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

LUẬN VĂN THẠC SĨ KỸ THUẬT
(*Theo định hướng ứng dụng*)

HÀ NỘI - 2020

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

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS.NGUYỄN ĐÌNH HÓA

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả luận văn

Nguyễn Hữu Đam

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn sâu sắc nhất đến người hướng dẫn khoa học TS. Nguyễn Đình Hóa, cảm ơn Thầy trong thời gian qua mặc dù công việc rất bận rộn nhưng đã dành cho tôi sự giúp đỡ và hướng dẫn tận tình, những kiến thức quý báu Thầy truyền đạt đã giúp tôi vượt qua những khó khăn để hoàn thành Luận văn này.

Tôi xin chân thành cảm ơn các Thầy cô giảng viên trong khoa Công nghệ thông tin và Sau Đại Học của Học Viện Công Nghệ Bưu Chính Viễn Thông đã tận tình giảng dạy và hướng dẫn trong suốt quá trình học tập và nghiên cứu ở Học viện.

Tôi xin cảm ơn những người thân trong gia đình, bạn bè, đồng nghiệp về sự động viên, quan tâm và giúp đỡ trong thời gian qua.

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
CMS	Cepstral Mean Subtraction	Lọc bỏ nhiễu CMS
DCT	Discrete Cosin Transform	Biến đổi gián đoạn Cosin
DFT	Discrete Fourier Transform	Biến đổi gián đoạn Fourier
FFT	Fast Fourier Transform	Biến đổi Fourier nhanh
HMM	Hidden Markov Model	Mô hình Markov ẩn
LPC	Linear Predictive Coding	Mã hoá dự báo tuyến tính
MFCC	Mel Scale Frequency Cepstral Coefficients	Các hệ số cepstral với thang tần số Mel
PLP	Perceptual Linear Prediction	Giác quan dự báo tuyến tính
F0	Fundamental Frequency	Tần số giao động của dây thanh
LDA	Linear Discriminant Analysis	Phương pháp phân tích tuyến tính
GMM	Gaussian Mixture Model	Mật độ xác suất sinh quan sát
HTK	Hidden Markov Model Toolkit	Công cụ cho mô hình HMM

DANH SÁCH HÌNH VẼ

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.

Xử lý tiếng nói trở thành một trong những lĩnh vực quan trọng trong xu hướng phát triển công nghệ của xã hội hiện nay. Đặc biệt, khi công nghệ thông tin ngày càng phát triển thì các ứng dụng của xử lý tiếng nói ngày càng trở lên cấp thiết. Mục đích của những nghiên cứu trong lĩnh vực xử lý tiếng nói là làm cho việc tương tác giữa người và máy ngày càng hiệu quả và tự nhiên hơn.

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.

Trên thế giới đã có rất nhiều hệ thống nhận dạng tiếng nói tiếng Anh đã và đang được ứng dụng rất hiệu quả như: Via Voice của IBM, Spoken Toolkit của CSLU (Central of Spoken Language Under-standing), Speech Recognition Engine của Microsoft, Hidden Markov Model toolkit của đại học Cambridge, CMU Sphinx của đại học Carnegie Mellon,... ngoài ra, một số hệ thống nhận dạng tiến nói tiếng Pháp, Đức, Trung Quốc,... cũng khá phát triển.

Ở Việt Nam, nhận dạng tiếng nói vẫn là một lĩnh vực khá mới mẻ. Đến nay tuy đã có nhiều nghiên cứu về nhận dạng tiếng nói tiếng Việt và đã đạt được một số thành tựu, nhưng nhìn chung vẫn chưa đạt được kết quả cần thiết để có thể tạo ra các sản phẩm mang tính ứng dụng cao. Có thể kể đến các công trình sau:

- **AILab**: Đây là công trình được phòng thí nghiệm Trí tuệ Nhân tạo - AILab thuộc Đại học Khoa học Tự nhiên tạo ra dựa trên các công nghệ tiên tiến nhất về nhận dạng và tổng hợp tiếng nói để đáp ứng nhu cầu của người dùng. Dựa trên công nghệ xử lí tiếng nói tiếng Việt, AILab đã xây dựng phần mềm iSago chuyên hỗ trợ tìm kiếm thông tin qua tiếng nói. Thông qua ứng dụng phần mềm người sử dụng có khả năng hỗ trợ giao tiếp với điện thoại di động trực tiếp bằng lời nói. Từ đó người sử dụng tìm kiếm thông tin nhà hàng, quán Bar, Café trên địa bàn TP. HCM. Khi người dùng đặt câu hỏi bằng tiếng nói, iSago sẽ truyền nội dung truy vấn này về server để xử lý và gửi lại kết quả tìm kiếm, dạng một danh sách: tên nhà hàng, địa chỉ. Phần mềm này cũng cho phép người dùng hiển thị địa chỉ tìm được dạng bản đồ hoặc nghe đọc địa chỉ trực tiếp bằng công nghệ tổng hợp giọng nói. Phần mềm được cung cấp miễn phí tại địa chỉ www.ailab.hcmus.edu.vn
- **Vietvoice**: Đây là phần mềm của một người dân Việt Nam ngụ tại Canada. Phần mềm có khả năng nói tiếng Việt từ các tập tin. Để chạy được chương trình, cần cài đặt Microsoft Visual C++ 2005 Redistributable Package (x86). Đối với người khiếm thị, phần mềm này cho phép sử dụng cách gõ tắt (nhấn nút Ctrl và một chữ) để chọn lựa một trong các tính năng hiển thị trên màn hình. Người dùng có thể cập nhật từ điển các chữ viết tắt và các từ ngữ tiếng nước ngoài.
- **Vspeech**: Đây là một phần mềm điều khiển máy tính bằng giọng nói do một nhóm sinh viên Đại học Bách Khoa TP. HCM viết. Phần mềm sử dụng thư viện Microsoft Speech SDK để nhận dạng tiếng Anh nhưng được chuyển thành tiếng Việt. Nhóm đã khá thành công với ý tưởng này, do sử dụng lại thư viện nhận dạng engine nên thời gian thiết kế rút ngắn lại mà hiệu quả nhận dạng khá tốt. Phần mềm Vspeech có các lệnh gọi hệ thống đơn giản như gọi thư mục My Computer, nút Start,... Phiên bản mới nhất có tương tác với MS Word 2003, lướt web với trình duyệt Internet

Explorer. Không có các chức năng tùy chỉnh lệnh và gọi tắt các ứng dụng. Phần mềm chạy trên nền Windows XP, Microphone và card âm thanh sử dụng tiêu chuẩn thông thường.

Tuy nhiên, việc ứng dụng nhận dạng giọng nói vào điều khiển máy tính còn nhiều hạn chế. Một số sản phẩm của nước ngoài về nhận dạng tiếng nói Tiếng Việt như: Nuance (Dragon Dictation và Dragon Search), Google search,... . Ở Việt Nam thì hầu như chỉ mới có bộ phần mềm Vspeech của nhóm sinh viên trường Đại học Bách Khoa TP. HCM, nhìn chung các phần mềm cũng đều vẫn có những hạn chế nhất định. Phần mềm Vspeech được phát triển từ mã nguồn mở Microsoft Speech SDK nhận dạng tiếng Anh, thông qua dữ liệu, phương thức trung gian, việc nhận dạng được chuyển trong Vspeech để nhận biết tiếng Việt.

Lĩnh vực nghiên cứu và xử lý tiếng nói đã và đang tiếp tục được nghiên cứu, phát triển và các ứng dụng của nó ngày càng trở nên phổ biến và quan trọng. Vì vậy nghiên cứu nhận dạng tiếng nói tiếng Việt là một vấn đề được các nhà nghiên cứu quan tâm, đầu tư công sức trong những năm gần đây. Tiếng Việt là ngôn ngữ đơn âm và có thanh điệu, có nhiều đặc thù khác biệt so với các ngôn ngữ nước ngoài. Việc nghiên cứu nhận dạng tiếng nói tiếng Việt là cần thiết. Các thành quả nghiên cứu nhận dạng tiếng nói của các ngôn ngữ nước ngoài cần được kế thừa và nghiên cứu để áp dụng vào trong tiếng Việt.

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… Sóng âm không thể truyền trong môi trường chân không. Khi kích thích dao động âm trong mối trường không khí thì những lớp khí sẽ bị nén và dãn. Trạng thái nén dãn lần lượt được lan truyền từ nguồn âm dưới dạng sóng dọc tới nơi thu âm. Nếu cường độ nguồn âm càng lớn thì âm thanh truyền đi càng xa [7].

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.

Công suất của tiếng nói, khi nói to nhỏ cũng khác nhau. Khi nói thầm công suất 10^{-3}mW, nói bình thường 10mW, nói to 10^{3}mW.

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. Thanh quản có hai nếp gấp gọi là dây thanh âm, dây thanh âm sẽ rung khi luồng không khí đi qua khe thanh môn là khe giữa hai dây thanh âm. Bộ phận miệng là một ống âm không đều. Bộ phận mũi cũng là một ống âm học không đều có diện tích và chiều dài cố định, bắt đầu từ lỗ mũi đến vòm miệng mềm.

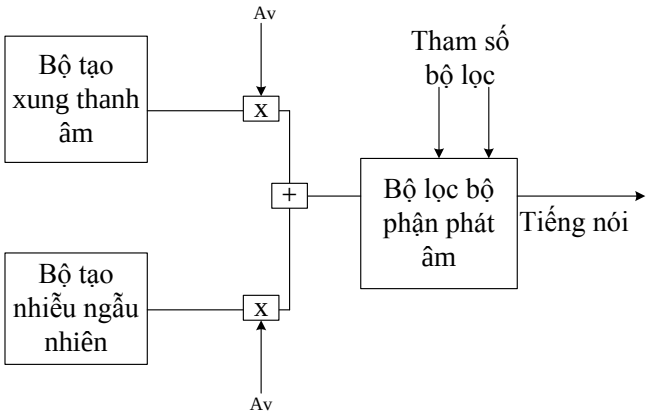
Quá trình tạo ra âm phi mũi: vòm miệng mềm ngăn chặn bộ phận mũi và âm thanh phát ra thông qua môi. Đối với quá trình tạo ra âm mũi: vòm miệng mềm hạ

xuống và bộ phận mũi liên kết bộ phận miệng, lúc này phía trước của bộ phận miệng khép lại hoàn toàn và âm thanh ra thông qua mũi. Đối với âm thanh nói giọng mũi, âm thanh phát ra cả mũi và môi. Âm thanh của tiếng nói có thể chia làm ba loại khác nhau:

- *Âm hữu thanh:* giống như âm khi chúng ta nói 'a' hay 'e' được tạo ra khi dây thanh âm căng lên và rung khi áp suất không khí tăng lên, làm thanh mồm mở ra rồi đóng lại khi luồng không khí đi qua. Những dây thanh âm rung tạo ra dạng sóng của luồng không khí có dạng xấp xỉ tam giác. Chu kì cao độ âm thanh của đàn ông trưởng thành thường từ 50Hz đến 250Hz, giá trị trung bình khoảng 120Hz. Đối với phụ nữ trưởng thành, giới hạn trên cao hơn nhiều, có thể lên đến 500Hz.
- *Âm vô thanh:* được tạo ra khi dây thanh âm không rung. Có hai loại âm vô thanh cơ bản: âm xát và âm hơi. Đối với âm xát như khi ta nói chữ 's', một số điểm trên bộ phận phát âm co lại khi luồng không khí ngang qua nó, hỗn loạn xảy ra tạo nên nhiễu ngẫu nhiên. Đối với âm bật hơi, như khi ta nói chữ 'h', hỗn loạn xảy ra ở gần thanh môn khi dây thanh âm bị giữ nhẹ một phần. Ngoài hai loại âm cơ bản nói trên, còn có một loại âm trung gian vừa mang tính chất nguyên âm, vừa mang tính chất phụ âm, được gọi là bán nguyên âm hay bán phụ âm. Ví dụ như 'i', 'u' trong từ 'ai' và 'âu'.
- *Phụ âm nổ:* ví dụ như âm 'p', 't', 'k' hay 'đ', 'b', 'g' trong tiếng Việt được tạo ra do loại kích thích khác.

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. Như vậy có thể thu được phổ tần số của tín hiệu tiếng nói bằng cách nhân phổ của nguồn âm thanh với đặc tính tần số của bộ lọc. Hình bên dưới minh họa tiếng nói hữu thanh và vô thanh. Các độ lợi A_V và A_N xác định cường độ của nguồn tạo âm hữu thanh và vô thanh.



Hình 1-1: Mô hình lọc nguồn tạo tiếng nói

Mô hình lọc nguồn cho quá trình tạo tiếng nói khá đơn giản nhưng không thể lọc được âm xát bằng cách đỉnh cộng hưởng của bộ phận phát âm như âm hữu thanh hay âm bật hơi, vì vậy mô hình lọc nguồn hoàn toàn không chính xác cho âm xát.

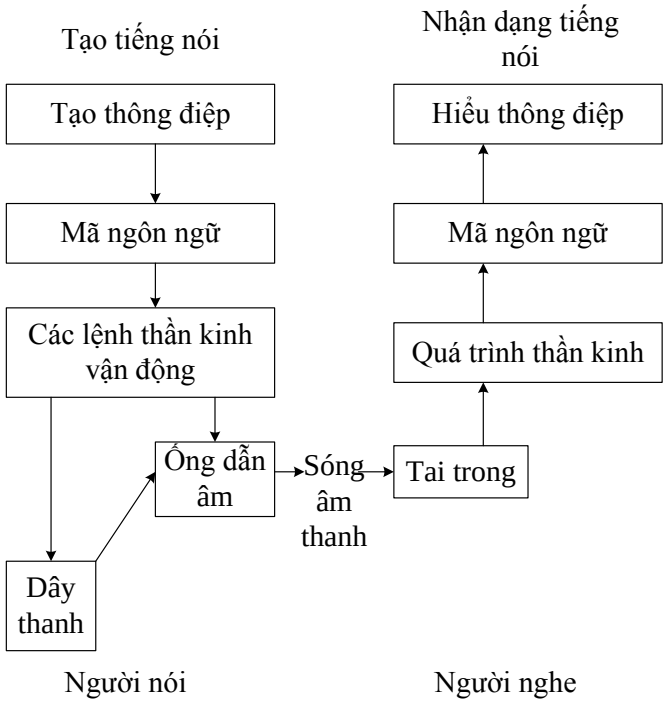
1.1.6. Hệ thống thính giác của người

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.

Khi nghe một sóng âm thuần túy tức âm đơn (sóng sine), những điểm khác nhau trên màng đáy sẽ rung động theo tần số của âm đơn đi vào tai. Điểm lệch lớn nhất trên màng đáy phụ thuộc vào tần số của âm đơn. Tần số cao tạo ra điểm lệch lớn nhất ở phía đáy và tần số thấp tạo ra điểm lệch lớn nhất ở phía đỉnh. Như vậy màng đáy làm nhiệm vụ phân tích tần số tín hiệu vào phức tạp thành những tần số khác nhau ở những điểm khác nhau dọc theo chiều dài của nó. Như vậy có thể xem mọi điểm là bộ lọc thông dải và có tần số trung tâm và băng thông xác định. Ngưỡng nghe của một âm đơn tăng lên khi có sự hiện diện của những âm đơn lân cận khác (âm mặt nạ) và chỉ có bằng tần hẹp xung quanh âm đơn mới tham gia vào hiệu ứng mặt nạ, băng tần này thường gọi là âm tần tới hạn. Giá trị của băng tần tới hạn phụ thuộc vào tần số của âm đơn cần thử. Tóm lại quá trình nghe của hệ thính giác là một dãy các bộ lọc băng thông, có đáp ứng phủ lắp lên nhau và 'băng thông hiệu quả' của chúng xấp xỉ với các giá trị của băng tần tới hạn.

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

Quá trình tạo tiếng nói bắt đầu khi người nói muốn chuyển tải thông điệp của mình cho người nghe thông qua tiếng nói. Hệ thống thần kinh sẽ chịu trách nhiệm chuyển đổi thông điệp sang dạng mã ngôn ngữ. Khi một mã ngôn ngữ được chọn lựa, các lệnh thần kinh vận động điều khiển đồng bộ các khâu vận động nhằm phát ra chuỗi âm thanh. Vậy đầu ra cuối cùng của quá trình là một tín hiệu âm học. Đối với quá trình thu nhận tiếng nói, người nghe xử lý tín hiệu âm thanh thông qua màng tai trong; nó có khả năng cung cấp một phân tích phổ cho tín hiệu tới. Quá trình thần kinh sẽ chuyển đổi tín hiệu phổ thành các tín hiệu hoạt động với thần kinh thính giác; có thể coi đây là quá trình lấy ra các đặc trưng. Cuối cùng các tín hiệu được chuyển thành mã ngôn ngữ và hiểu được thông điệp.

1.1.8. Mô hình lọc nguồn tạo tiếng nói

1.1.8.1. Nguyên âm

Các nguyên âm có tầm rất quan trọng trong nhận dạng tiếng nói; hầu hết các hệ thống nhận dạng dựa trên cơ sở nhận dạng nguyên âm đều có tính năng tốt. Các nguyên âm nói chung là có thời gian tồn tại dài (so với các phụ âm) và dễ xác định phổ. Chính vì thế dễ dàng cho việc nhận dạng tiếng nói, cả đối với con người và máy móc. Về mặt lý thuyết, các cực đại của biểu diễn phổ của tín hiệu nguyên âm chính là các tần số cộng hưởng (formants) tạo nên nguyên âm. Giá trị của các formant đầu tiên (2 hoặc 3 formant đầu tiên) là yếu tố quyết định cho phép chúng ta nhận dạng được nguyên âm. Do nhiều yếu tố biến thiên như sự khác nhau về giới tính, về độ tuổi, tình trạng tinh thần của người nói và nhiều yếu tố ngoại cảnh khác, đối với một nguyên âm xát định các giá trị formant cũng có sự biến thiên nhất định. Tuy nhiên sự khác biệt về các giá trị các formant giữa các nguyên âm khác nhau lớn hơn nhiều; và trong không gian formant chúng ta có thể xác định một cách tương đối các vùng riêng biệt cho từng nguyên âm.

1.1.8.2. Các âm vị khác

Nguyên âm đôi thì có sự biến thiên một cách liên tục các formant của biểu diễn phổ theo thời gian. Đối với âm vị loại này, cần phải đặc biệt chú ý đến việc phân đoạn theo thời gian khi nhận dạng. Các bán nguyên âm như /l/,/r/ và /y/ là tương đối khó trong việc biểu diễn đặc trưng. Các âm thanh này không được coi là nguyên âm nhưng gọi là bán nguyên âm do bản chất tựa nguyên âm của chúng. Các đặc trưng âm học của các âm thanh này chịu ảnh hưởng rất mạnh của ngữ cảnh mà trong đó chúng xuất hiện. Đối với các âm mũi thì miệng đóng vai trò như một khoảng cộng hưởng có tác dụng bẫy năng lượng âm tại một vài tần số tự nhiên. Các tần số cộng hưởng này của khoang miệng xuất hiện như các phản cộng hưởng, hay các điểm không của hàm truyền đạt. Ngoài ra, các phụ âm mũi còn được đặc trưng bởi những sự cộng hưởng mạnh hơn về phổ so với các nguyên âm. Các phụ âm xát vô thanh như /s/, /sh/. Hệ thống tạo ra các phụ âm xát vô thanh bao gồm một nguồn nhiễu tại một điểm thắt mà chia ống dẫn âm thành hai khoang. Âm thanh được bức

xạ tại khoang trước. Khoang sau có tác dụng bẫy năng lượng như trong trường hợp phụ âm mũi, và như vậy là đưa các phản cộng hưởng vào âm thanh đầu ra. Bản chất không tuần hoàn là đặc trưng cơ bản nhất của nguồn kích thích xác vô thanh. Điểm khác biệt của các âm xát hữu thanh như /v/, /th/ so với các phụ âm xát vô thanh là ở chỗ có hai nguồn kích thích liên quan tới việc tạo ra chúng. Như vậy đặc trưng của phụ âm xát hữu thanh là bao gồm cả hai thành phần kích thích tuần hoàn và nhiễu. Các âm dừng là các phụ âm /b/, /d/, /g/, /p/, /t/ và /k/ chúng có thời gian tồn tại rất ngắn. Các âm dừng có tính chất động vì thế các thuộc tính của chúng chịu ảnh hưởng rất nhiều bởi nguyên âm đi sau nó.

1.2. Giới thiệu về xử lý tiếng nói

Xử lý tiếng nói ngày nay đang là vấn đề được quan tâm nghiên cứu nhiều bởi khả năng ứng dụng trong nhiều lĩnh vực như: Công nghệ thông tin, Viễn thông, tự động hóa (chế tạo người máy có khả năng tương tác với con người)... qua đó giúp quá trình tương tác giữa người với máy trở nên hiệu quả và tự nhiên hơn.

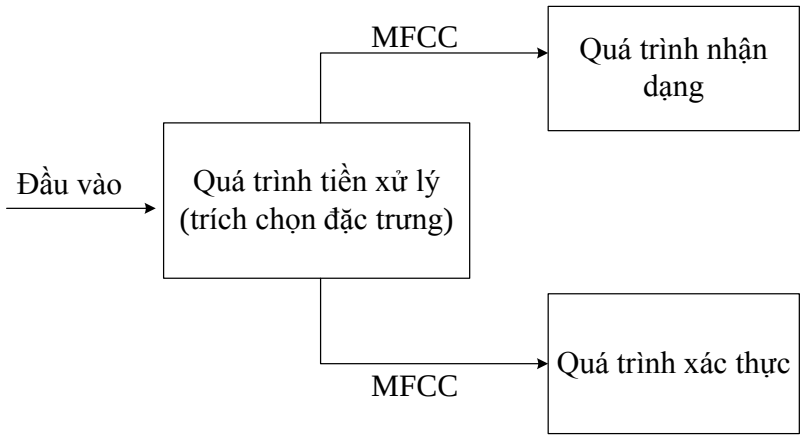
Quá trình xử lý tín hiệu tiếng nói là quá trình thu nhận, lưu trữ và truyền tín hiệu. Quá trình nhận dạng, tổng hợp tiếng nói hay xác thực người nói thông qua giọng nói là các ví dụ điển hình của quá trình xử lý tín hiệu tiếng nói.

1.2.1. Mục đích của xử lý tiếng nói

Thực hiện xử lý, mã hoá một cách có hiệu quả tín hiệu tiếng nói để truyền và lưu trữ tiếng nói.

Tổng hợp và nhận dạng tiếng nói tới giao tiếp người-máy bằng tiếng nói dựa vào các thông tin của quá trình tiền xử lý.

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

Thông tin đầu vào là tín hiệu tiếng nói do con người phát ra dưới dạng tương tự, sau đó tín hiệu này được số hóa (rời rạc, lượng tử và mã hóa dạng nhị phân). Quá trình tiền xử lý tiếng nói tiến hành xử lý tín hiệu tiếng nói cho kết quả là các tham số của tín hiệu tiếng nói (Các hệ số MFCC và LPC). Các tham số này trở thành đầu vào đối với tất cả các ứng dụng của xử lý tiếng nói. Như vậy tất cả các ứng dụng của xử lý tiếng nói đều cần phải dựa trên các kết quả của quá trình tiền xử lý. Kết quả của quá trình này góp phần quyết định tính chính xác và hiệu quả của các ứng dụng.

1.3. Nhận dạng tiếng nói

1.3.1. Bài toán nhận dạng tiếng nói

Nhận dạng tiếng nói tự động là một kỹ thuật nhằm làm cho máy "hiểu" được tiếng nói của con người. Thực chất đây là một quá trình biến tín hiệu tiếng nói do người phát ra thành tín hiệu số sau đó sử dụng một số giải thuật để đối chiếu giữa tín hiệu thu được với các dữ liệu tham chiếu để xác định xem tín hiệu thu được tương ứng với dữ liệu tham chiếu nào trong bộ tham chiếu (từ điển nhận dạng). Kết quả của việc nhận dạng sau đó có thể được sử dụng trong các ứng dụng khác như nhập số liệu, soạn thảo văn bản bằng lời nói, điều khiển tự động…

Mục tiêu của hầu hết các chương trình nhận dạng tiếng nói là kết quả nhận dạng đạt đến độ chính xác 100% mà không phụ thuộc vào một điều kiện nào cả. Tuy nhiên tất cả các nghiên cứu gần đây chỉ cho độ chính xác đến khoảng trên 90%

trong một số điều kiện cụ thể nào đó còn những chương trình nhận dạng mà không có điều kiện giới hạn gì thì độ chính xác chỉ đạt không quá 87%.

Các chương trình nhận dạng tiếng nói tự động hiện nay khá nhiều và hết sức đa dạng. Tuy nhiên chúng ta cũng có thể dựa vào một số đặc điểm để phân chúng thành một số dạng chủ yếu như:

- Nhận dạng các từ phát âm rời rạc/liên tục:

  Trong các chương trình nhận dạng các từ phát âm rời rạc yêu cầu người nói phải dừng một khoảng trước khi nói từ tiếp theo. Còn hệ thống nhận dạng các từ phát âm liên tục không yêu cầu điều kiện này.

- Nhận dạng tiếng nói độc lập/phụ thuộc người nói

  Đối với hệ thống nhận dạng phụ thuộc người nói đòi hỏi tiếng người nói phải có trong cơ sở dữ liệu của hệ thống còn hệ thống nhận dạng không phụ thuộc người nói thì người nói không nhất thiết phải có mẫu trước khi nhận dạng trong cơ sở dữ liệu.

- Nhận dạng với từ điển cỡ nhỏ/vừa/lớn

  Hiệu năng của một hệ thống nhận dạng với từ điển cỡ nhỏ thường cao hơn hiệu năng của các hệ thống nhận dạng có từ điển cỡ vừa và lớn.

- Nhận dạng trong môi trường nhiễu cao/thấp

  Hiệu năng của các hệ thống nhận dạng không nhiễu sẽ cao hơn hiệu năng của các hệ thống nhận dạng có nhiễu.

  Tín hiệu tiếng nói sau khi được số hóa sẽ phân thành các khung có độ dài khoảng từ 10ms đến 45ms qua bước phân tích và xác định các đặc tính sẽ cho ta một dãy các vector đặc trưng của tiếng nói. Các vector này sau đó sẽ được sử dụng để tìm kiếm các từ giống nhất trong từ điển dựa trên một số điều kiện ràng buộc nào đó về mặt âm thanh, ngữ nghĩa, từ vựng…

  Do tính chất của tiếng nói phụ thuộc vào nhiều yếu tố nên việc thu nhận, phân tích các đặc trưng của tiếng nói là việc không dễ dàng. Ở đây, chúng ta có thể nêu ra một số yếu tố khó khăn cho bài toán nhận dạng tiếng nói:

  - Khi phát âm, người nói thường nói nhanh chậm khác nhau.

- Các từ được nói thường dài ngắn khác nhau.
- Một người cùng nói một từ nhưng ở hai lần phát âm khác nhau thì cho kết quả phân tích khác nhau.
- Mỗi người có một chất giọng riêng được thể hiện thông qua độ cao của âm, độ to của âm, cường độ âm và âm sắc.
- Những yếu tố như nhiễu của môi trường, nhiễu của thiết bị thu…

1.3.2. Các phương pháp nhận dạng tiếng nói

Như đã đề cập trong phần trên, hiện nay có ba phương pháp chủ yếu được sử dụng trong nhận dạng tiếng nói là:

Phương pháp âm học - ngữ âm học

Phương pháp nhận dạng mẫu

Phương pháp ứng dụng trí tuệ nhân tạo

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. Theo lý thuyết này thì trong bất kỳ một ngôn ngữ nào cũng luôn tồn tại một số hữu hạn các đơn vị ngữ âm phân biệt và những đơn vị ngữ âm đó được đặc trưng bởi các thuộc tính vốn có trong tín hiệu tiếng nói, hoặc trong phổ của nó thông qua thời gian.

Nguyên lý hoạt động của hệ thống này như sau:

- ***Bước đầu tiên***: Tín hiệu tiếng nói sau khi số hoá được đưa qua một bộ "đo" các đặc tính của tiếng nói, mục đích là nhằm biểu diễn xấp xỉ các đặc tính của tiếng nói thay đổi theo thời gian. Bước này là cần thiết cho hầu hết các hệ thống nhận dạng theo các hướng tiếp cận khác nhau.
- ***Bước thứ hai***: Là bước tách các đặc tính của tiếng nói nhằm biến đổi các số đo phổ tín hiệu thành một tập các đặc trưng mô tả các đặc tính âm học của các đơn vị ngữ âm khác nhau. Các đặc trưng đó có thể là: Tính chất âm mũi, âm xát, vị trí các formant…
- ***Bước thứ ba***: Là bước phân đoạn và gán nhãn. Ở bước này hệ thống nhận dạng cố gắng tìm các vùng âm thanh ổn định và gán cho mỗi vùng này một

nhãn phù hợp với đặc tính của đơn vị ngữ âm. Đối với một hệ thống nhận dạng theo hướng âm học ngữ âm học thì bước này là tâm điểm và khó thực hiên nhất. Do đó có rất nhiều chiến lược đã được sử dụng để giới hạn phạm vi của các điểm phân đoạn và xác xuất gán nhãn.

- ***Bước cuối cùng***: Từ các khối ngữ âm thu được sau bước phân đoạn và gán nhãn, người ta dựa vào một số nguyên tắc lựa chọn để kết hợp các khối ngữ âm này thành các từ, câu nhận dạng.

Có rất nhiều vấn đề đối với một hệ thống nhận dạng tiếng nói theo hướng âm học - ngữ âm học những vấn đề này bằng nhiều cách khác nhau nó ảnh hưởng tới hiệu quả của một hệ thống nhận dạng. Những vấn đề đó là:

+ Cần có sự hiểu biết về các đặc tính âm học của các đơn vị ngữ âm. Sự hiểu biết này không thể đầy đủ cho tất cả nhưng đối với một số trường hợp đơn giản thì có thể cho kết quả tốt.
+ Sự chọn lựa các đặc trưng dựa của tiếng nói hầu hết tùy thuộc vào một khía cạnh cụ thể mà ta quan tâm. Chúng được chọn theo trực giác không tối ưu và đầy đủ ý nghĩa.
+ Việc thiết kế của các hệ thống phân lớp âm thanh cũng không tối ưu và hầu hết nó đều dựa trên cây nhị phân quyết định…

  Không có một thủ tục tự động lựa chọn ngưỡng chính xác để làm căn cứ cho việc gán nhãn. Trên thực tế không có một phương pháp lý tưởng để gán nhãn cho tập huấn luyện. Từ đó, hướng tiếp cận âm học - ngữ âm học muốn áp dụng được vào thực tế cần phải có thêm nhiều nghiên cứu nữa.

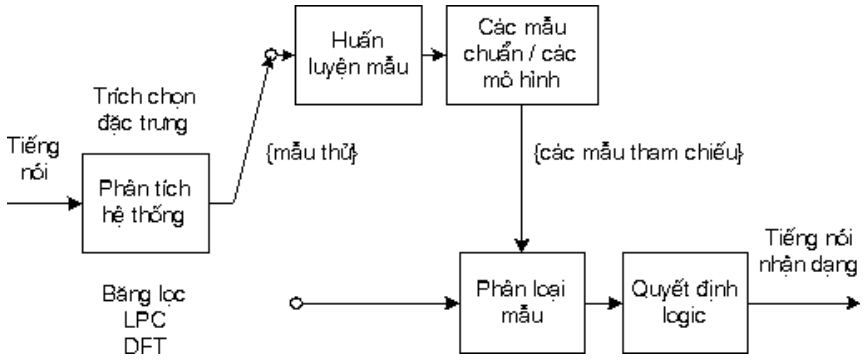
b. Phương pháp nhận dạng mẫu

Phương pháp nhận dạng mẫu sử dụng trực tiếp mẫu tiếng nói mà không cần phải xác định các đặc trưng hay phân đoạn một cách rõ ràng. Trong hầu hết các hệ thống, nhận dạng mẫu bao gồm hai bước.

- ***Bước đầu tiên***: là bước huấn luyện. Ở bước này dựa trên nhiều phiên bản khác nhau của mẫu cần nhận dạng, hệ thống tạo ra các mẫu tham chiếu dùng để so sánh với mẫu cần nhận dạng ở bước sau.

- ***Bước thứ hai***: là bước nhận dạng. Ở bước này mẫu cần nhận dạng được so sánh với các mẫu tham chiếu để xác định xem nó "giống" mẫu tham chiếu nào nhất. Mẫu tham chiếu giống nó nhất chính là kết quả nhận dạng.

  Tư tưởng của phương pháp này là nếu như có đủ các phiên bản khác nhau của mẫu cần nhận dạng thì thông qua bước huấn luyện hệ thống có thể xác định một cách chính các đặc trưng của mẫu. Việc xác định các đặc trưng thông qua bước huấn luyện được gọi là **phân lớp mẫu**. Hiện nay, có hai phương pháp nhận dạng mẫu được sử dụng rộng rãi đó là mô hình **Markov ẩn** và mô hình sử dụng **mạng nơron**. Sơ đồ khối của một hệ thống nhận dạng mẫu như sau:



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

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

- ✓ **Trích chọn các đặc trưng**: Ở bước này dựa trên một số biện pháp phân tích để xác định các đặc trưng của các mẫu. Đối với các hệ thống nhận dạng tiếng nói có hai phương pháp cơ bản là phương pháp phân tích hệ số phổ theo thang độ Mel (MFCC) và phương pháp phân tích mã hóa dự đoán tuyến tính (LPC).
- ✓ **Huấn luyện mẫu**: Ở bước này, hệ thống dựa trên các đặc trưng của các mẫu trong cùng một lớp được tạo ra ở bước trước để tạo nên các mẫu tham chiếu của hệ thống. Ví dụ trong hệ thống nhận dạng từ, để xây dựng nên một từ tham chiếu chúng ta phải thu từ đó lặp đi lặp lại nhiều lần, sau đó trích chọn các đặc trưng của những từ này nhằm tạo một từ tham chiếu cho hệ thống.

- ✓ **Phân lớp mẫu**: Trong bước này, mẫu cần nhận dạng được so sánh với các mẫu tham chiếu. Ở đây, cần một thủ tục để tính khoảng cách cục bộ, và quy chuẩn thời gian giữa các mẫu.
- ✓ **Quyết định logic**: Sau bước phân lớp mẫu ta có được điểm đánh giá sự "giống" nhau giữa mẫu cần nhận dạng và mẫu tham chiếu. Những thông số điểm này sẽ được sử dụng để đưa ra quyết định là mẫu nào "giống" với mẫu cần nhận dạng nhất.

Đặc điểm của một hệ thống nhận dạng mẫu:

Hiệu năng của hệ thống rất nhạy cảm với số mẫu dữ liệu có trong tập huấn luyện. Thông thường, khi mà số mẫu có trong tập huấn luyện càng nhiều thì hiệu năng nhận của hệ thống càng cao.

Mẫu tham chiếu rất nhạy cảm với môi trường thu âm và đặc tính của đường truyền do đặc tính phổ của tiếng nói chịu tác động của đường truyền và nhiễu nền. Không cần có những hiểu biết đặc biệt về ngôn ngữ chính vì vậy hệ thống này ít phụ thuộc vào kích thước từ điển, cú pháp và ngữ nghĩa.

Khối lượng tính toán trong thủ tục huấn luyện hoặc nhận dạng tỷ lệ tuyến tính với số mẫu dùng huấn luyện hoặc nhận dạng.

c. Phương pháp ứng dụng trí tuệ nhân tạo

Phương pháp này là sự lai tạo của hai phương pháp trên với mục đích khai thác tối đa ưu điểm của từng phương pháp. Phương pháp này điều chỉnh thủ tục nhận dạng theo cách mà con người sử dụng trí tuệ của mình trong việc quan sát, phân tích và cuối cùng đưa ra một quyết định dựa trên các thông số đặc trưng về âm học. Những kỹ thuật thường được sử dụng cùng với các phương pháp này là:

Sử dụng hệ chuyên gia để phân đoạn và gán nhãn do đó bước chủ yếu và khó nhất được thực hiện đơn giản hơn so với một hệ thống nhận dạng chỉ dựa vào hướng tiếp cận âm học - ngữ âm học thuần tuý.

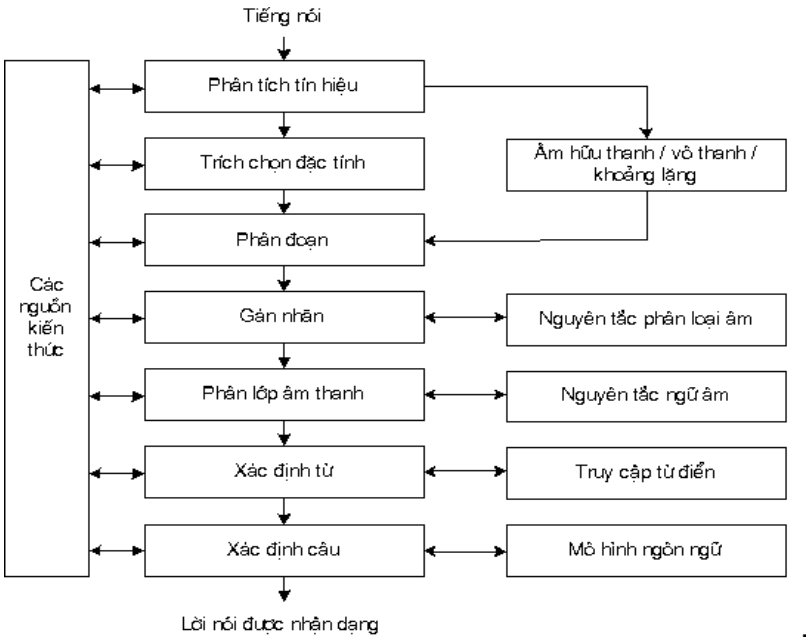
Sử dụng mạng nơron để học mối quan hệ giữa các đơn vị ngữ âm và tất cả các đầu vào đã nhận biết (bao gồm âm học, ngôn ngữ học, cú pháp, ngữ nghĩa…), sau đó sử dụng mạng này để nhận dạng.

Mục đích của việc sử dụng hệ chuyên gia là nhằm tận dụng các nguồn kiến thức của con người vào hệ thống nhận dạng. Các nguồn kiến thức đó bao gồm:

- *Kiến thức về âm học*: Nhằm để phân tích phổ và xác định đặc tính âm học của các mẫu tiếng nói đầu vào.
- *Kiến thức về từ vựng*: Sử dụng để kết hợp các khối ngữ âm thành các từ cần nhận dạng.
- *Kiến thức về cú pháp*: Nhằm kết hợp các từ thành các câu cần nhận dạng.
- *Kiến thức về ngữ nghĩa*: Nhằm xác định tính logic của các câu đã được nhận dạng.

Sự kết hợp các nguồn kiến thức phụ thuộc vào hệ chuyên gia mà hệ thống nhận dạng sử dụng.

Có nhiều cách khác nhau để có thể kết hợp các nguồn kiến thức. Cách thông dụng nhất là xử lý từ dưới lên, trong đó các tiến trình ở mức thấp nhất (như trích chọn đặc trưng, giải mã ngữ nghĩa) được đặt trên các tiến trình cao hơn (như giải mã từ vựng, mô hình ngôn ngữ) theo một tiến trình tuần tự nhằm giảm việc xử lý trong mỗi tầng xuống mức nhỏ nhất có thể. Sơ đồ khối của phương pháp này như sau:



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

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. Điều này làm cho sự thể hiện của âm tiết tiếng Việt trong chuỗi lời nói nổi bật và tách bạch hơn. Do đó nên việc vạch ra ranh giới giữa các âm tiết trong tiếng Việt dễ dàng hơn nhiều việc phân chia ranh giới âm tiết trong các ngôn ngữ châu Âu [6] (trong ngôn ngữ châu Âu, việc phân chia âm tiết có khi phải dùng phương pháp phân tích phổ). Việc tách bạch âm tiết còn được thể hiện ở chữ viết, mỗi âm tiết được viết tách ra thành một từ riêng biệt. Có thể nói so với các âm tiết châu Âu, tiếng Việt có tính độc lập cao hơn hẳn.

Trong các ngôn ngữ châu Âu thường gặp các hiện tượng nối âm (liaison), trong tiếng Việt không có hiện tượng nối âm như vậy.

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. Nói cách khác trong tiếng Việt ranh giới của âm tiết trùng với ranh giới của hình vị [5] (hình vị là đơn vị có nghĩa nhỏ nhất trong một ngôn ngữ). Chính vì vậy trong một phát ngôn, số lượng âm tiết trùng với số lượng hình vị.

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. Nói cách khác trong tiếng Việt ranh giới của âm tiết trùng với ranh giới của hình vị [5] (hình vị là đơn vị có nghĩa nhỏ nhất trong một ngôn ngữ). Chính vì vậy trong một phát ngôn, số lượng âm tiết trùng với số lượng hình vị.

1.4.1.3. Cấu trúc chặt chẽ

Mỗi âm tiết tiếng Việt ở dạng đầy đủ có 5 phần như Hình 1-6:

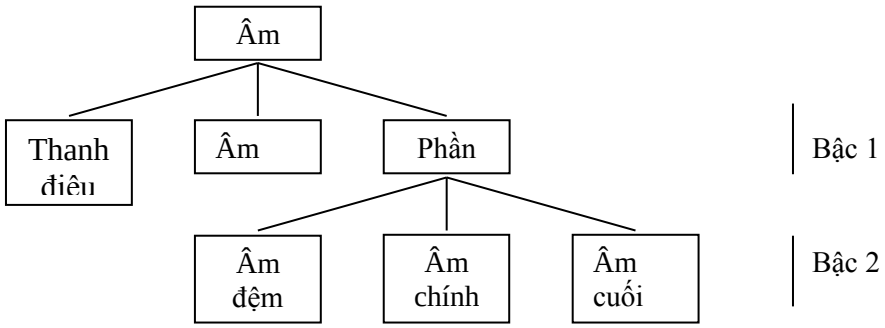
Cấu trúc tổng quát của một âm tiết tiếng Việt là (C1)(w)V(C2). Trong đó C1 là phụ âm đầu, (w) là âm đệm, V là âm chính và C2 là âm cuối.

<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]

Âm tiết tiếng Việt có cấu trúc gồm hai bậc: bậc một bao gồm các thành tố trực tiếp được phân định bằng những ranh giới có ý nghĩa ngữ âm học. Phần thứ hai bao gồm các yếu tố của phần vần chỉ có chức năng khu biệt thuần tuý. Quan hệ giữa các yếu tố ở bậc một là quan hệ lỏng lẻo, giữa các yếu tố của bậc hai có quan hệ chặt chẽ. Các thực nghiệm đã chứng minh rằng: tính độc lập của thanh điệu đối với các âm vị cụ thể lộ ra ở chỗ đường nét âm điệu và trường độ của nó không gắn liền với thành phần âm thanh của âm tiết.

Theo GS. Bảng và cộng sự [1] số lượng âm tiết phát âm được của tiếng Việt là 18958. So với các các ngôn ngữ thông thường trên thế giới có số lượng âm tiết vào khoảng 3000-5000. Điều này cho thấy tiếng Việt có số lượng âm tiết rất lớn, và chính vì thế ít có hiện tượng đồng âm, ít gây trở ngại cho việc nhận diện âm tiết. Theo [6], trong tiếng Việt có 6 thanh điệu, 21 âm đầu và 155 phần vần và phần vần đóng vai trò khu biệt lớn hơn cả so với các yếu tố khác trong Bậc 1.



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

1.4.2. Âm vị tiếng Việt

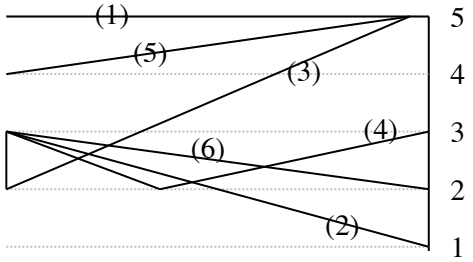
Âm vị là đơn vị đoạn tính nhỏ nhất có chức năng phân biệt nghĩa. Về mặt xã hội của ngữ âm, trong số các âm vị trong lời nói của ngôn ngữ, ta có thể tập hợp một số lượng có hạn những đơn vị mang những nét chung về cấu tạo âm thanh và về chức năng trong ngôn ngữ đó gọi là âm vị.

Có một cản trở khi nghiên cứu âm vị tiếng Việt là chưa có một qui định chính thức về pháp lý, hay một chuẩn chung của các nhà khoa học ngữ âm về một chuẩn tiếng Việt. Có thể quan niệm tạm thời coi "tiếng Việt chuẩn như một thứ tiếng chung được hình thành trên cơ sở tiếng địa phương của miền Bắc với trung tâm là Hà Nội mà cách phát âm của nó là cách phát âm Hà Nội với sự phân biệt /ţ-c/,/ş-s/,/ẓ-z/ và các vần ưu/iu, ươu/iêu" [5].

1.4.2.1. Thanh điệu

Âm vị tiếng Việt có hai loại âm vị đoạn tính và âm vị siêu đoạn tính. Âm vị đoạn tính là các đơn vị có thể chia cắt được trong chuỗi lời nói như nguyên âm, phụ âm. Âm vị siêu đoạn tính là loại đơn vị không có âm đoạn tính, không độc lập tồn tại, nhưng cũng có chức năng phân biệt nghĩa, nhận diện từ, đó là thanh điệu. Đây là đặc điểm riêng của tiếng Việt so với các ngôn ngữ Châu Âu. Một số ngôn ngữ khác như tiếng Hán, tiếng Thái cũng có đặc điểm này như tiếng Việt.

Thanh điệu được hình thành bằng sự rung động của dây thanh, tuỳ theo sự rung đó nhanh hay chậm, mạnh hay yếu, biến chuyển ra sao mà ta có các thanh điệu khác nhau. Thanh điệu tiếng Việt thuộc loại thanh lướt, có nghĩa là các thanh điệu phân biệt với nhau bằng sự di chuyển cao độ từ thấp lên cao hay từ cao xuống thấp.



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

Theo các nhà ngôn ngữ học thì thanh điệu có ảnh hưởng bao chùm lên toàn bộ âm tiết, mặc dù gánh nặng chủ yếu tập trung ở phần vần. Tiếng Việt có sáu thanh điệu. Nếu chia thang độ của giọng nói bình thường thành 5 bậc thì ta có thanh điệu tiếng Việt được miêu tả như trong Hình 1-8.

1.4.2.2. Âm đầu

Trong các sách giáo khoa tiếng Việt [3, 4], tiếng Việt có 21 âm vị là âm đầu. Các âm vị /p,r/ không được liệt kê là các âm vị đầu tiếng Việt và được coi là âm vị có nguồn gốc từ ngôn ngữ nước ngoài. Âm vị / ʔ/ âm tắc thanh hầu được liệt kê trong một số sách giáo khoa tiếng Việt như một phụ âm đầu. Trong những âm tiết như: "ai, ơi, ăn, oản, uống, oanh, uyên" có hiện tượng khép khe thanh lúc mở đầu khi chúng được phát âm lên. Tiếng bật do động tác mở khe thanh đột ngột được nghe rõ hoặc không rõ ở từng người, trong từng lúc, phụ thuộc vào phong cách và bối cảnh ngữ âm. Thừa nhận tồn tại âm tắc thanh hầu đưa đến xây dựng được một mô hình tổng quát của âm tiết tiếng Việt cân xứng hơn với ba thành tố luôn có mặt: thanh điệu, âm đầu, âm vần [6].

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. Khác với âm chính luôn nằm ở đỉnh âm tiết, âm đệm nằm ở đường cong đi lên của đỉnh âm tiết. Âm đệm không xuất hiện trước các nguyên âm tròn môi /u,o, ɔ/, nó chỉ xuất hiện trước các nguyên âm hàng trước. Độ mở của âm đệm phụ thuộc vào độ mở của các nguyên âm-âm chính đi sau.

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. Âm chính âm tiết có thể chia thành 4 nhóm:

Nhóm nguyên âm đơn, hàng trước, không tròn môi. Âm sắc của nhóm này thường là bổng. Có thể dài và thể ngắn. Thể ngắn có sự biến dạng ít nhiều về trường độ, âm sắc, cường độ, phát âm căng và ngắn.

Nhóm nguyên âm đơn, hàng sau tròn môi. Âm sắc trầm. Có thể dài và thể ngắn. Sự thể hiện thể ngắn có cấu âm không giữ đều

Nhóm nguyên âm đơn, hàng sau, không tròn môi. Âm sắc trầm vừa.Nguyên âm đôi. phát âm yếu dần, yếu tố đầu phát âm mạnh hơn yếu tố sau, do đó âm sắc của nguyên âm đôi là do yếu tố đầu quyết định. Nguyên âm chỉ có một thể dài và không bị biến dạng về âm sắc và trường độ.

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). Do đó có sự khác biệt lớn giữa âm /t/ trong phát âm hai từ "at" và "ta". Trong khi phát âm từ "ta", lối thoát của không khí được khai thông sau khi bị cản trở bằng một động tác mở ra tạo thành một tiếng động đặc thù. Trong khi phát âm từ "at", bộ phận cấu âm ở nguyên vị trí cấu âm và không khí không được thoát ra ngoài [5].

Trong nhiều trường hợp phụ âm cuối hầu như chỉ là một khoảng im lặng. Ví dụ như âm vị /k/ trong từ "tác". Do vậy âm vị /k/ được nhận diện chủ yếu làm biến đổi âm sắc của âm chính đi ở giai đoạn cuối.

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]

Bán nguyên âm cũng không thường xuyên được thể hiện rõ rệt mà chỉ được nhận diện bằng việc biến đổi âm sắc của âm chính. Về mặt này thì bán nguyên âm còn có tác dụng mạnh hơn là phụ âm cuối.

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.

Theo thống kê trong tiếng Việt có khoảng 6000 âm tiết. Nhìn về mặt ghi âm: âm tiết tiếng Việt có cấu tạo chung là: phụ âm - vần. Ví dụ âm *tin* có phụ âm *t*, vần *in*.

Phụ âm là một âm vị và âm vị này liên kết rất lỏng lẻo với phần còn lại của âm tiết (ví dụ hiện tượng nói lái).

Vần trong tiếng Việt lại được cấu tạo từ các âm vị nhỏ hơn, trong đó có một âm vị chính là nguyên âm.

Ngoài ra, tiếng Việt là ngôn ngữ có thanh điệu. Hệ thống thanh điệu gồm 6 thanh: bằng, huyền, sắc, hỏi, ngã, nặng.

Thanh điệu trong âm tiết là âm vị siêu đoạn tính (thể hiển trên toàn bộ âm tiết). Do đó đặc trưng về thanh điệu thể hiện trong tín hiệu tiếng nói không rõ nét như các thành phần khác của âm tiết.

Sự khác biệt về cách phát âm tiếng Việt rất rõ rệt theo giới, lứa tuổi và đặc biệt là theo vị trí địa lí (giọng miền Bắc, miền Trung và miền Nam khác nhau rất nhiều).

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

1.4.5.1. Thuận lợi

- Tiếng Việt là ngôn ngữ đơn âm, số lượng âm tiết không quá lớn. Điều này sẽ giúp hệ nhận dạng xác định ranh giới các âm tiết dễ dàng hơn nhiều. Đối với hệ nhận dạng các ngôn ngữ Ấn-Âu (tiếng Anh, tiếng Pháp...) xác định ranh giới âm tiết (endpoint detection) là vấn đề rất khó và ảnh hưởng lớn đến kết quả nhận dạng.
- Tiếng Việt là ngôn ngữ không biến hình từ. Âm tiết tiếng Việt ổn định, có cấu trúc rõ ràng. Đặc biệt không có 2 âm tiết nào đọc giống nhau mà viết khác nhau. Điều này sẽ dễ dàng cho việc xây dựng các mô hình âm tiết trong nhận dạng; đồng thời việc chuyển từ phiên âm sang từ vựng (lexical decoding) sẽ đơn giản hơn so với các ngôn ngữ Ấn-Âu.

1.4.5.2. Khó khăn

- Tiếng Việt là ngôn ngữ có thanh điệu (6 thanh). Thanh điệu là âm vị siêu đoạn tính, đặc trưng về thanh điệu thể hiện trong tín hiệu tiếng nói không rõ nét như các thành phần khác của âm tiết.

- Cách phát âm tiếng Việt thay đổi nhiều theo vị trí địa lí. Giọng địa phương trong tiếng Việt rất đa dạng (mỗi miền có một giọng đặc trưng).
- Hệ thống ngữ pháp, ngữ nghĩa tiếng Việt rất phức tạp, rất khó để áp dụng vào hệ nhận dạng với mục đích tăng hiệu năng nhận dạng. Hệ thống phiên âm cũng chưa thống nhất.
- Các nghiên cứu về nhận dạng tiếng Việt cũng chưa nhiều và ít phổ biến. Đặc biệt khó khăn lớn nhất là hiện nay chưa có một bộ dữ liệu chuẩn cho việc huấn luyện và kiểm tra các hệ thống nhận dạng tiếng Việt.

1.5. Kết luận

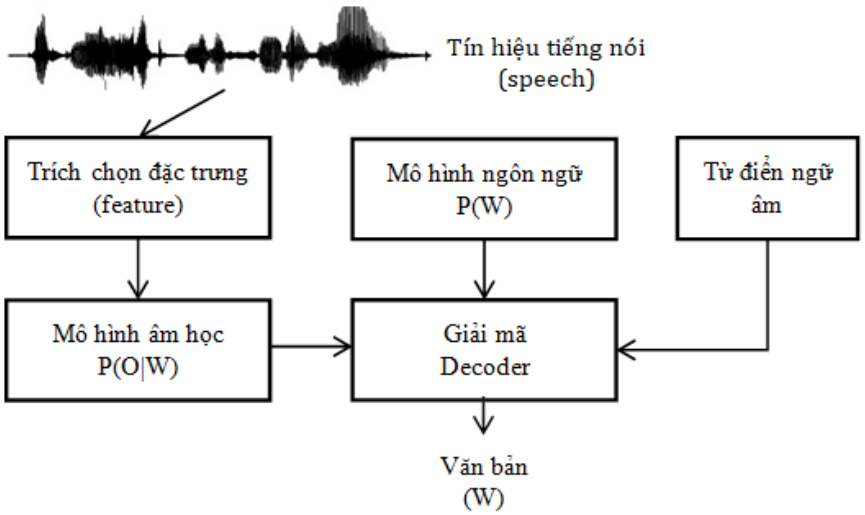
Chương 1 này đã giới thiệu một cách tổng quan về hệ thống nhận dạng tiếng nói và các đặc trưng âm thanh cần thiết cho quá trình nhận dạng từ vựng từ âm thoại, các cơ sở lý thuyết của một hệ thống nhận dạng tiếng nói. Qua chương này, chúng ta đã nắm được khái quát về hệ thống nhận dạng tiếng nói, các cơ sở lý thuyết của một hệ thống nhận dạng tiếng nói, các giai đoạn cơ bản của hệ thống nhận dạng tiếng nói. Các nghiên cứu hiện thời về nhận dạng tiếng nói đối với tiếng Việt, cơ sở dữ liệu tiếng nói, một bộ phận gắn liền với nhận dạng tiếng nói và đặc điểm của ngôn ngữ tiếng Việt cũng đã được trình bày.

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.

Vì tín hiệu tiếng nói là dạng chuỗi thông tin biến đổi theo miền thời gian nên HMM rất phù hợp để ứng dụng trong nhận dạng tín hiệu tiếng nói. Theo các nghiên cứu [8], [9 - 11] so sánh hiệu suất nhận dạng tiếng nói bằng các phương pháp trên, chỉ ra rằng phương pháp ứng dụng HMM cho kết quả tối ưu nhất đặc biệt là đối với các ứng dụng xử lý tín hiệu thời gian thực và dữ liệu huấn luyện lớn.

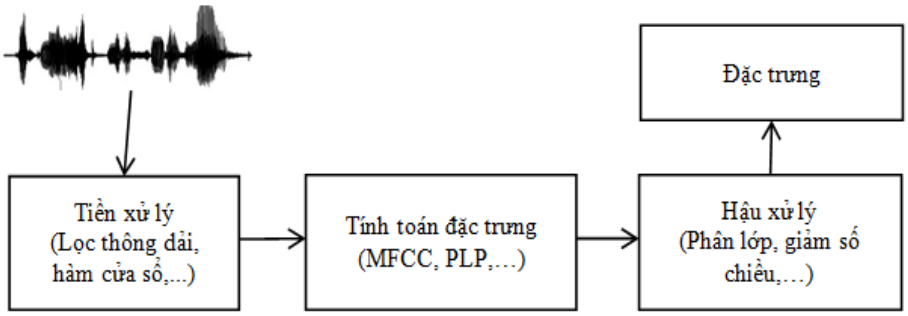
2.1. Các thành phần chính của một hệ thống nhận dạng tiếng nói



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 trích chọn đặc trưng áp dụng một số kỹ thuật nhằm làm giảm độ phức tạp của tín hiệu tiếng nói đầu vào, đồng thời rút trích các thông tin quan trọng và có ý nghĩa cho việc mô hình hóa và nhận dạng. Đầu ra thu được một chuỗi các vector đặc trưng (hay còn gọi là các quan sát) ký hiệu là *O*. 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.

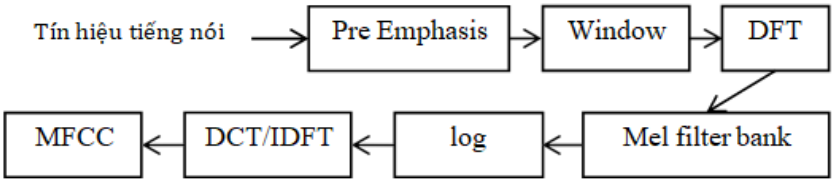
a. Khâu tiền xử lý: Có nhiệm vụ chính là lọc nhiễu, rút trích các tín hiệu nằm trong miền tần số mà tai người nghe được (0-10kHz), chia tín hiệu tiếng nói thành các khung có kích thước từ 10ms đến 30ms (còn gọi là hàm cửa sổ Window), độ lệch giữa hai khung liên tiếp thường nằm trong khoảng 10ms- 20ms.

b. Khâu tính toán đặc trưng: Biến đổi tín hiệu sang miền tần số qua phép biến đổi Fourier rời rạc (DFT), thực hiện các tính toán để thu được đặc trưng. Hai loại đặc trưng được sử dụng phổ biến trong nhận dạng tiếng nói là các hệ số đường bao phổ của tần số mel (Mel Frequency Cepstral Coefficient - MFCC) và mã dự báo tuyến tính giác quan (Perceptual Linear Prediction - PLP).

c. Khâu hậu xử lý: Để nâng cao chất lượng đặc trưng và giảm kích thước vector đặc trưng trước khi đưa vào mô hình ngôn ngữ. Một trong các phương pháp phân lớp và giảm số chiều thường được áp dụng trong nhận dạng tiếng nói là phương pháp phân tích tuyến tính LDA.

2.1.1.1. Đặc trưng MFCC

Đây là một trong những loại đặc trưng được sử dụng phổ biến trong nhận dạng tiếng nói. Ý tưởng chính của MFCC tính toán các giá trị phổ của tín hiệu cho băng tần trên miền tần số mà tai người dễ cảm thụ nhất. Sơ đồ khối các bước để tính toán đặc trưng MFCC trên tín hiệu tiếng nói đầu vào được trình bày ở hình 2-3 [Jurafsky 2008].



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**: Do tai người chỉ nhạy cảm với các tần số thấp nên một hàm tăng cường tín hiệu theo công thức (2.1) cho các tần số cao được áp dụng trước khi tín hiệu được đưa vào tính toán ở các bước sau.

  $$s(n)=x(n) - a*x(n-1) \quad (2.1)$$

  Trong đó *x(n)* là tín hiệu vào, *a* là hệ số (trong luận án này *a*=0.95)
- **Window:** Tạo các khung tín hiệu gọi là cửa sổ. Tín hiệu tiếng nói là loại tín hiệu liên tục và biến đổi theo thời gian. Tuy nhiên trong một khoảng thời gian ngắn từ 10ms đến 30ms có thể được coi là ổn định. Đối với các hệ thống nhận dạng từ vựng lớn phát âm liên tục thì đơn vị nhận dạng thường là một âm vị và độ dài phát âm của một âm vị cũng thường nằm trong khoảng thời gian này. Vì thế thay vì ta đi tính toán đặc trưng trên toàn bộ một phát âm thì ta chỉ tính toán trên từng khung cửa sổ (Window) có độ dài từ 10ms đến 30ms. Để không bị mất thông tin giữa hai khung liên tiếp thì các cửa sổ thường được xếp chồng lên nhau với khoảng cách từ 10ms đến 20ms. Hình 2-3 minh họa quá trình phân chia cửa sổ cho một tín hiệu tiếng nói với kích

thước cửa sổ là 25ms và khoảng cách giữa hai khung (độ dịch khung) là 10ms. Hàm cửa sổ áp lên mỗi khung thường là hàm Hamming với công thức sau:

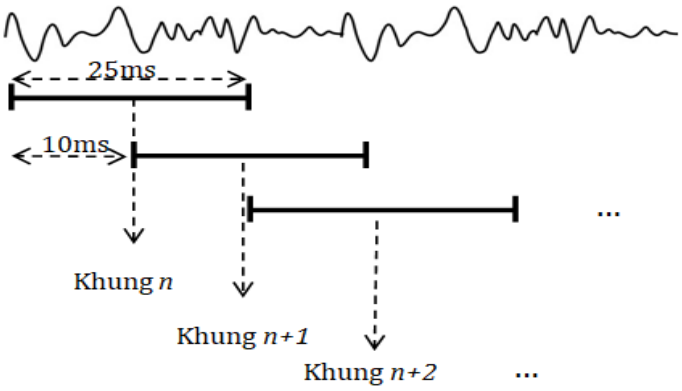
$$W(n)=\{0.54 - 0.46 \cos(\frac{2\pi n}{L})\} \quad (2.2)$$

Khi đó giá trị của tín hiệu sau khi áp dụng hàm cửa sổ là: $y(n)=W(n)s(n)$.

Trong đó L là kích thước của cửa sổ, $0 \le n \le L$, $s(n)$ giá trị của tín hiệu ở miền thời gian tại thời điểm n.

- **DFT:** Biến đổi Fourier rời rạc. Biến đổi DFT được áp dụng để trích chọn thông tin về phổ của tín hiệu đầu vào. Biến đổi này được thực hiện trên mỗi một khung đã được lấy qua hàm cửa sổ. Tính toán DFT được mô tả ở công thức (2.3).

$$X(k)= \sum_{n=0}^{L-1} y[n] e^{-j2\frac{\pi}{L}kn} \quad (2.3)$$



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:** Lọc và biến đổi sang tần số Mel. Tần số âm thanh thường dao động trong khoảng dưới 10kHz, tuy nhiên tai người chỉ nhạy cảm hay nghe rõ nhất trong khoảng 1kHz. Các hệ thống nhận dạng cố gắng mô phỏng

lại cách thức nghe của con người vì thế vấn đề đặt ra là cần biến đổi tín hiệu từ miền tần số Hz sang miền tần số mà con người dễ nghe nhất. Miền tần số này gọi là Mel (được đặt đề xuất bởi Steven and Volkmann, 1940). Công thức biến đổi được mô tả ở công thức (2.4).

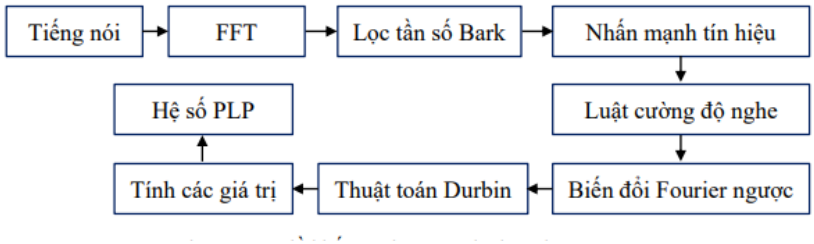
$$\text{mel(f)} = 2595\log_{10}(1+\frac{f}{700}) \quad (2.4)$$

Các bộ lọc băng tần được thiết kế trên miền tần số Mel này.

- **Logarithm (log) và biến đổi Cosine rời rạc (DCT)**: Hàm logarithm được áp dụng trên các giá trị DFT đo độ thính của tai người theo hàm logarithm, vì vậy việc áp dụng hàm log để đưa đặc trưng tính toán được gần giống với tín hiệu mà tai người nghe. Đồng thời việc sử dụng hàm log giúp cho đặc trưng tính toán ít bị ảnh hưởng bởi sự biến đổi ngẫu nhiên ở tín hiệu đầu vào. Sau đó các giá trị logarithm này được áp dụng hàm biến đổi Fourier ngược (hoặc có thể dùng công thức biến đổi Cosine rời rạc) như công thức (2.5) để thu được các giá trị MFCC.

$$\text{C[k]} = \sum_{n=0}^{L-1} \log(|X[k]|) e^{j\frac{2\pi}{L}kn} \quad (2.5)$$

Ngoài phương pháp trích chọn đặc trưng MFCC được trình bày ở trên, ta cũng có thể sử dụng phương pháp tính toán đặc trưng PLP dựa trên cơ sở phương pháp mã dự báo tuyến tính LPC (Linear Prediction Coding). Đặc trưng này được tạo ra dựa trên đặc tính vật lý của tai người khi nghe [H. Hermansky 1990]. Hình 2-5 miêu tả các bước xử lý tính toán PLP.



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

Trong đó:

- **Biến đổi Fourier nhanh (FFT):** Tương tự như phương pháp MFCC, tín hiệu tiếng nói được chia thành các khungvà được chuyển sang miền tần số bằng thuật toán FFT..
- **Lọc theo thang tần số Bark:** Tín hiệu tiếng nói được lọc qua các bộ lọc phân bố theo thang tần số phi tuyến, trong trường hợp này là thang tần số Bark theo công thức (2.6).

$$Bark(f) = 6\ln\{\frac{f}{1200} + [(\frac{f}{1200})^2 + 1]^{0,5}\} \tag{2.6}$$

- **Nhấn mạnh tín hiệu:** Dùng hàm cân bằng độ ồn (equal-loudnes). Bước này tương tự bước nhấn mạnh (preemphais) của phương pháp MFCC. Hàm này mô phỏng đường cong cân bằng độ ồn (Equal-Loudnes Curve) như công thức (2.7).

$$E(\omega) = \frac{(\omega^2 + 56,8*10^6)\omega^4}{(\omega^2 + 6,3*10^6)(\omega^2 + 0,38*10^9)(\omega^6 + 9,58*10^{26})} \tag{2.7}$$

- **Dùng luật cường độ nghe (Power Law of Hearing):** Dùng một phép ánh xạ phi tuyến để làm tăng đặc tính năng lượng của tín hiệu tương đồng với cách thức mà tai nghe âm thanh. Phép ánh xạ này mô tả ở công thức (2.8).

$$\Phi(f) = \Psi(f)^{0,33} \tag{2.8}$$

- **Biến đổi Fourier ngược (Inverse DFT):** Các hệ số tự tương quan được biến đổi Fourier ngược là giá trị đầu vào cho LPC.
- **LPC**: Thuật toán tính toán các hệ số dự báo tuyến tính theo thuật toán Levinson-Durbin [Levinson 1947].
- **Thuật toán Durbin:** Thuật toán Durbin được sử dụng để tính các hệ số dự báo tuyến tính như phương pháp LPC.
- **Tính các giá trị delta:** Phương pháp tính tương tự như phương pháp hệ số MFCC.

2.1.2. Kỹ thuật khử nhiễu CMS

Đây là một kỹ thuật thông dụng để khử nhiễu trong các hệ thống nhận dạng, được dùng kết hợp trong quá trình tính toán các đặc tính phổ của tiếng nói. Phương pháp này dựa trên giả thiết là các đặc tính tần số của môi trường là thường xuyên cố định hoặc biến đổi chậm. Các tham số cepstral của một phát âm được trừ đi giá trị trung bình của các tham số trong một khoảng thời gian nào đó và làm cho các giá trị này ít bị ảnh hưởng bởi môi trường:

$$\hat{O}(\tau) = O(\tau) - \frac{1}{T}\sum_{t=1}^{T} O(t) \tag{2.9}$$

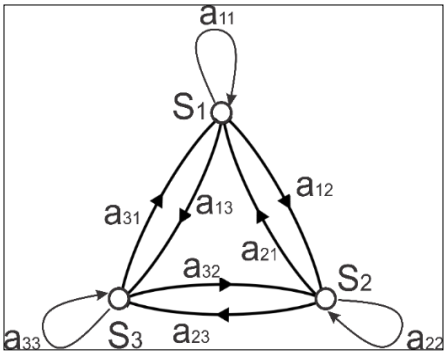
trong đó T là độ dài của vùng lấy giá trị trung bình, thường là độ dài của cả phát âm.

Kỹ thuật CMS có ưu điểm là đơn giản, thời gian tính toán nhanh, dễ áp dụng, khi áp dụng kỹ thuật khử nhiễu này vào nhận dạng tiếng nói, cần lưu ý đến tốc độ xử lí và bảo tồn các đặc trưng âm học của phụ âm, đặc biệt là các phụ âm vô thanh. Để đảm bảo thực hiện được trong thời gian thực, hiện nay, người ta thường áp dụng mô hình tham số thích nghi với nhiễu. Cụ thể như sau: Khi huấn luyện tham số, người ta lấy một mẫu sạch, không bị nhiễu, để huấn luyện, sau đó, người ta lấy các mẫu sạch này trộn với các loại nhiễu sinh bởi các mô hình toán học khác nhau và tham số mô hình sẽ được biến đổi bởi mẫu nhiễu nhờ các công cụ mô hình. Do đó, trong giai đoạn nhận dạng, khi tín hiệu thực được đưa vào hệ thống, người ta sẽ tính thẳng các đặc trưng và quyết định từ chính tín hiệu chứ không cần lọc.

2.2. Tổng quan về mô hình Markov ẩn HMM

2.2.1. Chuỗi Markov

Là dãy gồm N trạng thái $S_1, S_2, \ldots S_n$ 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]

Theo đó, ta có ma trận xác suất chuyển tiếp trạng thái có dạng:

$$A = \{a_{ij}\} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad (2.10)$$

Với tổng các xác suất hàng ngang của ma trận bằng 1.

Công thức tính xác suất tổng quát của một chuỗi các trạng thái:

$$\begin{aligned} &P(s_{i1}, s_{i2}, \ldots, s_{ik}) = P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) \\ &= P(s_{ik} \mid s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) = \ldots \\ &= P(s_{ik} \mid s_{ik-1}) P(s_{ik-1} \mid s_{ik-2}) \ldots P(s_{i2} \mid s_{i1}) P(s_{i1}) \end{aligned} \quad (2.11)$$

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:

- $O = \{o_1, o_2, \ldots, o_T\}$ là tập các vector quan sát.
- $S = \{s_1, s_2, \ldots, s_N\}$ là tập hữu hạn các trạng thái s gồm N phần tử.
- $A = \{a_{11}, a_{12}, \ldots, a_{nn}\}$ 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 $\sum a_{ij} = 1$, $\forall i\ j{=}k$
- $B = \{b_{2t}, b_{it}, \ldots, b_{(N-1)t}\}$ 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)

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

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 s_i .

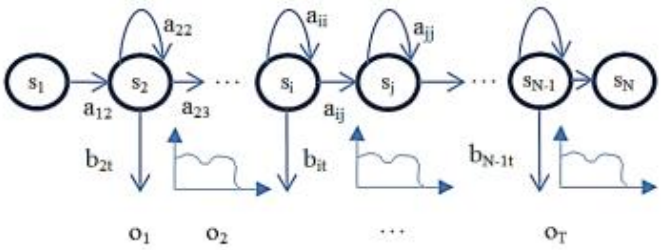
- $\Pi = \{\pi_i\}$ là tập xác suất trạng thái đầu, với $\pi_i = P(q_1 = s_i)$ với i=1..N là xác suất để trạng thái s_i là trạng thái đầu q_1.

Như vậy một cách tổng quát một mô hình HMM λ có thể được biểu điễn bởi λ=(A,B,Π). Trong lĩnh vực nhận dạng các mô hình HMM được áp dụng với hai giả thiết sau:

- Một là giả thiết về tính độc lập, tức không có mỗi liên hệ nào giữa hai quan sát lân cận nhau o_i và o_{i+1}, khi đó xác suất của một chuỗi các quan sát O={o_i} có thể được xác định thông qua xác suất của từng quan sát oi như sau: $P(O) = \prod_{i=t}^{T} P(o_i)$.

- Hai là giả thiết Markov, xác suất chuyển thành trạng thái s_t chỉ phụ thuộc vào trạng thái trước nó s_{t-1}.

Trong nghiên cứu này chúng tôi sử dụng mô hình HMM-GMM có cấu trúc dạng Left-Right liên kết không đầy đủ được minh hoạ như Hình 2-7



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 = (S_1, S_2, …, S_n) và trạng thái quan sát được tại thời điểm t là q_t.

b. M: Số hiện tượng quan sát được của mỗi trạng thái, ký hiệu hiện tượng quan sát được là V = {V_1, V_2, …,V_m}, tín hiệu quan sát được ở thời điểm t là Ot.

c. Xác suất chuyển tiếp trạng thái biểu diễn bởi ma trận A = {a_{ij}} 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: Π ={π_i}, với π_i là trạng thái S_i 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

Mô hình hỗn hợp Gauss (Gaussian Mixture Model - GMM) là mô hình dạng thống kê được xây dựng từ việc huấn luyện thông qua thông qua dữ liệu học thường được sử dụng trong các nghiên cứu về nhận dạng. GMM là hàm mật độ xác suất sinh quan sát của trạng thái (state output probability density function - pdf) trong mô hình HMM thường được ứng dụng trong thực tế. Trong mô hình HMM có N trạng thái thì mỗi trạng thái có một hàm mật độ xác suất mô tả xác suất để có quan sát O (observations) tại thời điểm t khi đang ở trạng thái j.

Hàm mật độ xác suất phân bố Gauss có dạng:

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

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

trong đó: X là véc tơ dữ liệu chứa các tham số của đối tượng cần biểu diễn.

π_i, i=1,…,M là trọng số của hỗn hợp.

μ_I: véc tơ trung bình của véc tơ D chiều.

$\sum_\iota$: ma trận hiệp phương sai kích thước DxD (thường chọn là ma trận đường chéo để giảm số tham số của mô hình)

Một mô hình hỗn hợp Gauss đa biến có tổng số M thành phần mật độ Gauss có dạng:

$$p(X \mid \lambda) = \sum_{i=1}^{M} \pi_i g_i(X \mid \mu_i, \Sigma_i) \tag{2.17}$$

Một mô hình GMM đầy đủ tham số hóa bởi véc tơ trung bình, ma trận hiệp phương sai và các trọng số hỗn hợp từ các thành phần hợp Gauss được biểu diễn gọn lại như sau:

$$\lambda = \{\pi_i, \mu_i, \sum_i\}, i = 1, 2, ..., M \quad (2.18)$$

Một quan sát O trong nhận dạng tiếng nói thường là một véc tơ gồm các hệ số rút trích MFCC tĩnh (static feature) và các hệ số động (dynamic features).

Giả thiết T là số lượng véc tơ đặc trưng MFCC của tín hiệu tiếng nói, M là số thành phần Gauss:

$$X = \{x_1, x_2, ..., x_T\} \quad (2.19)$$

Tương đồng với GMM sẽ là: $p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda)$

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

Với dãy quan sát $O = \{o_1, o_2, o_3, ...\}$ và mô hình Markov ẩn $\lambda = (A, B, \pi)$ đã được huấn luyện, chúng ta cần tính xác suất $P(O/\lambda)$.

Chúng ta có dãy quan sát $O = \{o_1, o_2, o_3, ...\}$ và mô hình Markov ẩn $\lambda = (A, B, \pi)$, chúng ta cần tính xác suất $P(O/\lambda)$.

Giả sử dãy quan sát có độ dài là T, vậy ta có một dãy các trạng thái tương ứng của mô hình Markov ẩn sinh ra nó : $q = q_1, q_2, q_3, ... q_T$. Ta có xác suất để dãy quan sát O được sinh ra bởi λ với dãy trạng thái q là:

$$P(O/q, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda)$$

với giả thiết các O_i là độc lập ta có:

$$P(O/q, \lambda) = b_{q_1}(o_1)\, b_{q_2}(o_2) ... b_{q_T}(o_T)$$

Mặt khác ta xác suất của dãy trạng thái q đối với mô hình λ là :

$$P(q/\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T}$$

Từ đó ta có xác suất của dãy quan sát *O*, đối với mô hình λ và dãy trạng thái *q* là :

$P(O, q|\lambda) = P(O/q, \lambda)\, P(q/\lambda)$

$$= \pi_{q_1} b_{q_1}(o_1)\, a_{q_1 q_2} b_{q_2}(o_2)\, a_{q_2 q_3} \ldots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

Xác suất của dãy quan sát *O* đối với mô hình λ sẽ là tổng của tất cả các xác suất dãy quan sát *O* đối với mô hình λ với mọi dãy trạng thái *q* có thể có :

$$P(O|\lambda) = \sum_q P(O,q|\lambda) = \sum_Q \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots a_{q_{T-1} q_T} b_{q_T}(o_T) \qquad (2.20)$$

Độ phức tạp tính toán của công thức (2.20) là $2T.\, N^T$, bởi vì tại mỗi thời điểm *t*, có *N* khả năng chuyển trạng thái, trong mỗi trạng thái có $(2T-1).N^T$ phép tính nhân, $N^T - 1$ phép tính cộng. Trong thực tế công thức này không thể thực hiện được do độ phức tạp quá lớn. Để khắc phục vấn đề này thuật toán tiến-lùi (forward-backward algorithms) được dùng để tính xác suất dãy quan sát *O* đối với mô hình λ.

Ta định nghĩa biến tiến (forward) $\alpha_t(i)$ là xác suất của dãy quan sát *O* tới thời điểm *t*: $O = o_1, o_2, \ldots, o_t$ tại trạng thái S_i được sinh bởi mô hình λ.

$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = S_i/\lambda)$

với giá trị khởi tạo $\alpha_1(i) = \pi_i b_i(o_1)$, $1 \le i \le N$

Các $\alpha_t(i)$ được tính bằng thuật toán đệ qui được miêu tả như sau:

1) Khởi tạo

$\alpha_1(i) = \pi_i b_i(o_1) \qquad 1 \le i \le N$

2) Tính các $\alpha_{t+1}(j)$ bằng phương pháp đệ qui

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}) \qquad (2.21)$$

$1 \le t \le T-1,\ 1 \le j \le N$

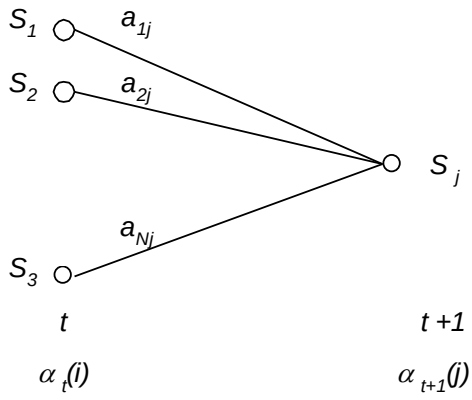
3) Kết thúc

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Bước 1) khởi tạo các $\alpha_1(i)$ tương ứng với các trạng thái i khác nhau. Tại bước 2 các $\alpha_{t+1}(j)$ được tính theo phương pháp đệ quy dựa vào các $\alpha_t(j)$ được tính trước đó. Hình 2-8 miêu tả các phép toán cần thiết để tính các $\alpha_t(j)$. Trạng thái của mô hình tại thời điểm $t+1$ là S_j và có tất cả N khả năng để dẫn tới trạng thái S_j từ các trạng thái S_i với xác suất là $\alpha_t(i)$. Xác suất này nhân với xác suất chuyển trạng thái a_{ij} sẽ cho ta xác suất mô hình ở trạng thái S_j tại thời điểm $t+1$ với điều kiện mô hình ở trạng thái S_i tại thời điểm t. Tổng các xác suất này với các i từ 1 đến N cho ta xác suất để mô hình ở trạng S_j tại thời điểm $t+1$, xác suất này nhân với xác suất $b_j(o_{t+1})$ sẽ cho ta xác suất dãy quan sát O tới thời điểm $t+1$ ở trạng thái S_j và đó chính là $\alpha_{t+1}(j)$.

Thuật toán quy nạp sẽ dừng ở bước 3) khi mà $t = T$. Khi đó tổng của các $\alpha_t(i)$ với i từ 1 đến N sẽ cho ta xác suất của dãy quan sát O đối với mô hình λ: $P(O/\lambda)$.

Độ phức tạp tính toán với cách tính theo các biến forward $\alpha_t(i)$ là N^2T , thấp hơn so với độ phức tạp $2T.N^T$ của bước trước.



Hình 2-8: Miêu tả các dãy phép toán được thực hiện để tính biến $\alpha_t(i)$ [4]

Biến lùi (backward) $\beta_t(j)$ được định nghĩa là xác suất của dãy O từ thời điểm $t+1$ đến T : $O = \{o_{t+1}, o_{t+2}, ..., o_T\}$, với điều kiện là mô hình ở trạng thái S_i tại thời điểm t .

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, ...,oO_T / q_t = S_i, \lambda), \quad 1 \le t \le T$$

Thuật toán tính biến lùi $\beta_t(i)$ cũng dựa trên phương pháp đệ qui giống như trường hợp của biến tiến $\alpha_t(i)$:

1) Khởi tạo

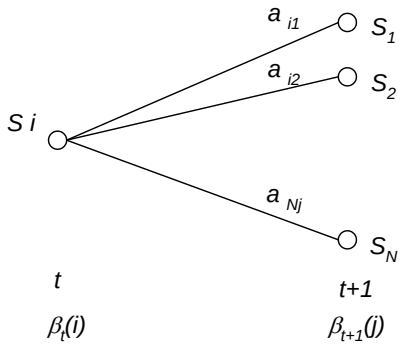
$\beta_T(i) = 1, \qquad 1 \le i \le N$

2) Tính các $\beta_t(j)$ bằng phương pháp đệ qui

$$\beta_t(j) = \sum_{i=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

$$t = T-1, T-2, ..., 1, \qquad 1 \le i \le T$$

Bước 1) khởi tạo các $\beta_T(i)$ bằng 1 cho tất cả các i. Các tính toán của bước 2) được mô tả trong Hình 2-9, trong đó $\beta_t(i)$ được tính dựa vào các $\beta_{t+1}(j)$ được tính trước đó. Trạng thái của mô hình tại thời điểm t là S_i và có tất cả N khả năng để dẫn tới trạng thái S_i từ các trạng thái S_j tại thời điểm $t+1$ với xác suất là $\beta_{t+1}(j)$. Xác suất này nhân với xác suất chuyển trạng thái a_{ij}, kết hợp với xác suất quan sát o_{t+1} tại trạng thái j sẽ cho ta xác suất mô hình ở trạng thái S_i tại thời điểm t với điều kiện mô hình ở trạng thái S_j tại thời điểm $t+1$. Tổng các xác suất này với các j từ 1 đến N cho ta xác suất để mô hình ở trạng S_i tại thời điểm t và đó là $\beta_t(i)$.



Hình 2-9: Miêu tả các dãy phép toán được thực hiện để tính biến $\beta_t(i)$ [4]

Bằng thuật toán tiến-lùi ta có thể tính xác suất $P(O/\lambda)$ như sau:

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i)$$

2.3.2. Bài toán giải mã

Với dãy quan sát $O= \{o_1, o_2, o_3, ...\}$ và mô hình Markov ẩn $\lambda=(A, B, \pi)$ làm thế nào chúng ta có thể tìm được dãy trạng thái tương ứng $q=\{q_1, q_2, q_3, ...\}$ tối ưu nhất theo một tiêu chuẩn nào đó.

Trong bài toán này, ta phải tìm dãy trạng thái $q=(q_1, q_2, ..., q_T)$ tối ưu tương ứng với một dãy quan sát $O=(o_1, o_2, ..., o_T)$ và mô hình $\lambda=(A, B, \pi)$ cho trước, để cho $P(O, S/\lambda)$ là lớn nhất.

Một phương pháp thông dụng hay được dùng để giải quyết bài toán này là dùng thuật toán tìm kiếm Viterbi. Đây là thuật toán dựa trên phương pháp lập trình động (Dynamic Programing Method) để tìm ra một dãy các trạng thái tối ưu duy nhất.

Thuật toán Viterbi:

Ta định nghĩa biến $\delta_t(i) = \max_{q_1,q_2,...q_{t-1},} P(q_1,q_2,...,q_t=i,o_1,o_2,...,o_t|\lambda)$ là biến có điểm cao nhất (best score) tại thời điểm t tương ứng với dãy trạng thái $q_1, q_2, ..., q_{t-1}$, kết thúc tại trạng thái S_i.

Các biến $\delta_t(i)$ được tính bằng phương pháp đệ qui dựa trên các tính toán trước đó

$$\delta_{t+1}(i)=[\max_{1\le i\le N}(\delta_t(i)a_{ij}]b_j(o_k)$$

Để lưu vết các trạng thái của dãy ta dùng mảng $\delta_t(i)$, khi thuật toán kết thúc các phần tử trong mảng chính là các trạng thái của dãy q cần tìm. Sau đây chi tiết thuật toán Viterbi

1) Khởi tạo

$$\delta_t(i) = \pi_i b_i(o_1), \quad 1 \le i \le N$$

$$\delta_1(i)=0$$

2) Tính toán đệ qui

$$\delta_{t+1}(i) = [\max_{1\le i\le N}(\delta_t(i).a_{ij}].b_j(o_k), \quad 2 \le t \le, \ 1 \le i \le N \qquad (2.22)$$

$$\psi_1(i) = \arg\max_{1\le i\le N}[\delta_{t-1}(i).a_{ij}] \qquad 2 \le t \le T, \ 1 \le i \le N$$

3) Kết thúc

$$P^* = \max_{1 \le i \le N}[\delta_T(i)]$$

$$q^*_T = \arg\max_{1 \le i \le N}[\delta_T(i).a_{ij}]$$

Truy hồi các trạng thái

$q^*_T = \delta_{t+1}(q^*_{t+1}), \quad t = T-1, T-2, \ldots, 1$

Kết thúc thuật toán các q^*_t chính là các trạng thái của dãy cần tìm.

Thuật toán Viterbi gần giống như thuật toán tính biến tiến $\alpha_t(i)$. Điểm khác nhau cơ bản giữa hai thuật toán này là công thức tính max (2.22) được dùng thay cho công thức tính tổng (2.21).

2.3.3. Bài toán huấn luyện

Làm thế nào chúng ta điều chỉnh các tham số A, B, π để có được xác suất $P(O/\lambda)$ lớn nhất

Đây là bài toán khó khăn nhất của mô hình Markov ẩn, chúng ta phải điều chỉnh bộ ba các tham số (A, B, π) để xác suất $P(O/\lambda)$ là lớn nhất. Trên thực tế không tồn tại một phương pháp thực sự tối ưu để $P(O/\lambda)$ là lớn nhất. Giải pháp cho bài toán này là thuật toán điều chỉnh tham số (re-estimation) Baum-Welch.

Ta định nghĩa biến $\gamma_t(i) = P(q_t = S_i / O, \lambda)$ là xác suất để mô hình ở trạng thái S_i vào thời điểm t với dãy quan sát O và mô hình λ đã cho. Với định nghĩa trên, biến $\gamma_t(i)$ được biểu diễn thông qua hai biến tiến và lùi như sau:

$$\gamma_t(i) = \frac{P(q_t = S_i, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \qquad (2.23)$$

Trong công thức trên $\alpha_t(i)$ là xác suất của dãy qua sát $o_1, o_2, \ldots, o_t$ và $\beta_t(i)$ là xác suất của dãy $o_{t+1}, o_{t+2}, \ldots, o_T$ với mô hình ở trạng thái S_i vào thời điểm t.

Từ công thức (2.23) ta rút ra được

$$\sum_{i=1}^{N}\gamma_t(i) = 1$$

Với $\gamma_t(i)$ ta có thể tìm được tại thời điểm t xác suất lớn nhất của dãy $o_1, o_2, \ldots, o_t$ là :

q_t = argmax[$\gamma_t(i)$], $1 \le i \le N$, $1 \le t \le T$

Ta định nghĩa biến $\xi_t(i, j)$ là xác suất mô hình ở trạng thái S_i tại thời điểm t và ở trạng thái S_j tại thời điểm t+1 với mô hình λ và dãy quan sát O cho trước.

$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j / O, \lambda)$

Hình 2-10 miêu tả mối quan hệ chuyển dịch giữa các trạng thái S_i và S_j. Từ định nghĩa các biến tiến lùi $\alpha_t(i)$ và $\beta_t(i)$ ta có

$$\xi_t(i, j) = \frac{P(q_t = S_i, q_{t+1} = S_j, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}$$

$$\gamma_t(i) = P(q_t = S_i / O, \lambda) = \frac{P(q_t = S_i, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

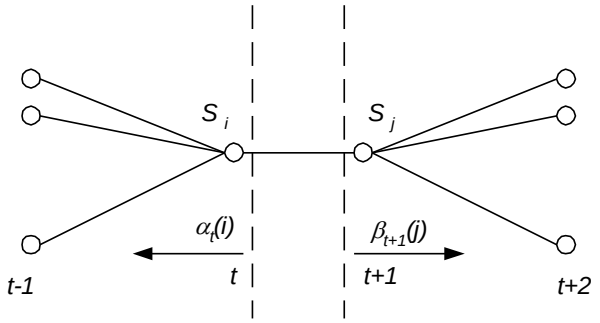
Với các định nghĩa trên ta có

$$\gamma_t(i) = \sum_{j=1}^{N} \varepsilon_t(i, j)$$

Từ định nghĩa công thức trên ta có thể nhận thấy:

$\sum_{t=1}^{T-1} \xi_t(i, j)$ = là khả năng để mô hình chuyển trạng thái từ S_i sang S_j

$\sum_{t=1}^{T-1} \gamma_t(i)$ = là khả năng để mô hình chuyển trạng thái từ S_i



Hình 2-10: Miêu tả các phép tính cần thiết để tính $\xi_t(i, j)$ [4]

Từ các quan sát trên ta có tập các công thức dùng để điều chỉnh (re-estimation) các tham số của mô hình Markov ẩn như sau :

$\bar{\pi}_i =$ khả năng mô hình ở trạng thái S_i tại thời điểm $(t=1) = \gamma_1(i)$ (2.24)

$\bar{a}_{ij} =$khả năng chuyển từ trạng thái S_i sang trạng thái S_j / khả năng chuyển từ trạng thái S_i

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (2.25)$$

$\bar{b}_j(v_k) =$ khả năng ở tại trạng thái S_i với ký hiệu quan sát v_k / khả năng ở tại trạng thái S_i

$$= \frac{\sum_{t=1, O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad (2.26)$$

Với một mô hình $\lambda=(A, B, \pi)$ đầu tiên chúng ta dùng các công thức (2.24), (2.25), (2.26) để tính toán bộ tham số mới $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$. Người ta đã chứng minh được rằng:

Hoặc là mô hình khởi điểm λ được định nghĩa chính xác là mô hình hội tụ và do đó $\lambda = \bar{\lambda}$.

Hoặc là mô hình mới có $P(O/\bar{\lambda}) > P(O/\lambda)$

Dựa vào chứng minh này chúng ta dùng $\bar{\lambda}$ thay thế cho λ là lặp lại các tính toán **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.** ta sẽ cải thiện được xác suất $P(O/\lambda)$ cho tới thời điểm thuật toán hội tụ.

Trong quá trình tính toán, sau mỗi lần lặp các biểu thức sau đây luôn được thoả mãn :

$$\sum_{i=1}^{N} \bar{\pi}_i = 1 \qquad 1 \le i \le N$$

$$\sum_{j=1}^{N} \bar{a}_{ij} = 1 \qquad 1 \le i \le N$$

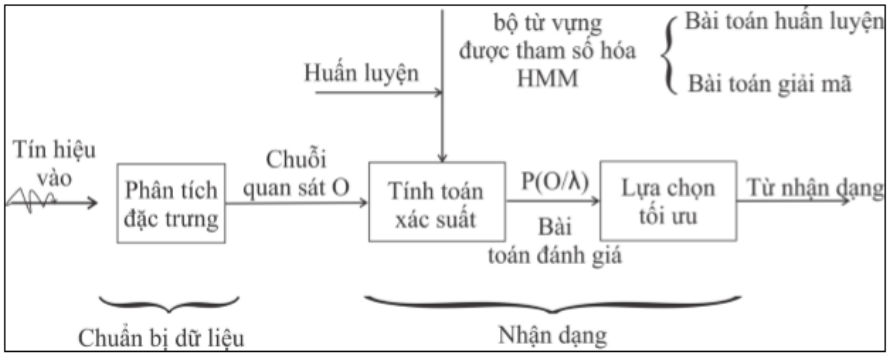
$$\sum_{k=1}^{N} \bar{b}_j(k) = 1 \qquad 1 \le j \le N$$

Các công thức (2.24), (2.25), (2.26) được gọi là công thức hiệu chỉnh lại tham số Baum-Welch và được chứng minh bằng thuật toán EM được trình bày trong [12].

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

2.4.1. Tổng quan

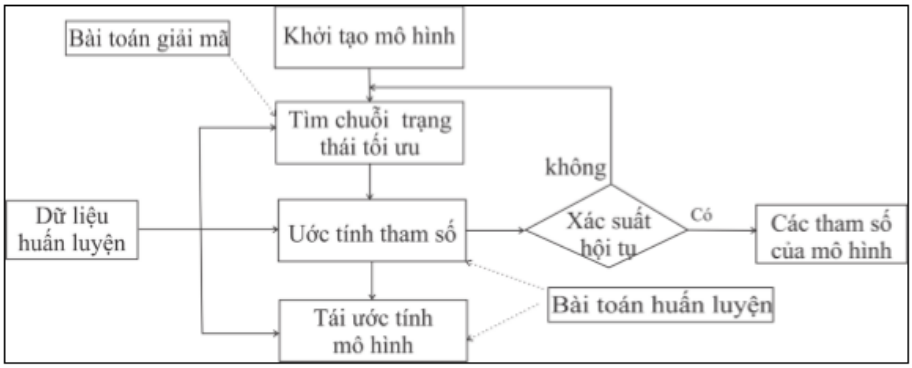
Giả sử ta có bộ từ vựng gồm V từ cần được nhận dạng, mỗi từ được mô hình bằng một HMM. Việc huấn luyện mô hình HMM cho mỗi từ cần có một bộ dữ liệu huấn luyện gồm K dữ liệu, mà mỗi dữ liệu đưa vào là chuỗi các vector đặc trưng của tiếng nói hay còn gọi là chuỗi quan sát đối với mô hình HMM. K càng lớn thì mô hình càng được huấn luyện đầy đủ.



Hình 2-11: Ứng dụng các bài toán trong nhận dạng từ rời rạc

2.4.2. Giai đoạn huấn luyện mô hình

Đầu tiên các tham số của mô hình được khởi tạo. Với một dữ liệu huấn luyện O ta dùng bài toán giải mã để phân lớp trạng thái (nghĩa là tìm chuỗi trạng thái tối ưu ứng với chuỗi quan sát cho trước), sau đó dùng bài toán huấn luyện để ước tính tham số và tính toán lại các tham số của mô hình HMM λ tương ứng. Quá trình giải mã và tính toán lại bộ tham số được lặp lại nhiều lần cho đến khi xác suất $P(O|\lambda)$ hội tụ thì sẽ thu được các tham số tối ưu của mô hình.



Hình 2-12: Các bước huấn luyện bằng HMM

2.4.3. Giai đoạn nhận dạng

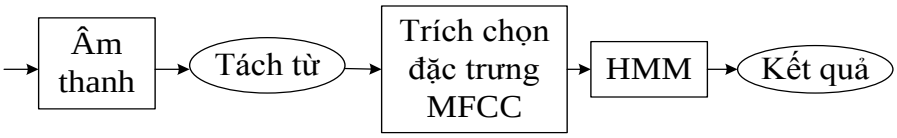
Tín hiệu tiếng nói cần nhận dạng được trích xuất vector đặc trưng, gọi là chuỗi quan sát O. Sau đó cần giải quyết bài toán đánh giá để tính V xác suất $P(O|\lambda_i)$ của V từ trong bộ từ vựng và chọn ra mô hình mô tả đúng nhất tín hiệu tiếng nói đưa vào, đó là mô hình λ_i có xác suất $P(O|\lambda_i)$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

2.5. Kết luận

Qua nội dung chương 2 ta đã nắm được giai đoạn đầu của hệ thống nhận dạng tiếng nói: các phương pháp xử lý tiếng nói. Các lý thuyết cơ bản về mô hình Markov ẩn, và ứng dụng trong nhận dạng tiếng nói đồng thời thuật toán giải mã trong các hệ thống nhận dạng liên tục cũng đã được đề cập chi tiết.

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

Một hệ thống nhận dạng và chuyển đổi nói chung thường bao gồm hai phần: phần huấn luyện(training phase) và phần nhận dạng (recognition phase). "Huấn luyện" là quá trình hệ thống "học" những mẫu chuẩn được cung cấp bởi những tiếng khác nhau (từ hoặc âm), để từ đó hình thành bộ từ vựng của hệ thống. "Nhận dạng" là quá trình quyết định xem từ nào được đọc căn cứ vào bộ từ vựng đã được huấn luyện. 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

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: Thực hiện việc ghi âm tín hiệu tiếng nói, tách tiếng nói khỏi nền nhiễu và lưu vào cơ sở dữ liệu.
- Trích chọn đặc trưng MFCC: Trích đặc trưng tín hiệu tiếng nói đã thu ở quá trình thứ nhất bằng phương pháp MFCC, đồng thời thực hiện ước lượng vector các vector đặc trưng này.
- Quá trình thứ ba: Xây dựng mô hình Markov ẩn với 6 trạng thái, tối ưu hóa các hệ số của HMM tương ứng với từng từ trong bộ từ vựng, tiến hành nhận dạng một từ được đọc vào micro và hiển thị kết quả.

Chi tiết các quá trình trên như sau:

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

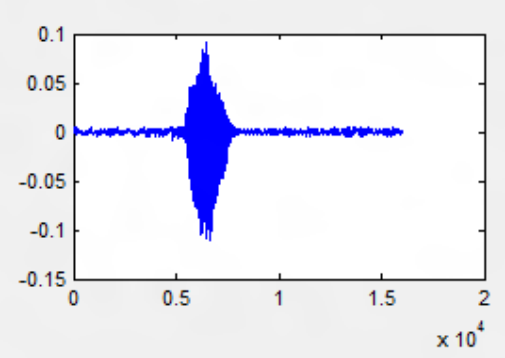
Thu thập và tiền xử lí tín hiệu tiếng nói ở giai đoạn huấn luyện được thực hiện bằng phương pháp thủ công là thu tín hiệu từ micro, dùng kỹ thuật xử lý đầu

cuối để phát hiện phần tín hiệu tiếng nói và phần tín hiệu nhiễu. Từ đó ta có thể tách tiếng nói ra khỏi nền nhiễu (chỉ thu tín hiệu tiếng nói mà không thu tín hiệu nhiễu nền).

Quá trình thu âm và tiền xử lý, chuẩn bị dữ liệu cho huấn luyện thực hiện như sau:

Bước 1: Thu âm từ micro (mặc định thu âm ở tần số 8000Hz trong thời gian khoảng 2s).

Ví dụ: Kết quả mẫu âm thanh thu được ở tần số 8000Hz trong thời gian khoảng 2s với từ 'hai':



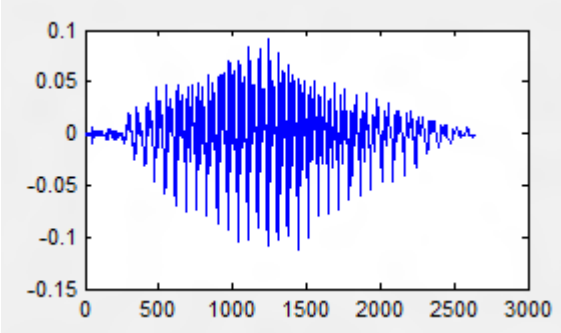
Hình 3-2: Từ 'hai' được thu âm – bao gồm nền nhiễu

Bước 2:

- Tiến hành chia mẫu âm thanh thu được thành các frame với kích thước mỗi frame khoảng 10ms.
- Kiểm tra ngưỡng của mỗi farm:
  - Nếu ngưỡng của frame nhỏ hơn hoặc bằng ngưỡng nền nhiễu: bỏ qua – **(xóa)**
  - Ngược lại, nếu ngưỡng của frame lớn hơn ngưỡng nền nhiễu → frame này có chứa tín hiệu tiếng nói → **giữ lại**

Bước 3: Lưu lại mẫu âm thanh chỉ bao gồm tín hiệu tiếng nói (.wav)

Ví dụ: Với từ 'hai' thu được ban đầu có kích thước là: 16000 (2s ở tần số 8000Hz), sau khi loại bỏ nhiễu, kích thước còn lại là: 2840



Hình 3-3: Từ 'hai' sau khi đã loại bỏ nền nhiễu

Với mục đích chuẩn bị bộ dữ liệu để huấn luyện nhận dạng, cơ sở dữ liệu bao gồm các tệp âm thanh lưu ở dạng wav, và các tệp văn bản chứa phiên âm chính tả của các tệp âm thanh. Mỗi tệp âm thanh có một tệp văn bản tương ứng phiên âm chính tả của phát âm. Các phiên âm ở mức âm vị được lưu trong các tệp có đuôi .phn. Các phiên âm ở mức âm vị bao gồm nhiều dòng, mỗi dòng chưa tên âm vị cùng với nhãn thời gian của âm vị đó trong tệp âm thanh.

Đến đây quá trình thu thập và tiền xử lí tín hiệu tiếng nói để xây dựng cơ sở dữ liệu đã hoàn thành nhiệm vụ. Đây là một phần rất quan trọng trong một hệ thống nhận dạng tiếng nói, nó ảnh hưởng rất lớn đến kết quả nhận dạng.

3.2. Trích chọn đặc trưng MFCC

Đến đây chúng ta đã có được các mẫu tiếng nói đã được khử nhiễu. Quá trình thứ hai sẽ thực hiện việc trích đặc trưng các mẫu tiếng nói đã thu ở quá trình thứ nhất. Có nhiều phương pháp trích đặc trưng khác nhau như: wavelets, LPC, MFCC… Ở đây chọn phương pháp MFCC (trích đặc trưng theo thang tần số Mel) do tốc độ tính toán cao, độ tin cậy lớn và đã được sử dụng rất hiệu quả trong các chương trình nhận dạng tiếng nói trên thế giới. Trong khuôn khổ của luận văn này, tác giả đã sử dụng bộ công cụ HTK (Hidden Markov Model Toolkit) phiên bản 3.4.1 để trích chọn đặc trưng MFCC.

- Các cấu hình cố định gồm:
  - Loại tham số phổ: đặc trưng MFCC.
  - Kích thước véc tơ tham số: Số chiều Vector đặc trưng MFCC_D_A_0 là 39 chiều, 13 hệ số tĩnh (MFCC_0), 13 hệ số delta và 13 hệ số acceleration.

MFCC_D_A_0 là 39 = 13 tĩnh (MFCC_0) + 13 hệ số delta +13 hệ số acceleration.Phương pháp CMS được áp dụng để lọc bỏ nhiễu.

- Các cấu hình thay đổi để so sánh hiệu suất hoạt động của hệ thống:
  - Số trạng thái HMM (State): 5 trạng thái.
  - Số phân bố Gauss trong mô hình hỗn hợp Gauss cho từng trạng thái HMM: từ 1 đến 5.

Việc thay đổi 2 tham số này liên quan đến độ phức tạp của mô hình HMM: số trạng thái HMM và số phân bố Gauss càng tăng thì mô hình càng phức tạp (càng có nhiều tham số).

Cấu hình các giá trị của thuộc tính MFCC trong HMM:

```
~o
<STREAMINFO> 1 39
<VECSIZE> 39<NULLD><MFCC_D_A_0><DIAGC>
~h "proto"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 39
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
<VARIANCE> 39
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
<GCONST> 0
<STATE> 3
<MEAN> 39
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
<VARIANCE> 39
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
<GCONST> 0
<STATE> 4
<MEAN> 39
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
<VARIANCE> 39
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
<GCONST> 0
<TRANSP> 5
 0.0 1.0 0.0 0.0 0.0
 0.0 0.6 0.4 0.0 0.0
 0.0 0.0 0.6 0.4 0.0
 0.0 0.0 0.0 0.7 0.3
 0.0 0.0 0.0 0.0 0.0
<ENDHMM>
```

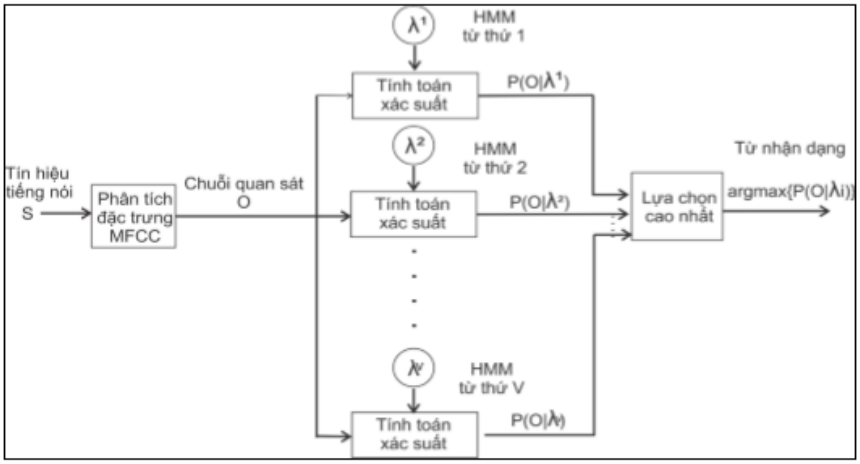
Hình 3- 4: Các giá trị của thuộc tính MFCC

Dòng 1 cho biết số chiều Vector đặc trưng MFCC_0_D_A là 39 = 13 tĩnh (MFCC_0) + 13 hệ số delta +13 hệ số acceleration. Có 5 trạng thái nhưng 2 trạng thái đầu và cuối không xét. Các vector kỳ vọng (mean) và phương sai (variance) đều bằng nhau.

3.3. Nhận dạng bằng mô hình HMM

Sau khi đã thực hiện xong hai quá trình trên thì chúng ta đã có một cơ sở dữ liệu các vector đặc trưng ứng với từng từ. Trong mô đun này chúng ta sẽ xây dựng

một mô hình Markov ẩn với dữ liệu huấn luyện là các vector đặc trưng có được từ quá trình hai. Sơ đồ nhận dạng bằng mô hình HMM được thể hiện như hình 3.4.



Hình 3- 5: Tổng quan mô hình nhận dạng

- Bước 1: Tín hiệu tiếng nói đưa vào được phân tích thành véc tơ đặc trưng MFCC (gọi là chuỗi quan sát O).
- Bước 2: Áp dụng bài toán đánh giá của HMM để tính toán các xác suất P(O| λ_i), là xác suất để mô hình HMM λ_i của từ thứ i trong tập từ vựng sinh ra chuỗi quan sát O.
- Bước 3: Ra quyết định nhận dạng: từ ứng với mô hình HMM có xác suất cao nhất được chọn là kết quả nhận dạng của tín hiệu tiếng nói đầu vào

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: Đối tượng thu âm gồm 158 người, 104 nam và 54 nữ. Tập dữ liệu huấn luyện bao gồm 296 câu, 1686 từ. Mỗi người thu âm 2 set với mỗi set gồm 10 từ phát âm rời rạc từ “không” đến “chín”.

- Dùng để kiểm thử hệ thống: Đối tượng thu âm gồm 38 người, 27 nam và 11 nữ. Tập dữ liệu kiểm tra có 74 câu, 342 từ. Mỗi người thu âm 2 set với mỗi sét gồm 10 từ phát âm rời rạc từ "không" đến "chín".

Để đảm bảo tính khách quan, người nói trong tập dữ liệu kiểm tra là độc lập với người nói trong tập dữ liệu huấn luyện. Môi trường thu âm trong phạm vi văn phòng có nhiễu nhẹ tạp âm như tiếng quạt, gió thổi... Thiết bị thu âm là laptop với micro chuẩn.

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

Phương pháp nhận dạng được sử dụng là phương pháp xây dựng một hệ thống nhận dạng bằng công cụ HTK. Đây là công cụ được sử dụng nhiều trong nhận dạng tiếng nói.

- Các cấu hình cố định gồm:
  - Loại tham số phổ: đặc trưng MFCC.
  - Kích thước véc tơ tham số: 39 chiều (gồm 13 hệ số tĩnh, 13 hệ số delta, 13 hệ số acceleration).
  - Ma trận phương sai: đường chéo (giả sử các chiều độc lập thống kê với nhau).
  - Ngữ cảnh: không phụ thuộc ngữ cảnh (vì là hệ thống nhận dạng từ phát âm rời rạc).
- Các cấu hình thay đổi để so sánh hiệu suất hoạt động của hệ thống:
  - Số trạng thái HMM: 5 trạng thái.
  - Số phân bố Gauss trong mô hình hỗn hợp Gauss cho từng trạng thái HMM: từ 1 đến 5.

Việc thay đổi 2 tham số liên quan đến độ phức tạp của mô hình HMM: Số trạng thái HMM và số phân bố Gauss càng tăng thì mô hình càng phức tạp (càng có nhiều tham số).

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.

---------- Overall Results ----------

SENT: %Correct=13.51 [H=10, S=64, N=74]

WORD: %Corr=77.29, Acc=47.00 [H=245, D=4, S=68, I=96, N=317]

==========

- Thử nghiệm với nhiều hàm Gaussian

Hệ thống của chúng ta làm việc với dữ liệu có độ đa dạng cao, do nhiều người nói, trong môi trường khác nhau, sử dụng các hệ thống điện thoại khác nhau. Một hàm Gaussian không đủ khả năng để mô hình hóa giọng nói của tất cả mọi người trong cơ sở dữ liệu. Một hàm phát xạ quan sát gồm nhiều hàm trộn là hàm Gaussian là cần thiết để nâng cao khả năng nhận dạng của hệ thống.

Trong lần thử nghiệm này 3 hàm Gaussian được sử dụng. Qua kiểm thử tra thử nhận dạng trên dữ liệu kiểm tra, hệ thống bao gồm 3 hàm Gaussian đã cho kết quả cải thiện tốt hơn so với hệ thống chỉ bao gồm một hàm Gaussian:

---------- Overall Results ----------

SENT: %Correct=14.86 [H=11, S=63, N=74]

WORD: %Corr=78.23, Acc=47.95 [H=248, D=3, S=66, I=96, N=317]

==========

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

Trong phần này một hệ thống nhận dạng được xây dựng trên toàn bộ hệ cơ sở dữ liệu. Dữ liệu kiểm tra dùng để đánh giá năng lực của hệ thống cũng chính là dữ liệu được dùng để huấn luyện. Hệ thống mới này sẽ cho một kết quả nhận dạng cao hơn rất nhiều, do các dữ liệu dùng để kiểm tra đã được huấn luyện trước đó.
Kết quả nhận dạng của hệ thống được huấn luyện với toàn bộ cơ sở dữ liệu như sau:

---------- Overall Results ----------

SENT: %Correct=20.27 [H=15, S=59, N=74]

WORD: %Corr=87.07, Acc=59.31 [H=276, D=4, S=37, I=88, N=317]

==========

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 quả này cho chúng ta một tiệm cận trên, một độ chính xác mà hệ thống nhận dạng có thể thực hiện được nếu như nó được cung cấp đầy đủ dữ liệu huấn luyện. Mặt khác hệ cơ sở dữ liệu dùng ở đây là hệ cơ sở dữ liệu có chất lượng kém như đã trình bày ở trên, do đó chúng ta có thể thấy một khả năng có thể xây dựng hệ thống nhận dạng phát âm liên tục có số lượng từ vựng lớn với độ chính xác cao hơn nếu chúng ta có cơ sở dữ liệu với chất lượng tốt.

Sở dĩ có sự khác biệt lớn về độ chính xác so với các hệ thống trước là do các khía cạnh sau:

Các giọng nói được dùng trong tập dữ liệu kiểm tra đã được hệ thống học trước đó. Do đó khi tiến hành nhận dạng hệ thống sẽ cho kết quả với độ chính xác cao hơn khi phải làm việc với giọng nói chưa được học. Khi số lượng người nói lớn và bao gồm các giọng nói đặc trưng bao phủ đại diện cho các giọng nói khác thì khi tiến hành nhận dạng với giọng nói lạ chưa được học, hệ thống vẫn có thể hoạt động cho kết quả tốt. Hệ thống được học với càng nhiều giọng nói thì khả năng nhận dạng của chúng đối với giọng nói lạ càng tốt.

Các từ có mặt trong dữ liệu kiểm tra đề đã được học trước đó. Với các từ có mặt trong dữ liệu kiểm tra nhưng không có mặt trong dữ liệu huấn luyện, để nhận

dạng chúng hệ thống phải tiến hành tổng hợp các âm vị tương ứng với các từ đó từ các âm vị đã được học. Việc tổng hợp này rõ ràng là không chính xác và là một trong các yếu tố làm giảm đáng kể độ chính xác nhận dạng của hệ thống.

3.5. Kết luận

Nội dung chương 3 đã trình bày cụ thể về quá trình xây dựng một hệ thống chuyển đổi tiếng Việt sang văn bản, các vấn đề lý thuyết trong các chương trước đã được áp dụng cụ thể trong thực tế. Qua chương này ta cũng đã nắm được phương pháp xây dựng một hệ thống nhận dạng và chuyển đổi âm thoại tiếng Việt sang văn bản.

Qua các thực nghiệm và phân tích kết quả một số nhận xét và đánh giá được rút ra như sau:

- ✓ Hàm phát xạ quan sát với nhiều hàm Gauss tạo ưu thế hơn hẳn so với hàm phát xạ quan sát với một hàm Gauss.
- ✓ Thử nghiệm với tập dữ liệu huấn luyện và tập dữ liệu kiểm tra trùng nhau cho độ chính xác cao hơn rất nhiều so với hệ thống trước: 78.23% so với 87.70% với các hệ số MFCC. Điều này cho thấy một khả năng có thể xây dựng hệ thống nhận với độ chính xác cao hơn nếu chúng ta có cơ sở dữ liệu đầy đủ và chất lượng tốt.
- ✓ Tỷ lệ lỗi nhận dạng nhầm còn nhiều. Một trong những nguyên nhân là chất lượng thu âm qua điện thoại thấp.

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.

PHỤ LỤC

Tổng quan về HTK

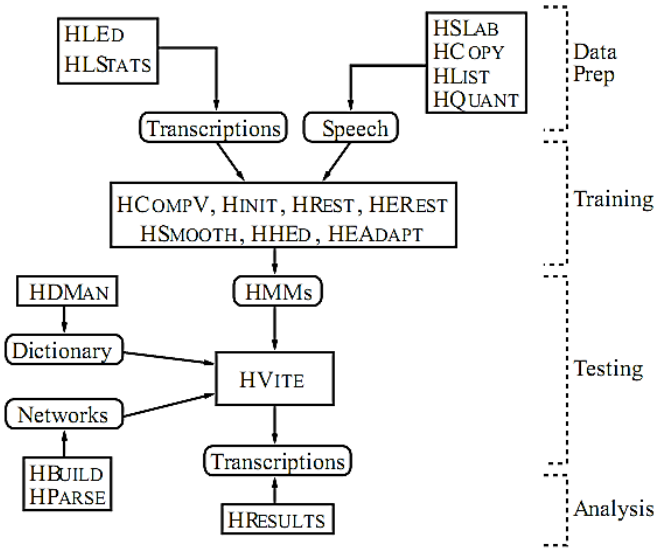
HTK (Hidden Markov Model Toolkit) là một bộ công cụ phát triển để xây dựng các mô hình Markov ẩn cho nhiều bài toán khác nhau, tuy nhiên HTK được thiết kế cho mục đích chính là phát triển các hệ thống nhận dạng tiếng nói. HTK là một bộ thư viện được viết trên ngôn ngữ C cung cấp các hàm liên quan đến trích chọn đặc trưng, xây dựng và huấn luyện mô hình HMM, bộ giải mã, huấn luyện thích nghi,… HTK được xây dựng đầu tiên bởi một nhóm nghiên cứu về học máy thuộc trường đại học Cambridge. Chức năng chính của HTK là dùng để huấn luyện các mô hình HMM dựa trên một tập các mẫu đã được gán nhãn trước. Sau đó HTK có thể sử dụng các mô hình HMM đã được huấn luyện để đoán nhận nhãn cho một tập mẫu khác [Young 2009].

Một cách tổng quát các công cụ của HTK có thể chia ra làm bốn nhóm dựa theo quy trình để xây dựng một hệ thống nhận dạng tiếng nói như hình 3-5.

Trong đó:

- **Data preparing:** Bước chuẩn bị cơ sở dữ liệu. Tại bước này HTK hỗ trợ việc ghi, soạn các file âm thanh thông qua hàm HSLab. Tính toán đặc trưng thông qua hàm Hcopy. Hcopy hỗ trợ tính toán các loại đặc trưng như MFCC, PLP, Fillter bank,…Soạn và tạo các phiên âm (transription) bằng hàm HLed.
- **Training**: Đầu tiên các mô hình HMM sẽ được khởi tạo các tham số ngẫu nhiên ban đầu theo cấu hình đã chọn bằng hàm HInit. Sau đó các mô hình này được huấn luyện ở mức đơn âm (monophone) bằng hàm HRest. Các mô hình cho các âm buộc hay còn gọi là âm phụ thuộc ngữ cảnh (triphone) được tạo ra bằng hàm Hhed dựa trên tập các mô hình đơn âm đã có, sau đó các mô hình này được huấn luyện lại bằng công cụ HERest.
- **Testing**: HTK cung cấp hai bộ nhận dạng là HVite và HDecode. HVite được sử dụng cho các hệ thống nhận dạng sử dụng mô hình ngôn ngữ ở mức 2-gram hoặc grammar. HDecode được sử dụng cho các hệ thống nhận dạng từ vựng lớn và sử dụng mô hình ngôn ngữ từ 3-gram trở lên.

- **Analysis**: Để đánh giá chất lượng nhận dạng của mô hình trên một tập mẫu đầu vào HTK cung cấp hàm HResults để tính toán các tham số độ chính xác theo từ (Word Accuracy - ACC) và độ chính xác theo câu (Sentence Accuracy).



Hình 3- 6: **Quy trình xây dựng một hệ thống nhận dạng tiếng nói trên HTK** [Young 2009]