình bày các bước thực hiện phân tích phổ tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

**Trường Đại Học Hàng Hải Việt Nam**  
**Khoa Công nghệ Thông tin**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**  
 -----\*\*\*-----

## THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>x</b>	Đề thi số:  <b>X</b>	Ký duyệt đề:  x
Thời gian: <b>60 phút</b>		

**Câu 1: (2 điểm)**

Trình bày quá trình chuẩn hóa văn bản trong hệ thống chuyển văn bản thành tiếng nói.

**Câu 2: (2 điểm)**

Trình bày giải pháp tổng hợp tiếng nói tiếng Việt bằng cách ghép phụ âm đầu và phần vần.

**Câu 3: (3 điểm)**

- Tìm biến đổi Fourier  $X(e^{j\omega})$  của dãy  $x(n) = \alpha^n u(-n-2)$  với  $|\alpha| > 1$
- Tìm biến đổi Fourier rời rạc N điểm  $X(k)$  của dãy  $x(n) = e^{j(2\pi/N)k}$  với  $0 \leq n \leq N-1$

**Câu 4: (3 điểm)**

- Phổ của tín hiệu tiếng nói là gì? Các loại tần số được sử dụng khi vẽ đồ thị phổ?
- Ảnh phổ của tín hiệu tiếng nói là gì? Trình bày các bước thực hiện phân tích phổ tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

**Trường Đại Học Hàng Hải Việt Nam**  
**Khoa Công nghệ Thông tin**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**  
 -----\*\*\*-----

## THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>x</b>	Đề thi số:  <b>X</b>	Ký duyệt đề:  x
Thời gian: <b>60 phút</b>		

**Câu 1: (2 điểm)**

Trình bày vấn đề tạo ngữ điệu khi tổng hợp tiếng nói.

**Câu 2: (2 điểm)**

Trình bày các vấn đề gặp phải khi tổng hợp tiếng nói bằng cách ghép nối các đơn vị âm.

**Câu 3: (3 điểm)**

a) Tìm biến đổi Fourier  $X(e^{j\omega})$  của dãy  $x(n)$ :

b) Tìm biến đổi Fourier rời rạc N điểm  $X(k)$  của dãy  $x(n)$ :

**Câu 4: (3 điểm)**

a) Phổ của tín hiệu tiếng nói là gì? Các loại tần số được sử dụng khi vẽ đồ thị phổ?

b) Ảnh phổ của tín hiệu tiếng nói là gì? Trình bày các bước thực hiện phân tích phổ tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi

**Trường Đại Học Hàng Hải Việt Nam**  
**Khoa Công nghệ Thông tin**  
**BỘ MÔN HỆ THỐNG THÔNG TIN**  
 -----\*\*\*-----

## THI KẾT THÚC HỌC PHẦN

Tên học phần: <b>XỬ LÝ TIẾNG NÓI</b> Năm học: <b>x</b>	Đề thi số:  <b>X</b>	Ký duyệt đề:  x
Thời gian: <b>60 phút</b>		

**Câu 1: (2 điểm)**

Trình bày hiểu biết của bạn về đặc điểm âm học của các loại nguyên âm, phụ âm. Lấy ví dụ.

**Câu 2: (2 điểm)**

Trình bày về nhận dạng tiếng nói bằng phương pháp dựa vào nhận dạng mẫu.

**Câu 3: (3 điểm)**

a) Tìm biến đổi Fourier  $X(e^{j\omega})$  của dãy  $x(n) = (n+1)\alpha^n u(n)$  với  $|\alpha| < 1$

b) Tìm biến đổi Fourier rời rạc N điểm  $X(k)$  của dãy  $x(n) = \sin(\frac{2\pi}{N} k_0 n)$  với  $0 \leq n \leq N-1$

**Câu 4: (3 điểm)**

a) Phổ của tín hiệu tiếng nói là gì? Các loại tần số được sử dụng khi vẽ đồ thị phổ?

b) Ảnh phổ của tín hiệu tiếng nói là gì? Trình bày các bước thực hiện phân tích phổ tín hiệu tiếng nói?

-----\*\*\*HẾT\*\*\*-----

Lưu ý: - Không sửa, xóa đề thi, nộp lại đề sau khi thi