HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Hữu Đam

NGHIÊN CỨU VỀ NHẬN DẠNG ÂM THANH VÀ ỨNG DỤNG TRONG CHUYỂN ĐỔI ÂM THOẠI SANG VĂN BẢN

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS.Nguyễn Đình Hóa

Phản biện 1: .....

Phản biện 2: .....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Nhận dạng tiếng nói của con người đã và đang thu hút sự quan tâm nghiên cứu của nhiều nhà khoa học khi mà công nghệ tự động hóa ngày càng có nhiều ứng dụng trong thực tiễn cuộc sống. Nghiên cứu nhận dạng tiếng nói Việt cũng được quan tâm nghiên cứu nhiều trong những năm gần đây, tuy vậy cho đến nay các kết quả vẫn chưa thỏa mãn những bài toán đặt ra từ thực tế cuộc sống do tính chất phức tạp về ngữ âm của tiếng Việt.

Hiện nay trên thế giới các công nghệ xử lý tiếng nói đã phát triển, các hệ thống ứng dụng xử lý tiếng nói đã được sử dụng ở nhiều nơi, độ chính xác của các hệ thống này ngày càng được cải thiện. Các ứng dụng của lĩnh vực xử lý tiếng nói rất phổ biến: nhận dạng tiếng nói, tổng hợp tiếng nói, xác thực người nói qua giọng nói và các thành tựu của chúng được áp dụng vào nhiều lĩnh vực trong thực tế.

Luận văn tập trung nghiên cứu các kỹ thuật nhận dạng tiếng nói, từ đó xây dựng ứng dụng nhận dạng một số từ, các số và cụ thể là nhận dạng âm thanh và ứng dụng trong chuyển đổi âm thoại sang văn bản sử dụng mô hình Markov ẩn dựa trên các đặc trưng MFCC. Ngoài ra, một số kỹ thuật khử nhiễu dữ liệu như CMS cũng được tích hợp để tăng tính hiệu quả của hệ thống. Các kỹ thuật nhận dạng giọng nói trong luận văn tập trung vào loại dữ liệu âm thanh tiếng Việt.

Cấu trúc của luận văn được trình bày trong ba chương gồm các nội dung chính như sau.

Chương 1 nghiên cứu và trình bày tổng quan về các đặc trưng âm thanh cần thiết cho quá trình nhận dạng từ vựng từ âm thoại. Trong chương này, một số phương pháp loại bỏ những thông tin không quan trọng, chẳng hạn như tiếng ồn của môi trường thu âm, nhiễu trên đường truyền, các đặc điểm riêng biệt của từng người nói,... cũng được mô tả sơ lược. Ngoài ra, nội dung chương cũng bao gồm các mô hình ngôn ngữ, các phương pháp hiện thời về nhận dạng tiếng nói, các đặc tính, cấu trúc cũng như khả năng biểu hiện ý nghĩa của tiếng Việt. Các nội dung nghiên cứu về âm vị tiếng Việt, thanh điệu, âm đầu, âm đệm, âm chính và âm cuối, và sự phân bổ của các âm vị trong tiếng Việt cũng được trình bày tại chương này.

Chương 2 này tập trung trình bày cơ sở lý thuyết của các thuật toán trong khâu tiền xử lý tiếng nói bao gồm: giải thuật phát hiện tiếng nói, các phương pháp tính hệ số và trích chọn đặc trưng MFCC và PLP, các kỹ thuật khử nhiễu như CMS và RASTA. Nội dung

chương đi sâu vào nghiên cứu và phân tích quá trình Markov sau đó sẽ đưa ra mô hình Markov ẩn và các trạng thái của mô hình Markov ẩn, đưa ra các bài toán cơ bản và và các giải pháp toán học cho các bài toàn cơ bản của mô hình Markov ẩn. Một số mô hình Markov ẩn khác nhau cũng được đi sâu nghiên cứu nhằm tìm kiếm khả năng mở rộng và nâng cao hiệu quả của hệ thống.

Chương 3 tập trung trình bày các kết quả thực nghiệm của hệ thống nhận dạng tiếng nói trong tiếng Việt và chuyển đổi âm thoại tiếng Việt sang văn bản. Nội dung chương được mở đầu bằng việc mô tả bộ cơ sở dữ liệu chuỗi tiếng Việt, từ đó trình bày quá trình huấn luyện hệ thống nhận dạng từ vựng, và cuối cùng là xây dựng chương trình nhận dạng từ vựng tiếng Việt và chuyển đổi âm thoại sang văn bản.

Chương 1 - TỔNG QUAN VỀ NHẬN DẠNG TIẾNG NÓI

1.1. Lý thuyết âm thanh và tiếng nói

1.1.1. Nguồn gốc âm thanh

Âm thanh là do vật thể dao động cơ học mà phát ra. Âm thanh phát ra dưới dạng sóng âm. Sóng âm là sự biến đổi các tính chất của môi trường đàn hồi khi năng lượng âm truyền qua. Âm thanh truyền được đến tai người là do môi trường dẫn âm. Sóng âm có thể truyền được trong chất rắn, chất lỏng, không khí. Có chất dẫn âm rất kém gọi là chất hút âm như: len, da, chất xốp…

1.1.2. Các đại lượng đặc trưng của dữ liệu âm thanh

1.1.2.1. Tần số của âm thanh

Là số lần dao động của phần tử khí trong một giây. Đơn vị là Hz, kí hiệu: f

1.1.2.2. Chu kì của âm thanh

Là thời gian mà âm thanh đó thực hiện một dao động hoàn toàn. Đơn vị là thời gian, kí hiệu là T.

1.1.2.3. Tốc độ truyền âm

Là tốc độ truyền năng lượng âm từ nguồn tới nơi thu. Đơn vị m/s. Tốc độ truyền âm trong không khí ở nhiệt độ từ 0- 20^0 C thường là 331 – 340 m/s.

1.1.2.4. Cường độ âm thanh

Là năng lượng được sóng âm truyền trong một đơn vị thời gian qua một đơn vị diện tích đặt vuông góc với phương truyền âm.

1.1.2.5. Thanh áp

Là lực tác dụng vào tai người nghe hoặc tại một điểm nào đó của trường âm thanh. Đơn vị : 1pa=1 N/m2 hoặc 1bar = 1dyn/cm2.

1.1.2.6. Âm sắc

Trong thành phần của âm thanh, ngoài tần số cơ bản còn có các sóng hài, số lượng sóng hài biểu diễn sắc thái của âm. Âm sắc là một đặc tính của âm nhờ đó mà ta phân biệt được tiếng

trầm, bổng khác nhau, phân biệt được tiếng nhạc cụ, tiếng nam nữ, tiếng người này với người khác.

1.1.2.7. Âm lượng

Là mức độ to nhỏ của nguồn. Đơn vị là W

1.1.3. Các tần số của âm thanh

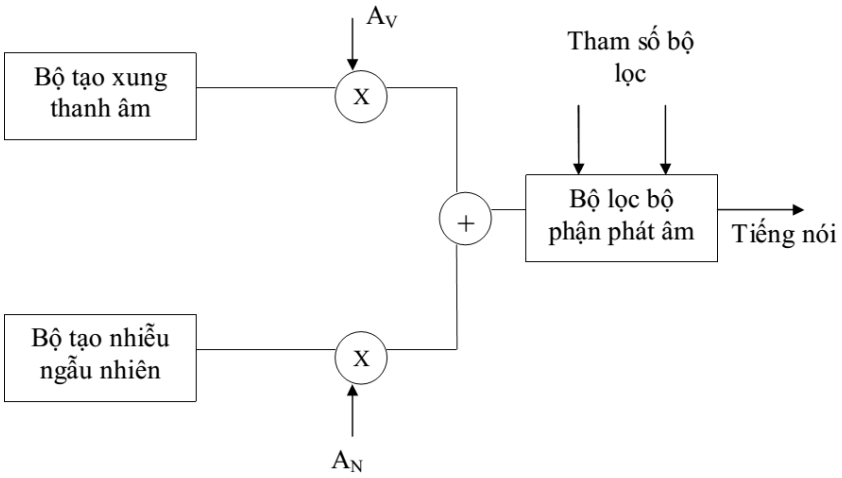
Theo [7], tần số cơ bản F0 là tần số giao động của dây thanh. Tần số này phụ thuộc vào giới tính và độ tuổi. F0 của nữ thường cao hơn của nam, F0 của người trẻ thường cao hơn của người già. Thường với giọng của nam, F0 nằm trong khoảng từ 80-250Hz, với giọng của nữ, F0 trong khoảng 150-500Hz. Sự biến đổi của F0 có tính quyết định đến thanh điệu của từ cũng như ngữ điệu của câu.

1.1.4. Cơ chế tạo lập tiếng nói của con người

Các cơ quan phát âm của con người chủ yếu gồm phổi, khí quản, thanh quản, bộ phận mũi và miệng.

1.1.5. Mô hình lọc nguồn tạo tiếng nói

Quá trình tạo tiếng nói là bộ lọc nguồn, trong đó tín hiệu từ nguồn âm thanh (cũng có thể là có chu kì hay nhiễu) được lọc bằng bộ lọc biến thiên theo thời gian có tính chất cộng hưởng tương tự với bộ phận phát âm.



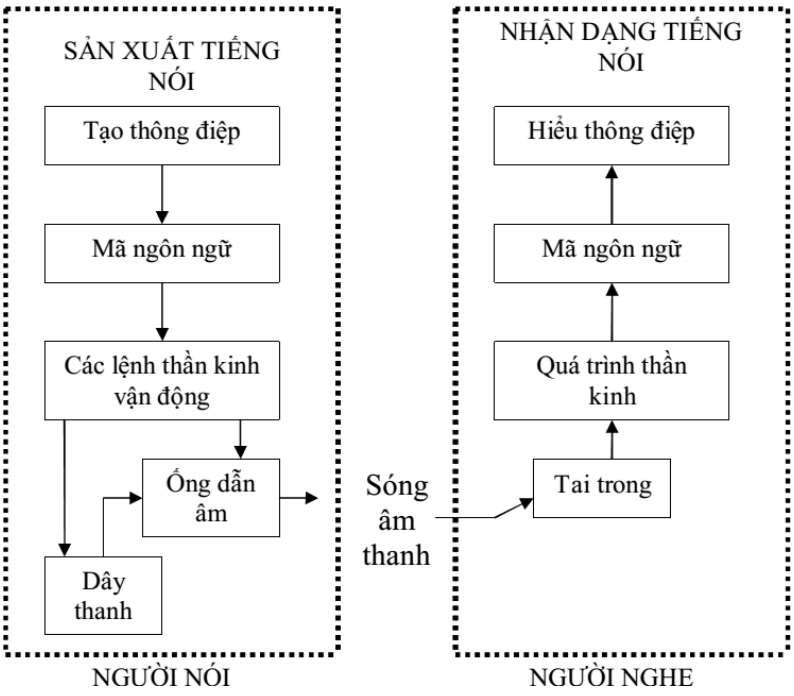
Hình 1-1: Mô hình lọc nguồn tạo tiếng nói [2]

1.1.6. Hệ thống thính giác của người

Quá trình nghe của người như sau: Sóng áp suất âm thanh tác động đến tai người, sóng này được chuyển thành chuỗi xung điện, chuỗi này được truyền tới não bộ thông qua hệ thần kinh, ở não chuỗi được xử lý và giải mã.

1.1.7. Quá trình tạo và thu nhận tiếng nói

Sơ đồ biểu diễn quá trình thu nhận tiếng nói của con người



Hình 1-2: Quá trình tạo và thu nhận tiếng nói [2]

1.1.8. Mô hình lọc nguồn tạo tiếng nói

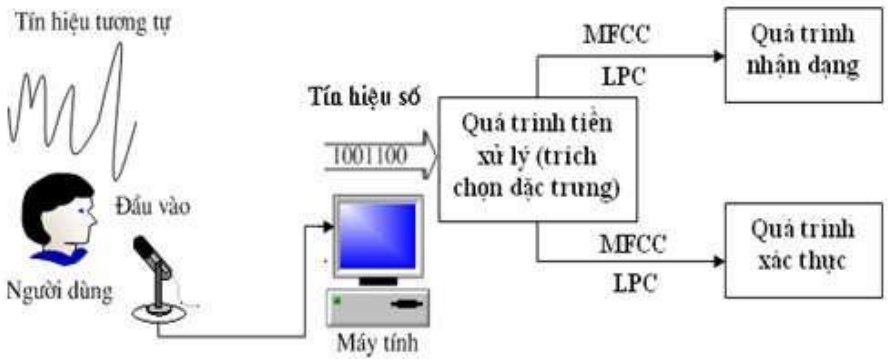
1.1.8.1. Nguyên âm

1.1.8.2. Các âm vị khác

1.2. Giới thiệu về xử lý tiếng nói

1.2.1. Mục đích của xử lý tiếng nói

Chúng ta có thể mô hình hóa cho bài toán xử lý tiếng nói như sau:



Hình 1-3: Mô hình bài toán xử lý tiếng nói [2]

1.3. Nhận dạng tiếng nói

1.3.1. Bài toán nhận dạng tiếng nói

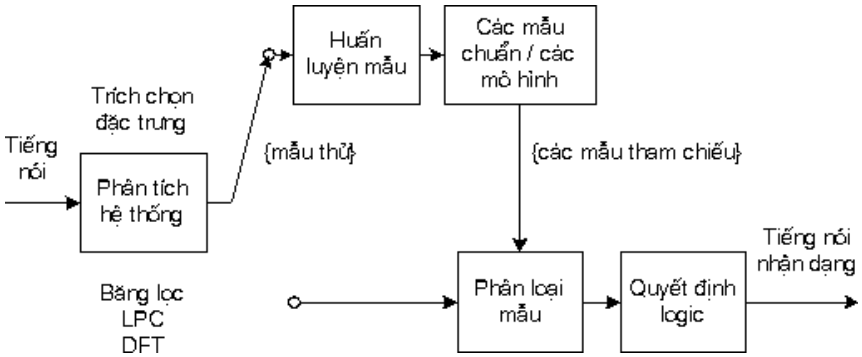
- Nhận dạng các từ phát âm rời rạc/liên tục
- Nhận dạng tiếng nói độc lập/phụ thuộc người nói
- Nhận dạng với từ điển cỡ nhỏ/vừa/lớn
- Nhận dạng trong môi trường nhiễu cao/thấp

1.3.2. Các phương pháp nhận dạng tiếng nói

a. Phương pháp âm học ngữ âm học

Hướng tiếp cận âm học và ngữ âm học dựa trên lý thuyết về âm học-ngữ âm học.

b. Phương pháp nhận dạng mẫu

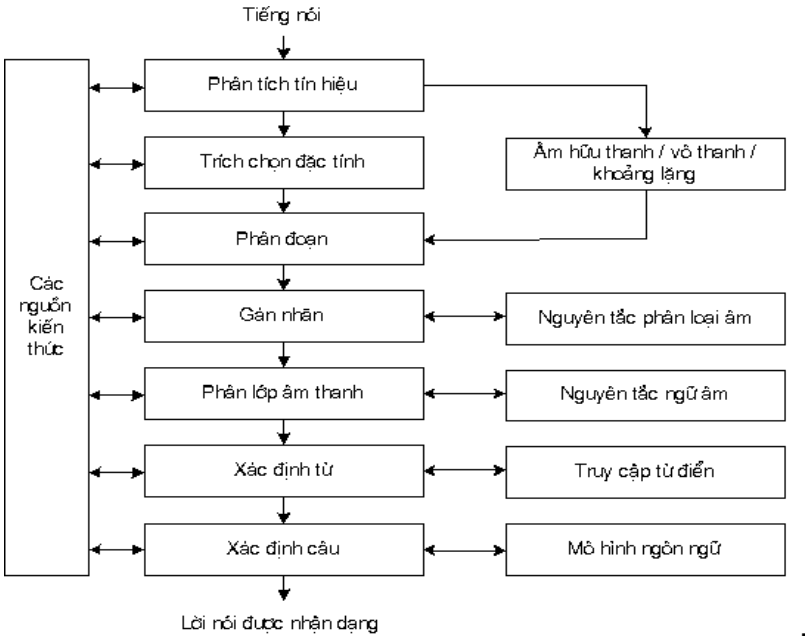


Hình 1-4: Hệ thống nhận dạng tiếng nói theo phương pháp nhận dạng mẫu [2]

Những bước cần thực hiện đối với một hệ thống nhận dạng mẫu là:

- ✓ **Trích chọn các đặc trưng**:
- ✓ **Huấn luyện mẫu**:
- ✓ **Phân lớp mẫu**:
- ✓ **Quyết định logic**:

Sơ đồ khối của phương pháp này như sau:



Hình 1-5: Tích hợp tri thức trong nhận dạng tiếng nói [2]

1.4. Nhận dạng tiếng Việt

1.4.1. Đặc điểm âm tiết tiếng Việt

1.4.1.1. Tính độc lập cao

Trong tiếng Việt, âm tiết được thể hiện khá đầy đủ, rõ ràng, được tách và ngắt thành từng khúc đoạn riêng biệt. Âm tiết nào của tiếng Việt cũng mang một thanh điệu và cấu trúc ổn định.

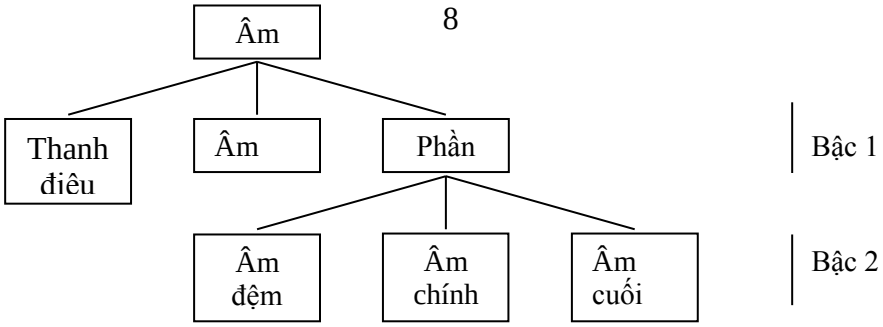
1.4.1.2. Khả năng biểu hiện ý nghĩa

Tuyệt đại đa số các âm tiết tiếng Việt đều có nghĩa. Gần như toàn bộ các âm tiết đều hoạt động như từ.

1.4.1.3. Cấu trúc chặt chẽ

<table>
<tr><td colspan="4">Thanh điệu</td></tr>
<tr><td rowspan="2">Âm đầu</td><td colspan="3">Vần</td></tr>
<tr><td>Âm đệm</td><td>Âm chính</td><td>Âm cuối</td></tr>
</table>

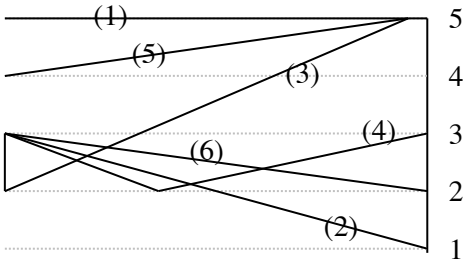
Hình 1-6: Cấu trúc của âm tiết tiếng Việt [6]



Hình 1-7: Cấu trúc hai bậc của âm tiết tiếng Việt [6]

1.4.2. Âm vị tiếng Việt

1.4.2.1. Thanh điệu



Hình 1-8: Các thanh điệu tiếng Việt 1. Không dấu, 2. Huyền, 3. Ngã, 4. Hỏi, 5. Sắc, 6.Nặng [6]

1.4.2.2. Âm đầu

1.4.2.3. Âm đệm

Âm đệm có chức năng tu chỉnh âm sắc của âm tiết lúc khởi đầu, làm trầm hoá âm tiết và khu biệt âm tiết này với âm tiết khác.

1.4.2.4. Âm chính

Âm chính là nguyên âm và có mặt trong mọi âm tiết qui định âm sắc của âm tiết. Âm chính tiếng Việt có tất cả 14 âm gồm 11 nguyên âm đơn và 3 nguyên âm đôi.

1.4.2.5. Âm cuối

Các âm cuối tiếng Việt có đặc điểm giống nhau là không buông (bộ phận cấu âm tiến đến vị trí cấu âm rồi giữ nguyên vị trí đó chứ không về vị trí cũ).

Âm chính	Âm phụ		Bán nguyên âm cuối		
	/u̯/	Ví dụ	/u̯/	/i̯/	Ví dụ
i	+	uy	+	-	iu
e	+	uê	+	-	êu
ɛ	+	oe	+	-	eo
i*h*e	+	uyên	+	-	yêu
u	-	ui	-	+	ui
o	-	ôi	-	+	ôi
ɔ	-	oi	-	+	oi
u*h*o	-	uôi	-	+	uôi.
ɯ	-	-	+	+	ưu, ưi
ɤ	+	quơ	-	+	-, ơi
ɤ̯	+	uân	+	+	âu,ay
a	+	oa	+	+	ao,ai
ă	+	ăn	+	+	au,ay
ɯ*h* ɤ	-	-	+	+	ưu,ươi

Hình 1-9: Phân bố giữa nguyên âm âm chính và các âm đệm và bán nguyên âm cuối [6]

1.4.3. Sự phân bố của các âm vị tiếng Việt

Các âm tiết tiếng Việt có cấu trúc chặt chẽ và các âm vị trong tiếng Việt kết hợp với nhau theo những qui luật. Hình 1-9 tổng kết sự phân bố giữa nguyên âm âm chính và các âm đệm và bán nguyên âm cuối [5].

1.4.4. Một số đặc điểm ngữ âm tiếng Việt

Theo [1], đặc điểm dễ thấy là tiếng Việt là ngôn ngữ đơn âm (monosyllable - mỗi từ đơn chỉ có một âm tiết), không biến hình (cách đọc, cách ghi âm không thay đổi trong bất

cứ tình huống ngữ pháp nào). Tiếng Việt hoàn toàn khác với các ngôn ngữ Ấn-Âu như tiếng Anh, tiếng Pháp là các ngôn ngữ đa âm, biến hình.

1.4.5. Những thuận lợi và khó khăn đối với nhận dạng tiếng Việt

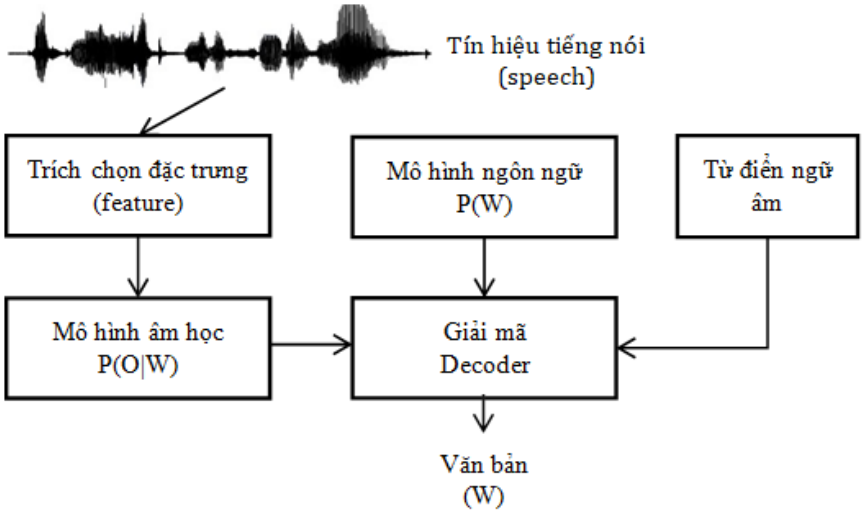
1.4.5.1. Thuận lợi

1.4.5.2. Khó khăn

Chương 2 - CÁC KỸ THUẬT NHẬN DẠNG TỪ VỰNG TRONG ÂM THOẠI TIẾNG VIỆT

Hiện nay có rất nhiều phương pháp nhận dạng tiếng nói. Mô hình Fujisaki được ứng dụng rộng rãi trong hệ thống của tiếng Nhật, mô hình MFGI được ứng dụng trong tiếng Đức, mô hình HMM (Hidden Markov Models), mô hình sử dụng mạng nơron,… Trong khuôn khổ Luận văn này tác giả lựa chọn mô hình HMM (Hidden Markov Models) để huấn luyện và nhận dạng tiếng nói. Mô hình Markov ẩn (HMM) là một mô hình thống kê, thích hợp ứng dụng trong việc nhận dạng mẫu: tiếng nói, hình ảnh và chữ viết…HMM được ứng dụng rộng rãi trong những năm gần đây vì hai lý do.Thứ nhất, mô hình có độ chính xác cao trong nhiều ứng dụng; Thứ hai,cấu trúc mô hình có thể thay đổi dễ dàng cho phù hợp với từng ứng dụng cụ thể.

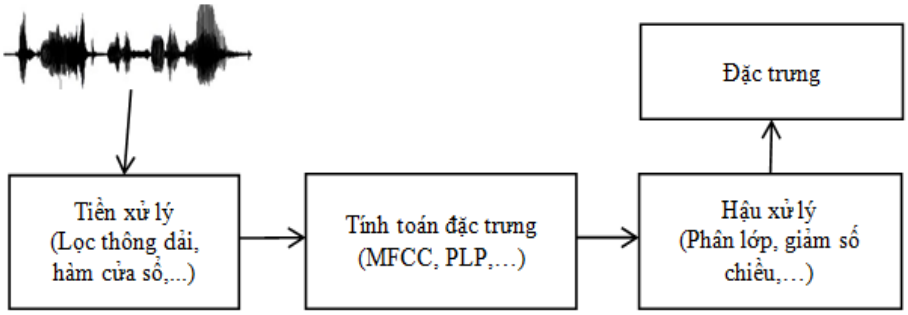
2.1. Các thành phần chính của một hệ thống nhận dạng tiếng nói



Hình 2-1: Sơ đồ khối tổng quan của một hệ thống nhận dạng tiếng nói [4]

Cấu trúc tổng quát của một hệ thống nhận dạng tiếng nói được mô tả ở hình 2-1

2.1.1. Trích chọn đặc trưng



Hình 2-2: Sơ đồ các bước trích chọn đặc trưng [4]

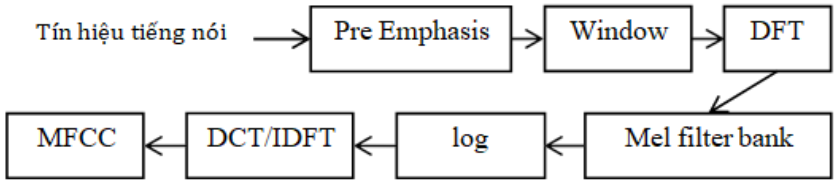
Khâu này có thể chia ra làm ba giai đoạn gồm tiền xử lý, tính toán đặc trưng và hậu xử lý như mô tả ở hình 2-2.

a. Khâu tiền xử lý:

b. Khâu tính toán đặc trưng:

c. Khâu hậu xử lý:

2.1.1.1. Đặc trưng MFCC

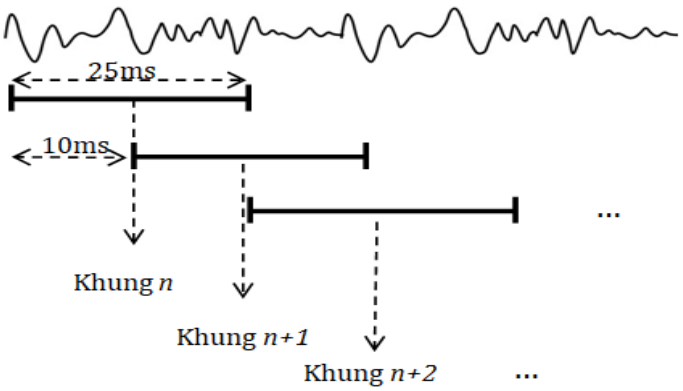


Hình 2-3: Sơ đồ khối các bước tính toán MFCC [4]

Về cơ bản, phương pháp trích chọn đặc trưng MFCC có các công đoạn chính như sau.

- **Pre Emphasis**
- **Window**

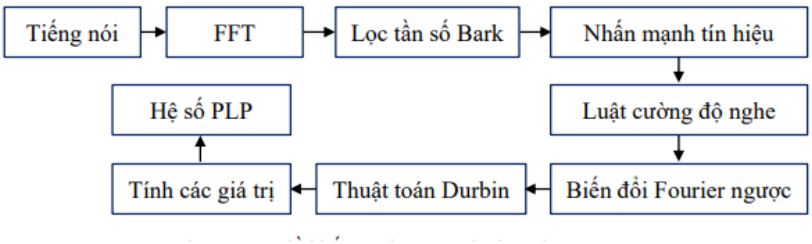
- **DFT**



Hình 2-4: Tạo khung trên tín hiệu tiếng nói [4]

Trong đó: *L* là kích thước của cửa sổ, *w[n]* giá trị của tín hiệu đầu vào sau khi qua hàm cửa sổ.

- **Mel Filter bank**
- **Logarithm (log) và biến đổi Cosine rời rạc (DCT)**



Hình 2-5: Sơ đồ khối các bước tính toán PLP [4]

Trong đó:

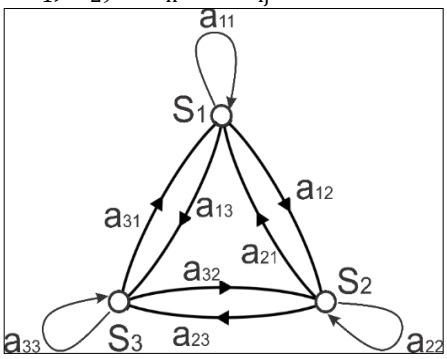
- **Biến đổi Fourier nhanh (FFT)**
- **Lọc theo thang tần số Bark**
- **Nhấn mạnh tín hiệu**
- **Dùng luật cường độ nghe (Power Law of Hearing**
- **Biến đổi Fourier ngược (Inverse DFT)**
- **LPC**
- **Thuật toán Durbin**
- **Tính các giá trị delta**

2.1.2. Kỹ thuật khử nhiễu CMS

2.2. Tổng quan về mô hình Markov ẩn HMM

2.2.1. Chuỗi Markov

Là dãy gồm N trạng thái S_1, S_2,…S_n với a_{ij} là xác xuất chuyển tiếp trạng thái từ S_i đến S_j.



Hình 2-6: Chuỗi Markov với 3 trạng thái *S*1, *S*2, *S*3 với các xác suất chuyển tiếp tương ứng *a*11 đến *a*33 [4]

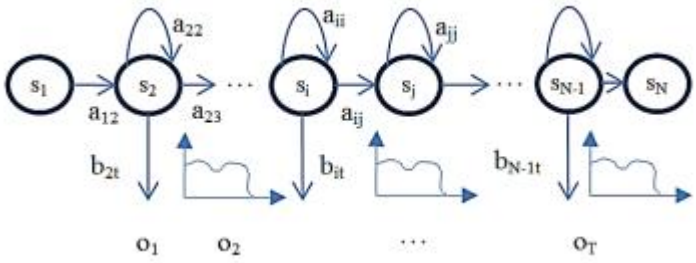
2.2.2. Mô hình Markov ẩn HMM

HMM là mô hình xác suất dựa trên lý thuyết về chuỗi Markov [Rabiner 1989] bao gồm các đặc trưng sau:

- $O = \{o_1, o_2, \dots, o_T\}$ là tập các vector quan sát.
- $S = \{s_1, s_2, \dots, s_N\}$ là tập hữu hạn các trạng thái s gồm N phần tử.
- $A = \{a_{11}, a_{12}, \dots, a_{nn}\}$ là ma trận hai chiều trong đó a_{ij} thể hiện xác suất để trạng thái s_i chuyển sang trạng thái s_j , với $a_{ij} \geq 0$ và $\sum a_{ij} = 1$, $\forall i\ j{=}k$
- $B = \{b_{2t}, b_{it}, \dots, b_{(N-1)t}\}$ là tập các hàm xác suất phát tán của các trạng thái từ s_2 đến s_{N-1}, trong đó b_{it}thể hiện xác suất để quan sát o_t thu được từ trạng thái s_i tại thời điểm t. Trong nhận dạng tiếng nói hàm b_{it} thường được sử dụng là hàm Gaussian với nhiều thành phần trộn (mixture) có dạng như công thức (2.12), trong trường hợp này ta gọi là mô hình kết hợp Hidden Markov Model và Gaussian Mixtrue Model (HMM-GMM)

$$b_i(o_t) = \sum_{k=1}^{M} c_{ik}\, N(o_t; \mu_{ik}, \Sigma_{ik}) \qquad (2.12)$$

  Trong đó: o_t là vector quan sát tại thời điểm t, M là số thành phần trộn của hàm Gaussian, c_{ik}, μ_{ik}, Σ_{ik} theo thứ tự là trọng số, vector trung bình và ma trận phương sai (covariance matrix) của thành phần trộn thứ k của trạng thái si .
- $\Pi = \{\pi_i\}$ là tập xác suất trạng thái đầu, với $\pi_i = P(q_1 = s_i)$ với i=1..N là xác suất để trạng thái s_i là trạng thái đầu q_1.



Hình 2-7: Mô hình HMM-GMM Left-Right với N trạng thái [4]

2.2.3. Các thành phần của HMM

Một HMM λ (N, M, A, B, Π) gồm 5 thành phần [3]:

a. N: Số trạng thái, với tập các trạng thái: S = (S1, S2, …, SN) và trạng thái quan sát được tại thời điểm t là qt.

b. M: Số hiện tượng quan sát được của mỗi trạng thái, ký hiệu hiện tượng quan sát được là V = {V1, V2, …,VM}, tín hiệu quan sát được ở thời điểm t là Ot.

c. Xác suất chuyển tiếp trạng thái biểu diễn bởi ma trận A = {aij} từ trạng thái Si đến Sj.

$$a_{ij} = P\left[q_{t+1} = S_j \mid q_t = S_i \right], 1 \le i, j \le N \quad (2.13)$$

$a_{ij} > 0 \forall i, j$ với điều kiện một trạng thái S_j có thể đến được từ mọi trạng thái Si và thỏa ràng buộc $\sum_{j=1}^{N} a_{ij} = 1$.

d. Phân bố xác suất (probability distribution) quan sát được tại trạng thái j:

B = {bj(k)}

$$bj(k) = P\left[v_k - t \; q_t = S_i \right], 1 \le j \le N \; 1 \le k \le M \quad (2.14)$$

thỏa ràng buộc $\sum_{k=1}^{M} b_j(k) = 1$

A và B là tham số quan trọng nhất trong mô hình HMM.

e. Phân bố xác suất trạng thái đầu tiên: $\Pi = \{\pi i\}$, với πi là trạng thái Si chọn.

$$\pi_i = P\left[q_1 = S_i \right], 1 \le i \le N \quad (2.15)$$

thỏa điều kiện $\sum_{i=1}^{N} \pi_i = 1$

Trong các thành phần trên, giá trị M và N được chọn đầu tiên và không thay đổi, chúng được sử dụng để tính 3 giá trị còn lại. Các bước tạo dữ liệu:

- Chọn trạng thái ban đầu với xác suất là π.
- Đặt t = 1
- Chọn $O_t = v_k$, với $B = \{b_j(k)\}$
- Chuyển sang một trạng thái mới, sử dụng ma trận $A = \{a_{ij}\}$
- Đặt t = t+1, quay lại bước ba nếu t < T.

Mô hình HMM được biểu diễn bởi bộ tham số: $\lambda = (A, B, \pi)$

Với chuỗi quan sát là: $O = O_1O_2\ldots O_T$

Trong đó: O_t: một hiện tượng của V; T: số trạng thái quan sát.

2.2.4. Hàm mật độ xác suất hỗn hợp Gauss

Hàm mật độ xác suất phân bố Gauss có dạng:

$$g_i(X \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)\right\} \quad (2.16)$$

các trọng số hỗn hợp cần thỏa điều kiện $\sum_{i=1}^{M} \pi_i = 1$

2.3. Ba bài toán cơ bản của mô hình Markov ẩn

Việc ứng dụng HMM trong nhận dạng tiếng nói dựa trên việc giải được ba bài toán cơ bản sau [1].

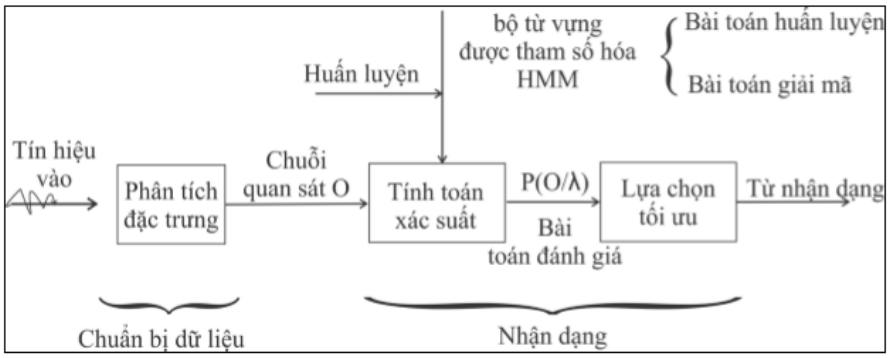
2.3.1. Bài toán đánh giá

2.3.2. Bài toán giải mã

2.3.3. Bài toán huấn luyện

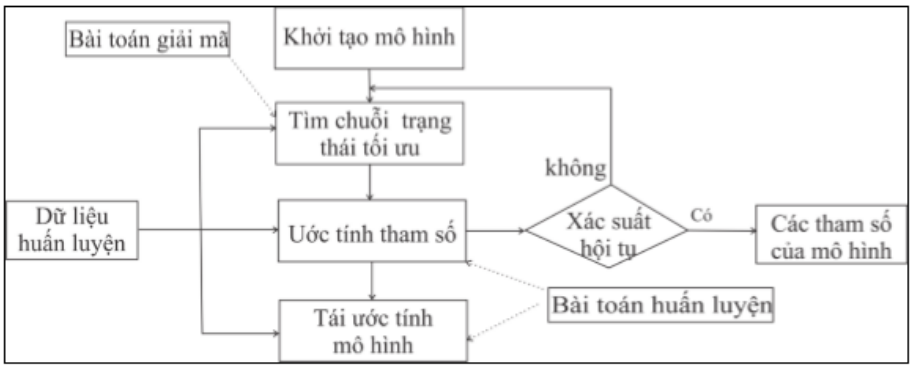
2.4. Ứng dụng của HMM trong nhận dạng tiếng nói rời rạc

2.4.1. Tổng quan



Hình 2-8: Ứng dụng các bài toán trong nhận dạng từ rời rạc [2]

2.4.2. Giai đoạn huấn luyện mô hình



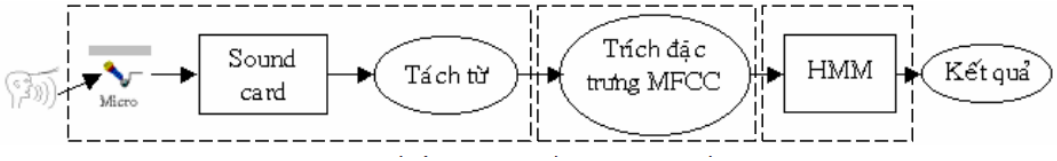
Hình 2-9: Các bước huấn luyện bằng HMM [2]

2.4.3. Giai đoạn nhận dạng

Tín hiệu tiếng nói cần nhận dạng được trích xuất vector đặc trưng, gọi là chuỗi quan sát O. Sau đó cần giải quyết bài toán đánh giá để tính V xác suất $P(O|\lambda_i)$ của V từ trong bộ từ vựng và chọn ra mô hình mô tả đúng nhất tín hiệu tiếng nói đưa vào, đó là mô hình λ_i có xác suất $P(O|\lambda_i)$ lớn nhất trong tập V mô hình, từ đó suy ra lệnh (từ đơn) ứng với tín hiệu đầu vào.

Chương 3 - XÂY DỰNG HỆ THỐNG CHUYỂN ĐỔI ÂM THOẠI TIẾNG VIỆT SANG VĂN BẢN

Sơ đồ tổng quát của hệ thống nhận dạng tiếng nói được thể hiện trên hình 3-1



Hình 3-1: Sơ đồ tổng quát của hệ thống nhận dạng và chuyển đổi [2]

Để thuận tiện cho việc nhận dạng và chuyển đổi hiển thị kết quả, trong giới hạn của luận văn này và từ sơ đồ trên tôi chia chương trình xây dựng hệ thống chuyển đổi thành ba quá trình riêng biệt:

- Thu thập và tiền xử lí tín hiệu tiếng nói
- Trích chọn đặc trưng MFCC
- Quá trình thứ ba

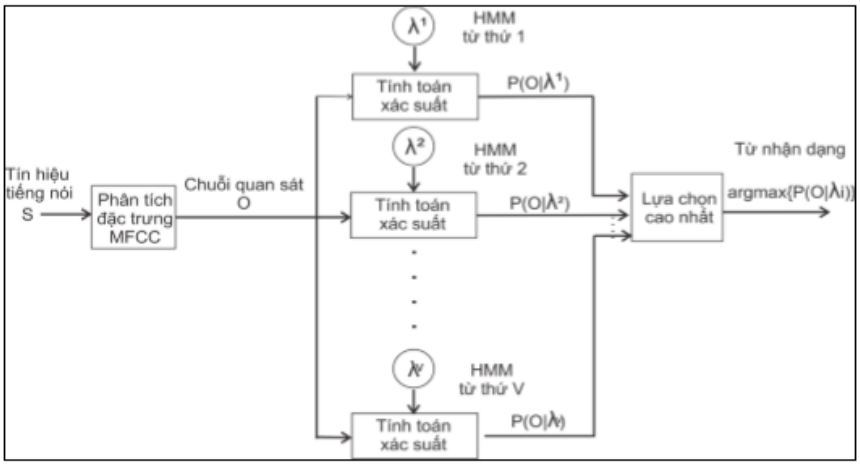
Chi tiết các quá trình trên như sau:

3.1. Thu thập và tiền xử lí tín hiệu tiếng nói

3.2. Trích chọn đặc trưng MFCC

3.3. Nhận dạng bằng mô hình HMM

Sơ đồ nhận dạng bằng mô hình HMM được thể hiện như hình 3.4.



Hình 3- 2: Tổng quan mô hình nhận dạng [2]

3.4. Xây dựng dữ liệu huấn luyện và kiểm thử hệ thống hiển thị kết quả.

Để tiếp tục tiến hành quá trình xây dựng hệ thống, ta cần chuẩn bị cơ sở dữ liệu huấn luyện để cài đặt, đánh giá hiệu suất hoạt động của hệ thống.

3.4.1 Thu âm dữ liệu

Dữ liệu thu âm được chia làm hai phần:

- Dùng để huấn luyện
- Dùng để kiểm thử hệ thống

3.4.2 Đặc tính file dữ liệu

Dữ liệu lưu theo định dạng chuẩn file *.wav của Microsoft, tần số lấy mẫu là 16 kHz, đơn kênh (mono), thời lượng mỗi file từ một đến hai giây, có bao gồm khoảng lặng (silence) ở đầu và cuối file.

3.4.3 Cấu hình hệ thống nhận dạng

3.4.4 Kết quả thực nghiệm

Kết quả đạt được với hệ thống có độ chính xác 77,29% ở mức từ và 13.51% ở mức câu, nhận thấy chất lượng nhận dạng ở mức câu còn thấp, nguyên nhân do dữ liệu giọng nói thu âm bằng điện thoại có lẫn nhiều tạp âm như tiếng ho, tiếng cười, "à, ờ"… đối với máy tính trường hợp như vậy gây ra những khó khăn đặc biệt trong nhận dạng tiếng nói.

- Thử nghiệm với nhiều hàm Gaussian

  Kết quả đạt được với hệ thống có độ chính xác 78.23% ở mức từ và 14.86% ở mức câu, so với 77,29% ở mức từ và 13.51% ở mức câu ở hệ thống sử dụng một hàm Gaussian.

- Thử nghiệm với dữ liệu kiểm tra và dữ liệu huấn luyện trùng nhau

Kết quả nhận dạng của hệ thống đã được cải thiện rõ ràng với có độ chính xác 87.70% ở mức từ và 20.27% ở mức câu.

KẾT LUẬN VÀ KIẾN NGHỊ

Với kết quả kiểm tra độ chính xác nhận dạng như trên thì có thể thấy rằng việc áp dụng mô hình Markov ẩn trong nhận dạng tiếng Việt đã cho kết quả khá tốt. Tuy chưa thật sự hoàn hảo nhưng những kết quả thu được tương đối khả quan.Tuy vẫn còn một số hạn chế như

- Dữ liệu huấn luyện chưa đầy đủ, số từ đem huấn luyện chưa nhiều, chưa thu được từ nhiều người, nhiều nơi; môi trường thu âm còn nhiều nền nhiễu (tiếng ồn),…
- Một số thông số có ảnh hưởng đến độ chính xác nhận dạng như: hàm khởi tạo, số nút ẩn, giá trị kích hoạt trọng số,… có thể được lựa chọn chưa tối ưu.

Các nguyên nhân trên muốn khắc phục được đều cần phải có thời gian, và cần phải bỏ công sức nghiên cứu nhiều hơn nữa. Để hệ thống có thể được ứng dụng rộng rãi hơn cần phải cải tiến và mở rộng thêm. Với thiết kế đã được đưa ra thì hướng phát triển tiếp của tác giả có thể là:

- Tăng số lượng từ trong từ điển nhận dạng.
- Có thể vừa thu âm, vừa nhận dạng (không phải chờ đến khi thu âm xong mới nhận dạng).
- Nhận dạng câu (có khả năng phán đoán được từ gần đúng).

Do thời gian làm Luận văn không có nhiều nên tác giả chưa có điều kiện để tìm hiểu hết những hướng tiếp cận mới trong nhận dạng tiếng nói. Hi vọng rằng trong thời gian tới tác giả Luận văn có thể hoàn thiện hơn nữa các nội dung đã đề ra.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Vũ Kim Bảng, Triệu Thị Thu Hương, Bùi Đăng Bình (2001). "Âm tiết tiếng Việt khả năng hình thành và thực tế ứng dụng", *Toàn văn Báo cáo Khoa học, Hội nghị kỷ niệm 25 năm thành lập Viện Công nghệ Thông tin,* tr 525-533.

[2] Ngô Văn Cương: "Nghiên cứu kỹ thuật nhận dạng tiếng nói tiếng Việt và ứng dụng" – Luận văn Thạc sỹ.

[3] Võ Xuân Hào, ĐH Quy Nhơn - 2009: "Giáo trình ngữ âm tiếng Việt hiện đại".

[4] Nguyễn Văn Huy: "Nghiên cứu mô hình thanh điệu trong nhận dạng tiếng Việt từ vựng lớn phát âm liên tục".

[5] Đỗ Xuân Tho (1997), Lê Hữu Tỉnh, *Giáo trình tiếng Việt 2*, Nhà xuất bản Giáo dục.

[6] Đoàn Thiện Thuật (1999), Ngữ âm Tiếng Việt, Nhà xuất bản Đại học Quốc gia Hà nội.

[7] Phạm Văn Sự, Lê Xuân Thành – Học viện Công nghệ bưu chính viễn thông: "Bài giảng xử lý tiếng nói" – 2010.

Tiếng Anh

[8] Ling Feng. "Speeech Recognition", Technical University of Denmark Informatics and Mathematical Modelling, Kgs. Lyngby, 2004.

[9] Prashanth Kannadaguli,Vidya Bhat. "A Comparison of Gaussi an Mixture Modeling (GMM) and Hidden Markov Modeling (HMM) based approaches for Automatic Phoneme Recognition in Kannada", Department of Electronics and Communication Engineering Manipal Institute of Technology, Manipal, India, 2015.

[10] Mariano Marufo da Silva, "Diego A. Evin, Sebastián Verrastro. "Speaker-independent embedded speech recognition using Hidden Markov Models", 978-1-5090-2938-©2016 IEEE, 2016.

[11] Devi Handaya, Hanif Fakhruroja, Egi Muhammad Idris Hidayat, Carmadi Machbub. "Comparison of Indonesian Speaker Recognition Using Véc tơ Quantization and Hidden Markov Model for Unclear Pronunciation Problem", 2016 IEEE 6th International Conference on System Engineering and Technology (ICSET), Oktober 3-4, 2016 Bandung – Indonesia, 2016.

[12] Rabiner L., Juang B.H. (1993). Fundamentals of Speech Recognition. Prentice Hall, ISBN 0-13-01517-2.

[13] Hermansky, H. and Daniel, P.W. Ellis and Sangita, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." Acoustics, Speech, and Signal Processing (ICASSP). Istanbul: IEEE, 2000. 1635-1638.

[14] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech." Acoustical Society of America Journal, 1990: 1738–1752

[15] Levinson, N. "The Wiener RMS error criterion in filter design and prediction." J. Math. Physics, 1947: 261–278.

[16] Jurafsky, Daniel and Martin, James H. Speech and Language Processing - 2nd Edition. Prentice Hall, ISBN-13: 978-0131873216, ISBN-10: 0131873210, 2008.

[17] Rabiner, L. and Juang, B. "An introdution to Hidden Markov Models." IEEE, V.77, No.2, 1989: 257-286.

[18] Young, Steve. *The HTK Book.* UK: Cambridge University Engineering Department, 2009.